Pattern Recognition Exercise Book (B4B33RPZ/BE5B33RPZ)

RPZ team

November 21, 2023

Problems are organized by lecture topics. There are representative problems from tests / exam in previous years, problems recommended for working through lecture / lab materials and more advanced problems.

$\oplus$ – Problems aligned with the lab.
$\odot$ – Problems from previous years tests with solution.
$\star$ – More advanced problem, not to be expected at the exam.

Contents

0 Probability

Problem 0.1 (Bayes Theorem)
Prove the Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

using the axioms of probability:
Axiom 1: $0 \leq P(A) \leq 1$, with $P(A) = 1$ if A is certain.
Axiom 2: If events (A_i), $i = 1, 2, \ldots$ are pairwise incompatible (exclusive) then $P(\bigcup_i A_i) = \sum_i P(A_i)$.
Axiom 3: $P(A \cap B) = P(B|A)P(A)$.

Problem 0.2 (Marginal/Conditional Probabilities)
Consider the same example as in the lecture. The joint probability p_{XK} is given by the table:

	cloudiness			
	1	2	3	4
rain	0.02	0.12	0.09	0.04
no rain	0.38	0.28	0.06	0.01

a) Compute marginal probabilities $p_K(k)$ for $k = \{\text{rain}, \text{no rain}\}$ and $p_X(x)$ for $x = \{1, 2, 3, 4\}$.

b) Compute the probability that the cloudiness is less or equal than 2 given that there was a rain.

Problem 0.3 (Bayes Theorem)
Suppose we have a test for cancer with the following statistics:

- The test was positive in 98% of cases when subjects had cancer;
- The test was negative in 97% of cases when subjects did not had cancer;
- Suppose that 0.1% of the entire population have this disease.

A patient takes a test. Denote the variables as: $C \in \{y, n\}, T \in \{+, -\}$.

a) Compute the probability that a person who test positive has this disease.

b) Compute the probability that a person who test negative does not have this disease.

1 Bayesian Decision Theory

Problem 1.1 (Umbrella Rain)
Consider the setup as in Problem 0.2. You have three possible decisions $D = \{$umbrella, no umbrella, 100$\}$ to make on a given day:

- umbrella : you take an umbrella with you,
- no umbrella: you do not take an umbrella with you and if it rains, you will get wet,
- 100: you do not take an umbrella with you but you make a fixed decision that if it rains, you will buy a new umbrella for 100 CZK.

Let the loss (cost) matrix $W(k, d)$ be as follows:

	umbrella	no umbrella	100
rain	0	10	5
no rain	5	-2	0

Compute:

a) The chance of rain given the cloudiness 2?

b) The expected cloudiness on a rainy day?

c) The risk of not having umbrella if the cloudiness is 2 (called partial risk)?

d) The risk of not having umbrella ever?

e) The risk of always carrying an umbrella?

f) The optimal strategy $q^*(x)$

Problem 1.2 (Coarse Decision Space)
Assume weather classes: $K = \{sunny, cloudy, rain, hailstorm\}$. You want to go for a walk, but plan to stay inside if the weather is not $k = sunny$. Given a measurement x from your UltimateWeatherSensor™, you calculated the posterior probabilities of the current weather as $p_{K|X}(\cdot \mid x) = (0.4, 0.2, 0.2, 0.2)$. The task is to decide whether it is sunny, $D \in \{sunny, not\ sunny\}$. What is the optimal Bayesian decision in the following cases (explain):

a) The cost of a correct decision is zero and the cost of a wrong decision is a constant $C > 0$ (normal person).

b) Mistakenly deciding $d = sunny$ costs twice less than mistakenly deciding $d = not\ sunny$ (an active person that does not care *that* much about getting wet).

Problem 1.3 ($\oplus$ 0-1 loss, 2 classes, Gaussian Conditionals)
Recall the optimal decision strategy q minimizes the risk:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) W(k, q(x)). \tag{2}$$

Consider 0-1 loss function:

$$W(k, d) = \begin{cases} 1, & \text{if } k \neq d \\ 0, & \text{if } k = d. \end{cases} \tag{3}$$

a) Prove: $q(x) = \arg\max_d p(d|x)$.

b) Let additionally $K = \{0, 1\}$. Prove $q(x)$ takes the form: $\frac{p(x|k=0)}{p(x|k=1)} \lessgtr \theta$.

c) Consider Gaussian Measurements:

$$p(x|k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x - \mu_k)^2}{2\sigma_k^2}). \tag{4}$$

Prove $q(x)$ takes the form $ax^2 + bx + c \lessgtr 0$.

Problem 1.4 (Error Correcting Codes)
A digital signal transmitting system reads 3 binary digits and for i-th digit outputs the probability that the digit is 1, the resulting probabilities are $0.3, 0.4, 0.7$. It is know that the true digits form an error correcting code where the last digit is always the sum of the first two digits modulo 2.

a) Recognize which number is encoded by the first two digits.

b) Decide whether this packet of 3 digits has to be requested again considering that the cost of skipping an error is $100\times$ more than requesting to repeated the packet.

Problem 1.5 (Gaussian, 3 classes)
We need to classify objects into three classes $k \in \{1, 2, 3\}$. The classes are equally probable a priory. Observations x of objects in class 1 follow the distribution $\mathcal{N}(0, 1^2)$. Recall $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Similarly, in classes 2, and 3, the observation are distributed as $\mathcal{N}(0, 2^2)$ and $\mathcal{N}(3, 2^2)$, respectively.
What is the optimal Bayesian decision $d \in \{1, 2, 3\}$ for the two observations $x = 1$ and $x = 0$ in the following cases:

- if the loss matrix is

$$W_a = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

- if the loss matrix is

$$W_b = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

What is the probability of incorrect decision ($d \neq k$) for the first case and observation $x = 1$?

Problem 1.6 ($\star$ Exam with Bernoulli Chain)
A student prepares for the exam in RPZ. There are K topics in total, one for each lecture. Because the lectures are sequential, he prepares sequentially. He learns the first topic with probability q. If he already learned k topics, he learns the next one again with probability q or otherwise stops preparing.
At the exam he gets a randomly drawn topic. Assume the student answered well on the topic with number x. The task is to recognize whether he/she has prepared at least half of the topics (assume K is even). Model the problem as a Bayesian decision:

a) In this problem, what is the hidden state, observation, decision?

b) What is the probability that he/she learned at least half of the topics?

c) Derive the optimal Bayesian decision strategy.

2 Non-Bayesian Tasks: Minimax, Neuman-Pearson, Wald

Problem 2.1 (Student Wants to Marry - *Lecture Example*)
An aging student at CTU wants to marry. He can't afford to miss recognizing a girl when he meets her, therefore he sets the threshold on overlooking an opportunity as $\bar{\epsilon}_D = 0.2$. At the same time, he wants to minimize mis-classifying boys for girls. The exact setup is as follows:

- Hidden states $K = \{\text{D}, \text{N}\} \equiv \{\text{F}, \text{M}\}$ (female, male)
- Measurements $X = \{\text{short}, \text{normal}, \text{tall}\} \times \{\text{ultralight}, \text{light}, \text{avg}, \text{heavy}\}$
- Prior probabilities do not exist
- Conditional probabilities $p(x|k)$ are given as follows:

$p(x|\text{F})$

short	.197	.145	.094	.017
normal	.077	.299	.145	.017
tall	.001	.008	.000	.000
	u-light	light	avg	heavy

$p(x|\text{M})$

short	.011	.005	.011	.011
normal	.005	.071	.408	.038
tall	.002	.014	.255	.169
	u-light	light	avg	heavy

Find the optimal strategy when you

a) formulate the task as a Neuman-Pearson problem

b) formulate the task as a minimax problem

Problem 2.2 (Neyman-Pearson Continuous Measurement)
Suppose that you have a two-class decision problem $y \in \{1, 2\}$ with real-valued features $x \in [0, 1]$ and that only the class conditional probabilities $p(x|y = 1) = 1$ and $p(x|y = 2) = x + 0.5$ are given.

a) Write down formally the Neyman-Pearson problem formulation.

b) Find the optimal Neyman-Pearson strategy for this decision problem when $y = 2$ is the dangerous state and the probability of overlooked danger shouldn't be higher than 0.1.

Problem 2.3 ($\odot$ Minimax - *Test Example*) Suppose that you have a two-class decision problem $y \in \{1, 2\}$ with the real-valued features $x \in [-1, 1]$ and that only the class conditional probabilities $p(x|y{=}1) = \max(-x, x) = |x|$ and $p(x|y{=}2) = \min(1 + x, 1 - x) = 1 - |x|$ are given.

a) Write down formally the Minimax problem formulation.

b) How many likelihood thresholds are in the solution? Why?

c) Find the optimal Minimax strategy for this decision problem. Any informal solution, e.g. geometric, will be awarded by 0 points.

Problem 2.4 ($\oplus$ Worst Bayes / Minimax)
Consider a binary classification problem ($K{=}\{1, 2\}$) with continuous features $x \in \mathbb{R}$. Suppose you have obtained the optimal Bayesian strategy q for the case when the proportion of classes was given by $p(k{=}1) = \pi^*$ (let's call it training distribution prior). Suppose at the test time the proportion of classes changes. What is the worst case performance of the strategy q?

a) How does the risk $R(q) = \sum_k \int p(x, k) W(k, q(x)) dx$ vary as a function of the parameter $\pi = p(k{=}1)$?

b) Show that the maximum over π is achieved either at $\pi = 0$ or at $\pi = 1$.

c) Express the value of the risk in the worst case. Assuming also $0{-}1$ loss, compare the worst risk to the objective of the minimax problem.

d) When the risk $R(q)$ viewed as a function of π is a constant function? Assume this is the case and q is in the form of the likelihood ratio test. Show that q then is the solution to the minimax problem.

3 Parameter Estimation, Maximum Likelihood

Problem 3.1 (Bernoulli Coin)
You observed random and independent draws of an unfair coin, the draws were (H, H, H, T, T).

a) Define the probability model and the likelihood of all observations. Find the maximum likelihood estimate of the heads probability of the coin, p_{ML}.

b) Let the heads probability be parametrized as $p(H) = \frac{1}{1+e^{-\eta}}$ for $\eta \in \mathbb{R}$. What is the maximum likelihood estimate of η? Does there always hold $p_{\mathrm{ML}} = \frac{1}{1+e^{-\eta_{\mathrm{ML}}}}$, why?

Problem 3.2 (Binomial Socks)
You have red and blue socks in the drawer. Let the unknown proportion of red to the total number of socks be denoted by π. You draw them randomly with replacement.

a) What does it mean to draw with replacement?

b) What is the probability of drawing two red socks in a row?

c) In $N = 10$ draws you got $R = 2$ red socks. What is the probability to draw R red socks in N trials?

d) Compute the maximum likelihood estimate of π.

Problem 3.3 ($\star$ Gaussian)
The density of a multivariate Normal distribution is given by

$$p(x) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \tag{5}$$

where $x, \mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succcurlyeq 0$. You are given i.i.d. observations $(x_i)_{i=1}^N$.

a) Find the maximum likelihood estimate of the mean μ. *Hint: optimal μ does not depend on Σ.*

b) Find the maximum likelihood estimate of the covariance matrix Σ.
Hints: Use log-likelihood; the mean μ is known from above and can be substituted in the end; You can solve for the precision matrix $\Lambda = \Sigma^{-1}$, it is simpler to differentiate with respect to it; use the following identities: $\det(\Sigma^{-1}) = \det(\Sigma)^{-1}$, $\frac{\partial \det(\Lambda)}{\partial \Lambda} = \Lambda^{-1}\det(\Lambda)$.

Problem 3.4 (Exponential Lamp Lifetime)
At a lamp factory, bulbs are tested in order to know their lifetime. They tested N bulbs and obtained a set of measurements $\mathcal{T} = \{t_1, ..., t_N\}$, where t_i is the time that the light bulb stood before it burned out. We will assume that the measurements follow the exponential distribution:

$$p(t) = \lambda e^{-\lambda t}, \;\; t \in [0, \infty).$$

a) Find the maximum likelihood estimate of the parameter λ.

b) Let us consider an alternative parameterization

$$p(t) = \tfrac{1}{\theta} e^{-\frac{t}{\theta}}.$$

Verify that the expected lifetime of a lamp is θ. Find the maximum likelihood estimate of θ. How does it relate to the rate λ?
Hint: The expected lifetime is the mean of the distribution defined as $\mathbb{E}[t] = \int_0^\infty tp(t)dt$. It can be calculated using integration by parts. An alternative solution is to note that θ_{ML} is the sample mean $\frac{1}{n}\sum_i t_i$ and with $n \to \infty$ the former approaches the true parameter θ and the later the distribution mean.

c) Assume that our prior knowledge about λ is expressed by the distribution $p(\lambda) = e^{-\lambda}$. Derive the MAP estimate of λ, λ_{MAP}. What is the expected lifetime of a bulb for $\hat{\lambda}_{\mathrm{MAP}}$?

Problem 3.5 (Exponential Hard Drives)
The reliability of hard drives is defined by the probability density function $p(t) = \lambda e^{-\lambda t}, t \in (0, \infty)$.

a) What is the maximum likelihood estimate of the failure rate λ if in an experiment with three hard drives the following lifetimes have been observed: $t_1 = 56, t_2 = 120$ and $t_3 = 424$?

b) If the test described above finished at time $T = 300$ and one hard drive would still be running, i.e. $t_1 = 56$ and $t_2 = 120$ are known but about the time of failure of the third hard drive we know only that $t_3 \geq T$. Formulate the likelihood function in this case. Find the ML estimate of λ.

Problem 3.6 ($\star$ German Tank)
During the second world war, British intelligence service had collected information about serial numbers of German tanks ever seen. Suppose serial numbers $x_1, x_2, \ldots x_n$ have been seen. Assume that x_i are independent and follow a uniform distribution with the density

$$p(x) = \frac{1}{\theta}\delta_{\{0 \leq x \leq \theta\}}, \tag{6}$$

where θ is the total number of tanks produced by Germany (we assume $x, \theta \in \mathbb{R}$ for simplicity).

a) What is the maximum likelihood estimate of θ, θ_{ML}?

b) Does the intuition suggest that θ_{ML} underestimates the real number of tanks Germany has?

c) Assume that apriori Germany had capacity to manufacture up to M tanks. What is the maximum a posteriori estimate of θ, θ_{MAP}?

d) Treating θ as a random variables, what is the posterior distribution of θ given the observations. What is the Bayesian estimate of θ minimizing the mean squared error, θ_{MSE}?

4 Nearest Neighbour, Non-Parametric Density Estimation

Problem 4.1 (Piece-wise constant density)
Let $\{x_i\}_{i=1}^n$ be independent observations with $x_i \in [0, 1]$. The domain $[0, 1]$ is partitioned into K equal size segments denoted Δ_k. The piece-wise density model is defined on $[0, 1]$ as

$$p(x) = \begin{cases} d_1, & \text{if } x \in \Delta_1 \\ \ldots & \\ d_K, & \text{if } x \in \Delta_K, \end{cases}$$

where $(d_k \geq 0 \mid k = 1 \ldots K)$ are parameters.

a) Estimate the parameter vector d using the maximum likelihood.
Hint: use the constraint that the density must integrate to 1.

Problem 4.2 (K-nearest neighbours)
Describe the K-NN algorithm and list its pros and cons.
With the following training set with data points (x, y) (measurement, class), classify point $x = 5$ using 1-NN, 3-NN and 5-NN classifier.

$$\mathcal{T} = \{(0, A), (-1.5, A), (10, B), (2, A), (4.5, A), (3, B), (6, B), (9, B), (1.5, A), (11, B)\}$$

Problem 4.3 (K-D trees)
Describe the algorithm for building a K-D tree.
Make a K-D tree (alternating X- and Y- cuts) from the following data:
(2,3), (4,7), (5,4), (7,2), (8,1), (9,6)
Describe how to search for **exact** nearest neighbour using a K-D tree.

Problem 4.4 (Parzen Windows)
Given the measurements $X = \{1, -1, 1, 3, 2, 0\}$, plot the non-parametric estimate of a distribution $p(x)$ using the Parzen window method with a kernel function $K(x, y) = k(x - y)$ and $k(z)$ defined as:

$$\begin{aligned} k(z) &= 1/h \quad \text{for } |z| \leq h/2 \ , \\ k(z) &= 0 \qquad \text{for } |z| > h/2 \ , \end{aligned}$$

for $h = 2$.

Problem 4.5 (Parzen Window Re-weighting)
Suppose we have training points $\{x_i\}_{i=1}^n$ and found a Parzen density estimate

$$p(y) = \frac{1}{n} \sum_i K(y - x_i)$$

using a fixed kernel K. Here all kernel copies have equal weights $\frac{1}{n}$. Consider giving kernels at different positions a different weight π_i:

$$p(y; \pi) = \sum_j \pi_j K(y - x_j),$$

where $\pi_j \geq 0$ must sum to 1 to ensure p is a density. Re-estimate coefficients π by maximizing the following lower bound on the log likelihood:

$$\sum_i \log p(x_i; \pi) \geq \sum_i \sum_j \frac{K_{i,j}}{K_i} \log \pi_j K_i,$$

where $K_{ij} = K(x_i - x_j)$ and $K_i = \sum_i K_{i,j}$. Later we will see that this is the first iteration of the EM algorithm initialized with $\pi_j = \frac{1}{n}$.

5 Logistic Regression

Problem 5.1 ($\oplus$ Logistic Regression from Bayes Decision)
Consider a recognition problem with two hidden states $K = \{-1, 1\}$ and $\mathbf{x} \in \mathbb{R}^d$. We know that the optimal decision expresses in many cases using the likelihood ratio.

a) Assume that the log odds is linear: $\log \frac{p(k=1|x)}{p(k=-1|x)} = w^\mathsf{T} x$. Knowing also that $\sum_k p(k|x) = 1$, find probabilities $p(1|x)$, $p(-1|x)$.

b) Assume that $p(k|x)$ is logistic (as derived in a), $p(x)$ exists but is unknown and does not depend on parameters. Given the training data points $\{x_i, k_i\}_{i=1}^n$, express the negative log likelihood of the data (up to an additive constant).

c) Plot the miss-classification indicator $[\![k \neq \operatorname{sign}(w^\mathsf{T} x)]\!]$ as a function of $z = kw^\mathsf{T} x$. Plot the function $\log(1 + e^{-kw^\mathsf{T} x})$ as a function of same z. Using convexity of this function, show that the negative log likelihood in b) is convex.

Problem 5.2 ($\oplus$ Properties of Logistic Sigmoid Function)
The logistic (sigmoid) function is $\sigma(z) = \frac{1}{1+e^{-z}}$. The following will be helpful for logistic regression but also for multi-layer neural networks.

a) Show that $\sigma(-z) = 1 - \sigma(z)$;

b) Show that $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1 - \sigma(z))$;

c) Compute $\frac{\partial}{\partial z} \log \sigma(z)$ (using b);

d) Compute $\frac{\partial}{\partial x} \log \sigma(w^\mathsf{T} x)$ (using c);

e) Show that $-\log \sigma(z)$ is convex (using monotonicity of the first derivative or non-negativity of second derivative)

6 Linear Classifier, Perceptron

Problem 6.1 (Linear Classifier)
Let $x \in \mathbb{R}^d$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$. Consider a linear classifier defined as

$$q(x) = \begin{cases} 1, & \text{if } a^\mathsf{T} x + b \geq 0, \\ -1, & \text{if } a^\mathsf{T} x + b < 0. \end{cases}$$

a) Find the distance from a given point x to the decision boundary of the classifier.
(*Hint: find y on the decision boundary (i.e. satisfying $a^\mathsf{T} y + b = 0$) such that $\|x - y\|^2$ is minimal. Then $d = \|x - y\|$ is the thought distance.*)

b) Let $k_i \in \{1, -1\}$ be the true class of $x_i \in \mathbb{R}^d$ (data point x_i is a vector with coordinates $x_{i,j}$, $j = 1 \ldots d$). Define $\bar{x}_i \in \mathbb{R}^{d+1}$ and $w \in \mathbb{R}^{d+1}$ such that q classifies x_i correctly iff $w^\mathsf{T} \bar{x}_i \geq 0$.

Problem 6.2 (Perceptron)
Consider training points $(x_i, k_i)_{i=1}^N$, where $x_i \in \mathbb{R}^d$, $k_i \in \{-1, 1\}$ and let $\bar{x}_i$ be as derived in Problem 6.1 (b). Consider the approximation to the empirical loss:

$$\tilde{R}(w) = \frac{1}{N} \sum_i \max(-w^\mathsf{T} \bar{x}_i, 0), \tag{7}$$

a) Plot this approximation for one data point (x, y) as a function of $z = w^\mathsf{T} \bar{x}$.

b) On the same graph, plot also the empirical loss of one data point as a function of $z = w^\mathsf{T} \bar{x}$.

c) On the same graph, plot also the log likelihood of data point (x, y) in the logistic regression model as a function of z. (see Problem 5.1).

Problem 6.3 (Perceptron)
We will see connection between Perceptron algorithm and stochastic gradient descent for risk approximation (7). Note, the common Perceptron algorithm considers data points sequentially. For a training data as in Problem 6.2, let $l_i(w) = \max(-w^\mathsf{T} \bar{x}_i, 0)$ be the approximate error of one data point and so $\tilde{R}(w) = \frac{1}{N} \sum_i l_i(w)$.

a) Apply stochastic gradient descent to $\tilde{R}$. A step of SGD picks a data point i at random and performs an update

$$w^{t+1} = w^t - \varepsilon \nabla_w l_i(w).$$

(Compute the gradient and simplify what possible).

b) Show that when starting with $w^0 = 0$, the classification boundary at step t is invariant of the step size ε.
(*Hint: inspect what the algorithm would do if starting from w^0 with different values of ε, e.g. $\varepsilon = 1$ and some $\varepsilon \neq 1$*).

Problem 6.4 (Perceptron)
A training set is given in the format $T = \{(\boldsymbol{x}_i; k_i)\}$, where $i = 1 \ldots 5, \boldsymbol{x}_i \in \mathbb{R}^2$, and $k \in \{1, -1\}$:

$$T = \{(-2, 1; -1), (0, 0; -1), (0, 2; 1), (0, -3; -1), (2, 2, 1)\}.$$

a) Find a linear classifier by the Perceptron algorithm, i.e. find a vector $\boldsymbol{w} \in \mathbb{R}^2$ and offset $b \in \mathbb{R}$ such that $y = \boldsymbol{w}^\mathsf{T} \boldsymbol{x} + b$ is positive for samples of class $k = 1$ end negative for $k = -1$. More specifically, what are the vector $\boldsymbol{w}$ and offset b after ten steps of the Perceptron algorithm?

Problem 6.5 (Perceptron) For the iterations of the Perceptron algorithm expressed as:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \boldsymbol{x}^t,$$

where $\boldsymbol{x}^t$ is the data point selected at step t as misclassified when using weights $\boldsymbol{w}^t$.

a) Starting with $\boldsymbol{w}^0 = 0$, show that for any t there holds $\|\boldsymbol{w}^t\|^2 \leq t \max_i \|\boldsymbol{x}_i\|^2$.
(*Hint: use induction)*

7 Support Vector Machines

Problem 7.1 (Soft Margin SVM Loss) Consider the primal soft margin SVM formulation:

$$\begin{aligned}\min_{\boldsymbol{w},b,\xi}\ & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i \\ & (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b)y_i \geq 1 - \xi_i \quad \forall i, \\ & \xi_i \geq 0 \quad \forall i.\end{aligned}$$

When in doubts, consult lecture slides. The tasks follow the lecture explanation.

a) Assuming all variables but ξ_i for one i are fixed, find the optimal solution for ξ_i. *Hint: write the respective minimization problem in one variable ξ_i, solve it graphically.*

b) Using the optimal values for ξ_i, reformulate the SVM problem as unconstrained optimization

$$\min_{\boldsymbol{w},b}\ \frac{1}{2C}\|\boldsymbol{w}\|^2 + \sum_i \max(1 - (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b)y_i, 0). \tag{8}$$

What optimization methods do you know, that can be applied to solve it?

c) Denoting $z = (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - b)y_i$, plot the function $\max(1 - z, 0)$. C.f. the loss function of Perceptron in Problem 6.2, and logistic regression in Problem 5.1.

d) ($\star$) When $C \to \infty$, does formulation (8) become equivalent to the optimization problem of Perceptron?

e) ($\star$) For a data point i, chosen at random, find the gradient of the function

$$\frac{1}{2nC}\|\boldsymbol{w}\|^2 + \max(1 - (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b)y_i, 0)$$

in $\boldsymbol{w}$ and b and write a gradient descent step.

Problem 7.2 (Hard Margin SVM)
Let $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. Consider a linear classifier defined as

$$q(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b).$$

a) We want the data points with $y_i = +1$ to satisfy $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b \geq \varepsilon$ with $\varepsilon > 0$ and data points with $y_i = -1$ to satisfy $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b \leq -\varepsilon$, i.e. to be on the right side of the decision boundary while also separated away from zero. Write the inequalities for the two cases $y = \pm 1$ in the unified form.

b) Compute the value of the margin, i.e. distance between the two hyperplanes

$$\begin{aligned}\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b &= \varepsilon \\ \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b &= -\varepsilon.\end{aligned}$$

Hint: Consider two points $\boldsymbol{x}$ and $\boldsymbol{y}$ on the respective hyperplanes and the normal vector to the hyperplanes $\boldsymbol{n} = \boldsymbol{w}/\|\boldsymbol{w}\|$. Project $\boldsymbol{x} - \boldsymbol{y}$ onto $\boldsymbol{n}$.

c) Formulate the SVM problem (to maximize the margin while satisfying the correct classification constraints with the margin) as a quadratic minimization.

Problem 7.3 (SVM: Vector Notation, Dual) Consider the primal soft margin formulation

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i \\ & (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + b)y_i \geq 1 - \xi_i \quad \forall i, \\ & \xi_i \geq 0 \quad \forall i,\end{aligned}$$

where $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is the training data with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ and $\boldsymbol{\xi} \in \mathbb{R}^n$. Let $\bar{\boldsymbol{X}} \in \mathbb{R}^{d,n}$ be the matrix of all data points multiplied by their class sign, so that $\bar{\boldsymbol{X}}_{:,i} = \boldsymbol{x}_i y_i$. Write the SVM problem using a matrix notation as

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w} + C\mathbf{1}^\mathsf{T}\boldsymbol{\xi} \\ & \bar{\boldsymbol{w}}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} \geq \mathbf{1}^\mathsf{T} - \boldsymbol{\xi}^\mathsf{T} \\ & \boldsymbol{\xi} \geq 0,\end{aligned}$$

where inequalities are coordinate-wise.

a) For the constraint $\boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} \geq \mathbf{1}^\mathsf{T} - \boldsymbol{\xi}^\mathsf{T}$ introduce a non-negative vector of Lagrange multipliers $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ to express the constraint as a barrier:

$$\max_{\boldsymbol{\alpha}\geq 0} -(\boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} - \mathbf{1}^\mathsf{T} + \boldsymbol{\xi}^\mathsf{T})\boldsymbol{\alpha} = \begin{cases} 0, & \text{if} \quad \boldsymbol{w}^\mathsf{T}\bar{\boldsymbol{X}} + b\boldsymbol{y}^\mathsf{T} - \mathbf{1}^\mathsf{T} + \boldsymbol{\xi}^\mathsf{T} \geq 0, \\ \infty, & \text{otherwise.} \end{cases}$$

You should obtain the problem reformulation in the form

$$\min_{\boldsymbol{w},b;\xi\geq 0} \max_{\boldsymbol{\alpha}\geq 0}(\dots).$$

b) Swap the minimization and maximization and solve analytically for $\boldsymbol{w}, b$ and $\boldsymbol{\xi}$ to obtain maximization in $\boldsymbol{\alpha}$ only.
Hint for $\boldsymbol{w}$: use critical point conditions to obtain that $\boldsymbol{w} = \bar{\boldsymbol{X}}\boldsymbol{\alpha}$.
Hint for $\boldsymbol{\xi}$: use that

$$\min_{\boldsymbol{\xi}\geq 0}(C\mathbf{1} - \boldsymbol{\alpha})^\mathsf{T}\boldsymbol{\xi} = \begin{cases} 0, & \textit{if} \quad C\mathbf{1} - \boldsymbol{\alpha} \geq 0, \\ -\infty, & \textit{otherwise}, \end{cases}$$

in order to eliminate $\boldsymbol{\xi}$ by introducing the constraint $\boldsymbol{\alpha} \leq C$.
Hint for b: use that

$$\min_b b\boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = \begin{cases} 0, & \textit{if} \quad \boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0, \\ -\infty, & \textit{otherwise}, \end{cases}$$

in order to eliminate b by introducing the constraint $\boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0$.
You should obtain the dual formulation:

$$\begin{aligned}\max_{\boldsymbol{\alpha}} \ & -\frac{1}{2}\boldsymbol{\alpha}^\mathsf{T}\bar{\boldsymbol{X}}^\mathsf{T}\bar{\boldsymbol{X}}\boldsymbol{\alpha} + \mathbf{1}^\mathsf{T}\boldsymbol{\alpha}, \\ & 0 \leq \boldsymbol{\alpha} \leq C, \\ & \boldsymbol{y}^\mathsf{T}\boldsymbol{\alpha} = 0.\end{aligned}$$

Problem 7.4 ($\odot$ Support Vectors) Consider soft-margin SVM for 3 given data points $(\boldsymbol{x}_i, y_i)$, $i = 1 \dots 3$ and suppose that the optimal dual solution is the vector $\boldsymbol{\alpha} = (0, C/2, C)$.

a) Which of the three points are support vectors?

b) Express the optimal $\boldsymbol{w}$.

c) It is known that for the optimal dual solution $\boldsymbol{\alpha}$ the optimal primal solution $(\boldsymbol{w}, b, \boldsymbol{\xi})$ must satisfy the following conditions (complementary slackness):

$$(y_i(\boldsymbol{w}^\mathsf{T}\boldsymbol{x_i} + b) - 1 + \xi_i)\alpha_i = 0 \quad \forall i,$$
$$\xi_i(\alpha_i - C) = 0 \quad \forall i.$$

Using these conditions, find the optimal b and $\boldsymbol{\xi}$.

Problem 7.5 (Kernel SVM) Suppose the input features $\boldsymbol{x}_i$ are first lifted to a higher dimension using a lifting function $\phi(\boldsymbol{x})$.

a) Write the dual SVM formulation in this case.

b) Denote the kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$. Does there always exist a corresponding ϕ for any given mapping $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$?

c) How the classification of the test data can be expressed knowing only the black box computing $k(x, x')$ and not knowing the underlying lifting $\phi(x)$?
Hint: use the support vector form, $\boldsymbol{w} = \bar{\boldsymbol{X}}\boldsymbol{\alpha}$ as derived in Problem 7.3 b).

Problem 7.6 (Kernels and Feature Maps)

a) Express the kernel function $k(\mathbf{x}, \mathbf{x}')$ that corresponds to feature map $\phi(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$ assuming 2-D vectors $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

b) Derive the feature map $\phi(\mathbf{x})$ that corresponds to kernel function $k(\mathbf{x}, \mathbf{x}') = (1 + x^\mathsf{T} x')^2$ for $\mathbf{x} \in \mathbb{R}^2$.

c) Let us have a kernel function $k(\mathbf{x}, \mathbf{x}') = (1 + x^\mathsf{T} x')^d$ and $\mathbf{x} \in \mathbb{R}^D$ for $d, D \in \mathbb{N}$. Do we know how to compute feature map for given d and D? Compare the computation of the explicit feature map and the kernel function.

8 AdaBoost

Problem 8.1 (AdaBoost Loss)
Consider a classifier of the form:

$$H(x) = \text{sign}(f(x)), \text{ where } f(x) = \sum_{t=1}^{T} \alpha_t h_t(x).$$

a) Verify that the empirical risk on the training set $(x_i, k_i)_{i=1}^n$ can be written as

$$\hat{R} = \sum_i [\![f(x_i) y_i < 0]\!].$$

b) Consider the empirical risk for one data point (x, y) and show that it is upper bounded by $e^{-f(x)y}$. Plot the empirical risk and the exponential upper bound as functions of $z = f(x)y$. We will derive AdaBoost algorithm by minimizing

$$\tilde{R} = \sum_i e^{-f(x_i)y_i}. \tag{9}$$

What are potential advantages of this loss function for optimization in α and as a surrogate for empirical risk?

c) Consider that T is given and $(h_t(x) \mid t = 1 \dots T)$ is a fixed "lifting". Identify $H(x)$ as a linear classifier. Compare the exponential loss of AdaBoost with the loss functions of Perceptron in Problem 6.2, logistic regression Problem 5.1 and SVM Problem 7.1.

Problem 8.2 (AdaBoost Greedy Objective)
Consider that we have already decided and fixed the part

$$f_{t-1}(x) = \sum_{k=1}^{t-1} \alpha_k h_k(x)$$

and consider adding one more term $\alpha_t h_t(x)$ to the classifier. Let us verify that the added part of the classifier should minimize the same exponential upper bound but with re-weighted data.

a) Write the minimization of (9) in h_t, α_t, grouping all constant factors (not depending on α_t, h_t) together. Show that the problem can be written in the form

$$\min_{\alpha_t, h_t} \sum_i D_t(i) e^{-\alpha_t y_i h_t(x_i)}, \tag{10}$$

where D_t are some coefficients satisfying $D_t(i) > 0$ and $\sum_i D_t(i) = 1$. Find these coefficients and identify them as weights assigned to data points by the already constructed part of the classifier.

Problem 8.3 (AdaBoost Greedy Optimization)
Continue from the greedy problem formulation (10).

a) Expand $e^{-\alpha_t y_i h_t(x_i)}$ using that $h_t(x_i)y_i$ may take only two values, ± 1.
Hint: you should obtain greedy step subproblem:

$$\min_{\alpha_t, h_t} \sum_i D_t(i) \Big([\![h_t(x_i) = y_i]\!] e^{-\alpha_t} + [\![h_t(x_i) \neq y_i]\!] e^{\alpha_t} \Big).$$

b) Using that $[\![h_t(x_i) = y_i]\!] = 1 - [\![h_t(x_i) \neq y_i]\!]$, group factors involving α_t together, to obtain

$$\min_{h_t, \alpha_t} e^{-\alpha_t} \sum_i D_t(i) + (e^{\alpha_t} - e^{-\alpha_t}) \sum_i D_t(i) [\![h_t(x_i) \neq y_i]\!]. \tag{11}$$

Recall that $\sum_i D_t(i) = 1$.

c) Show that for any $\alpha > 0$, the minimization in h_t reduces to

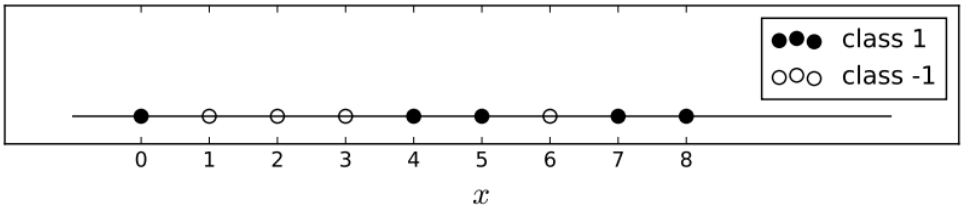
$$h_t = \arg\min_h \sum_i D_t(i) [\![h(x_i) \neq y_i]\!],$$

i.e. h_t minimizes the empirical classification error with data re-weighted by D_t.

d) Given the optimal h_t find the optimal $\alpha_t \in \mathbb{R}$ by minimizing (11). For brevity, denote $\varepsilon_t = \sum_i D_t(i) [\![h_t(x_i) \neq y_i]\!]$, the weighted error achieved by h_t. *Hint: differentiate in* α_t.

e) Under which conditions on ε_t the optimal α_t satisfies the assumption $\alpha_t > 0$ made in (c)?

Problem 8.4 (1-D Adaboost Classifier)
Adaboost learning algorithm. Consider the following 1-D data:



and the following set of weak classifiers: $h(x) = \text{sign}(ax + b) \;\; (a, b \in R)$. Use this example to explain how Adaboost works (make one full iteration, ending with first data re-weighting.)

9 Neural Networks, Backpropagation

Problem 9.1 (What is Backpropagation?)
Find the most consistent matching of concepts on the left and descriptions on the right:

a) Backpropagation	1 A way to learn neural networks.
b) Gradient	2 Method to optimize training loss.
c) Chain rule	3 Is necessary to find a step direction for gradient descent.
d) Training loss minimization	4 A rule to compute gradient of composite functions.
e) SGD	5 Computationally efficient automatic differentiation for scalar-valued composite functions.

Problem 9.2 (What is Gradient, etc.?)
Let $\boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. Find the most consistent matching of concepts on the left and descriptions on the right:

a) Gradient of f

b) Derivative of f

c) Jacobian of f

1 A linear mapping approximating f locally around a point.

2 Expression of the derivative in coordinates as a matrix.

3 Column vector of partial (or total) derivatives in case f is scalar-valued, i.e. $m = 1$.

Problem 9.3 (Backprop for Linear Layer)
Consider a linear layer $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ expressed in coordinates as

$$y_i = \sum_j W_{i,j} x_j. \tag{12}$$

Let $\mathcal{L}(\boldsymbol{y})$ be a real-valued loss function as a function of the layer output $\boldsymbol{y}$. Apply the total derivative rule

$$\frac{d\mathcal{L}}{d\boldsymbol{x}} = \Big(\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}}\Big)\Big(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big)$$

in order to compute the gradient of the loss $\mathcal{L}$ in $\boldsymbol{x}$.

a) Write the total derivative rule in coordinates using $\frac{\partial \mathcal{L}}{\partial y_j}$ and $\frac{\partial y_j}{\partial x_i}$ and substitute (12).

b) Now write the result in the vector form.
Answer: the gradient in x (column vector of derivative coordinates) is $(\nabla_{\boldsymbol{x}} \mathcal{L}) = W^{\mathsf{T}}(\nabla_{\boldsymbol{y}} \mathcal{L})$.

c) Find the gradient of the loss in $\boldsymbol{W}$, i.e. in all coordinates $W_{i,j}$. Note this needs to recall the value of $\boldsymbol{x}$.

10 K-Means Clustering

Problem 10.1 (K-Means Alternating Optimization) Consider a set of points $\{x_i \in \mathbb{R}^d \mid i = 1 \dots N\}$. Let $\tau(i) \in \{1 \dots K\}$ denote the cluster assignment of point i. Let $(c_k \in \mathbb{R}^d \mid k = 1 \dots K)$ denote cluster centers. Consider the problem

$$\min_{\tau, c} \sum_k \sum_{i \mid \tau(i) = k} \|x_i - c_k\|^2 = \min_{\tau, c} \sum_i \|x_i - c_{\tau(i)}\|^2.$$

a) For fixed cluster centers c, solve for optimal partition τ.

b) For fixed partition τ, solve for optimal cluster centers c.

c) Based on the above point show that K-means algorithm never increases the objective function. Show that if the centroids change their position, the objective function strictly decreases. Conclude that the algorithm converges in centroid positions.

Problem 10.2 ($\star$ K-Means, Equivalent reformulation) Let $\mathcal{X}$ be a set of points, $(c_k \mid k = 1 \dots K)$ cluster centers and $\{\mathcal{T}_k \mid k = 1 \dots K\}$ a disjoint partitioning of $\mathcal{X}$. Show that K-means clustering objective is equivalent to sum of all pairwise inter-class distances:

$$\min_{\mathcal{T},c} \sum_k \sum_{x \in \mathcal{T}_k} \|x - c_k\|^2 = \min_{\mathcal{T}} \sum_k \frac{1}{2\|\mathcal{T}_k\|} \sum_{x,y \in \mathcal{T}_k} \|x - y\|^2.$$

Problem 10.3 (K-medians, Villages) There are N villages on a plane at coordinates $p_i = (x_i, y_i)$, $i = 1 \dots N$. Find the positions of K wells to supply water to the villages such that the total construction cost is minimized. The cost of pipes is C_p per meter. (every village should be connected by a pipe to some well).

a) Assume that pipes can be laid without restrictions. Formulate the optimization problem and reduce to K-medians.

b) Consider that the cost of drilling a well is C_w. Propose an algorithm to find the optimal value of the number of wells K.

c) What changes if the pipes can only be laid along vertical or horizontal axis (each pipe is allowed to contain both horizontal and vertical segments)?

11 EM algorithm

Problem 11.1 (K-means / EM)
Let $\{x_i | i = 1 \dots n\}$ be observed points from a joint model $p(x, k)$ where $k \in \{1 \dots K\}$ is a hidden state (not observed). The model $p(x, k)$ is defined as $p(x, k) = p(x|k)p(k)$, where $p(x|k)$ is a Normal density $\mathcal{N}(\mu_k, I)$ with mean parameter μ_k for each k and $p(k)$ is uniform. Assume that initial estimates of μ_k are given.

a) Obtain K-means algorithm as follows. Given current means μ, find the most likely k_i, the assignment of mixture components to points maximizing the posterior distribution $p(k_i|x_i)$. Given current assignment k_i, find the maximum likelihood estimate of μ, i.e. maximizing the joint likelihood of x and k.

b) Obtain EM algorithm as follows. Given current means μ, find the soft assignment $\alpha_i(k) = p(k_i|x_i)$. Given weights $\alpha_i(k)$, find the parameters μ maximizing the "weighted" log-likelihood:

$$\sum_i \sum_k \alpha_i(k) \log p(x_i, k).$$

12 PCA, LDA

Problem 12.1 (PCA: Max variance / Least errors) Let $\{\boldsymbol{x}_i \in \mathbb{R}^D \mid i = 1 \dots N\}$ be the data points. We will assume the data is already centered (zero mean). Let $\{\boldsymbol{u}_j \in \mathbb{R}^D \mid j = 1 \dots M\}$ be M orthonormal vectors. Let $S = \frac{1}{N} \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$ be the $D \times D$ normalized data scatter matrix. Work out PCA derivation following the next steps.

a) Show that sum of variances along M orthonormal directions

$$\sum_{j=1}^{M} \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{u}_j^\mathsf{T} \boldsymbol{x}_i)^2 \tag{13}$$

can be expressed as $\sum_j \boldsymbol{u}_j^\mathsf{T} S \boldsymbol{u}_j$.

b) Consider the reconstruction error of the data when projecting to the subspace of $\{\boldsymbol{u}_j\}$:

$$\sum_i \|\boldsymbol{x}_i - \sum_j \boldsymbol{u}_j^\mathsf{T} \boldsymbol{x}_i \boldsymbol{u}_j\|^2. \tag{14}$$

Show that it is equal to $\sum_i \|\boldsymbol{x}_i\|^2 - \sum_j (\boldsymbol{u}_j^\mathsf{T} \boldsymbol{x}_i)^2$ and thus minimizing the reconstruction error is the same as maximizing the variance of projected data (13).

Problem 12.2 (PCA: Solution) Use Problem 12.1, to formulate the PCA problem as

$$\max_{\{\boldsymbol{u}\}} \sum_j \boldsymbol{u}_j^\mathsf{T} S \boldsymbol{u}_j \quad \text{s.t. } \boldsymbol{u}_j^\mathsf{T} \boldsymbol{u}_k = \delta_{j,k}. \tag{15}$$

a) Find the necessary conditions for the optimum by considering only the constraints $\|\boldsymbol{u}_j\|^2 = 1$ for all $j = 1 \dots M$ (no orthogonality constraints) and solving for stationary point. You should obtain

$$S\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j,$$

where λ_j is the Lagrange multiplier for $\|\boldsymbol{u}_j\|^2 = 1$.

b) Observe that when $\boldsymbol{u}_j$ are distinct eigenvectors and λ_j are eigenvalues of S, the necessary conditions as well as ortonormality constraints are satisfied.

c) Show that the objective (15) for $\boldsymbol{u}_j$ being the first M eigenvectors of S equals

$$\sum_{j=1}^{M} \lambda_j,$$

and thus in order to maximize it we should select the eigenvectors corresponding to M largest eigenvalues of S.

Hints: Every $D \times D$ symmetric real-valued matrix has D orthogonal eigenvectors and has eigen-decomposition $U \Lambda U^\mathsf{T}$ with real eigenvalues. A positive semidefinite symmetric matrix has non-negative eigenvalues.

Problem 12.3 (LDA: Solution) Let S_b be between-class scatter matrix (symmetric, positive semidefinite) and S_w be within-class scatter matrix (symmetric, positive semidefinite, full rank). Solve the LDA problem

$$\max_{\boldsymbol{v}} \frac{\boldsymbol{v}^\mathsf{T} S_b \boldsymbol{v}}{\boldsymbol{v}^\mathsf{T} S_w \boldsymbol{v}}. \tag{16}$$

- Note that the objective is invariant to scaling of $\boldsymbol{v}$. Fix the scale ambiguity by choosing a convenient constraint $\boldsymbol{v}^\mathsf{T} S_w \boldsymbol{v} = 1$ and simplify the problem to a constrained quadratic program.
- Find the necessary condition for optimum from stationary point conditions of the Lagrangian. You should obtain the equation in $\boldsymbol{v}, \lambda$

  $$S_b \boldsymbol{v} = \lambda S_w \boldsymbol{v}, \tag{17}$$

  Convert it to an eigenvalue problem.
- Use the identity (17) to simplify the objective (16) and conclude which eigenvector we should chose to solve LDA.

Problem 12.4 (LDA: data preparation) Let $\boldsymbol{x}_i^k$ be data points for $k = 1, 2$ (two classes). Let

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \boldsymbol{v}^\mathsf{T} \boldsymbol{x}_i^k; \qquad s_k = \sum_{i=1}^{N_k} (\boldsymbol{v}^\mathsf{T} \boldsymbol{x}_i^k - \mu_k)^2.$$

Show that μ_k and s_k can be expressed knowing only class means $\bar{\boldsymbol{x}}_k$ and class covariance matrices Σ_k.

13 Decision Trees

Problem 13.1 (Information Gain maximization)
Consider decision tree for deciding whether or not to go to the cinema. The attributes of training data are: (1) movie length (short/long), (2) country of origin (American/other), (3) the weather (nice/rainy). Find the first splitting atribute by maximization of the information gain (IG). Consider the following training data (note that attributes are all binary, target decision is also binary.)

Sample index	Short	American	Nice weather	Decision: Go?
1	1	1	0	1
2	1	0	0	1
3	1	1	1	0
4	0	1	0	1
5	1	0	1	0
6	0	0	0	0
7	1	1	1	0
8	1	0	1	1

14 Answers

Problem 0.1

$$\begin{aligned}\frac{P(B|A)P(A)}{P(B)} &= \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A|B)P(B)}{P(B)} \\ &= P(A|B)\end{aligned}$$

Problem 0.2

a) $p_K(\text{rain}) = 0.27$, $p_X(1) = 0.4$.
b) $P(x \leq 2|\text{rain}) \approx 0.52$.

Problem 0.3

Given the info you have, complete the table of joint probabilities $P_{C,T}$ and the marginal probabilities P_T.
a) $P_{C|T}(y|+) = \frac{P_{C,T}(y,+)}{P_T(+)} \approx 3.1\%$.
b) $P_{C|T}(n|-) = \frac{P_{C,T}(n,-)}{P_T(-)} \approx 99.9\%$.

Problem 1.1

a) $0.12/0.40 = 30\%$

b) $1 * (0.02/0.27) + 2 * (0.12/0.27) + 3 * (0.09/0.27) + 4 * (0.04/0.27) \approx 2.56$

c) $10 * (0.12/0.40) + (-2) * (0.28/0.40) = 1.6$

d) $10 * 0.27 + (-2) * 0.73 = 1.24$

e) $0 * 0.27 + 5 * 0.73 = 3.65$

f) $q^*(1) = no\,umbrella$ (partial risk -1.4), $q^*(2) = 100$ (partial risk 1.5), $q^*(3) = umbrella$ (partial risk 2), $q^*(4) = umbrella$ (partial risk 1).

Problem 1.2

a) The optimal decision is $not\ sunny$ (because $0.4 \cdot C < (0.2 + 0.2 + 0.2) \cdot C$).

b) The optimal decision is $sunny$ (because $(0.2 + 0.2 + 0.2) \cdot \frac{C}{2} < 0.4 \cdot C$).

Problem 1.4

a) We observe a sequence x of transmitted over a noisy channel sequence of the true bits k. The per-digit most likely solution [0, 0, 1] would be incorrect with respect to the error correcting code. In fact the only possible sequences that could have been sent (with the error-correcting code) are:

$$\begin{aligned}&[0,0,0]\\&[0,1,1]\\&[1,0,1]\\&[1,1,0].\end{aligned} \tag{18}$$

So there are only 4 possible hidden states. We need to compute and compare their posterior probabilities $p(k|x)$. For the purpose of selecting the best posterior, it is sufficient to compare only the numbers $p(x|k)$ as the hidden states are equiprobable a priori.

state	$p(x\|k)$
$[0,0,0]$	$0.7 \cdot 0.6 \cdot 0.3 = 0.126$
$[0,1,1]$	$0.7 \cdot 0.4 \cdot 0.7 = 0.196$
$[1,0,1]$	$0.3 \cdot 0.6 \cdot 0.7 = 0.126$
$[1,1,0]$	$0.3 \cdot 0.4 \cdot 0.3 = 0.036$

The best option with the error correcting is $[0, 1, 1]$.

b) For this part we gonna need the complete probability $p(k|x)$. It is obtained by renormalizing the four values, i.e., dividing them by their sum, 0.484. The probability of correct decoding is thus $0.196/0.484 \approx 0.405$ and of incorrect decoding respectively $1 - 0.196/0.48 \approx 0.595$. The decision to keep the message has the (partial) risk of $0.595 \cdot 100 \cdot C$ while the decision to request a repeat has the risk of $0.405 \cdot C$ only. We decide to ask for a repeat.

Problem 1.5

Lets first compute the probability densities for the given measurements $x \in \{0, 1\}$ by evaluating the normal distributions $p(x|k)$ (approximately).

	k = 1	k = 2	k = 3
x = 0	0.399	0.199	0.065
x = 1	0.242	0.176	0.121

Then we convert to the joint $p(x, k)$ (multiply by $p(k)$)

	k = 1	k = 2	k = 3
x = 0	0.133	0.066	0.022
x = 1	0.081	0.059	0.040

Finally we divide by $p(x) = p(x,1) + p(x,2) + p(x,3)$ to get the posterior $p(k|x)$.

	k = 1	k = 2	k = 3
x = 0	0.602	0.299	0.099
x = 1	0.450	0.328	0.222

In the first part with loss W_a (0-1 loss), we can simply pick the maximum posterior, resulting in $d = 1$ for both $x = 0$ and $x = 1$.

In the second part with loss W_b we compute the partial risks $\sum_k p(k|x)W(k,d)$.

	d = 1	d = 2	d = 3
x = 0	0.697	1.303	0.901
x = 1	0.878	1.122	0.778

We pick the decision with minimal partial risk: $q^*(x = 0) = 1$ and $q^*(x = 1) = 3$.

Finally, the probability of incorrect decision with W_a and $x = 1$ is $0.299 + 0.099 = 0.398$.

Problem 2.2

$q^*(x) = \begin{cases} 1 & : \quad x \in [0,T] \\ 2 & : \quad otherwise \end{cases}$, where $T = \frac{3}{2\sqrt{5}} - \frac{1}{2} \approx 0.17082$.

Problem 2.3

a)

We define the objective function of the minimax task as

$$\arg\min_{q:X\to Y} \max_{y\in Y} \sum_{x:q(x)\neq y} p(x \mid y), \tag{19}$$

where $Y = \{1, 2, \ldots, N\}$ are classes, X is a set of observations x, $p(x \mid y)$ are conditionals that are known $\forall y \in Y$, and $q : X \to Y$ is a strategy.

b)

For a 2-class 2-decision Minimax problem there is always only one likelihood threshold. This comes from the fact that the decision $q^*(x) = d^*$ is the solution of the system of inequalities

$$\gamma(x)c_1(d^*) + c_2(d^*) \leq \gamma(x)c_1(d) + c_2(d), \qquad d \in D \setminus \{d^*\}, \tag{20}$$

where c_1, c_2 are constants, and $\gamma(x)$ is the likelihood ratio. The system is linear with respect to the likelihood ratio. Therefore, in case of two decisions, e.g. $D = \{1, 2\}$, there will be only one threshold.

c)

First, we plot the probability distributions and the likelihood ratio:

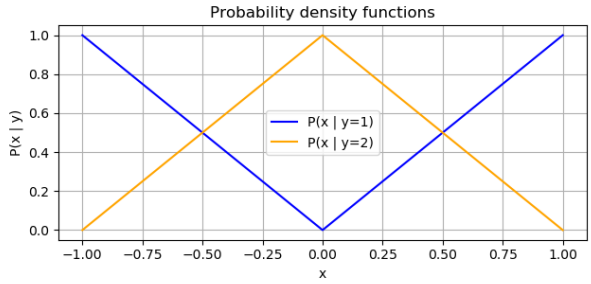


Fig. 1.a: Probability density functions.

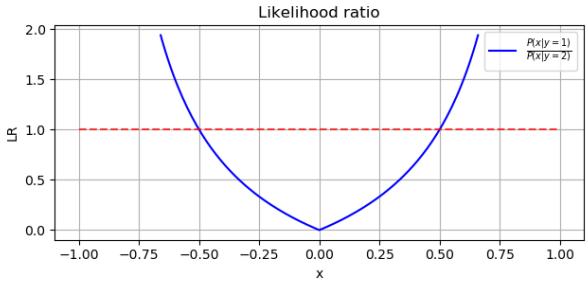


Fig. 1.b: Likelihood ratio $\gamma(x) = \frac{p(x|y=1)}{p(x|y=2)}$.

The minimax task

$$\arg\min_{q:X\to Y} \max_{y\in Y} \sum_{x:q(x)\neq y} p(x \mid y), \tag{21}$$

can be rewritten as

$$\arg\min_{q(x)} \max \left\{ \int_{X_2} p(x \mid y=1)\,dx, \int_{X_1} p(x \mid y=2)\,dx \right\} \tag{22}$$

where

$$X_1 \subseteq X, x \in X_1 : q(x) = 1, \tag{23}$$
$$X_2 \subseteq X, x \in X_2 : q(x) = 2, \tag{24}$$
$$X_1 \cup X_2 = X, \tag{25}$$
$$X_1 \cap X_2 = \emptyset. \tag{26}$$

We are looking for $t \in \langle -1, 1\rangle$, where

$$X_1 = \langle -1, -t\rangle \cup \langle t, 1\rangle, \tag{27}$$
$$X_2 = \langle -t, t\rangle. \tag{28}$$

Therefore the task is now

$$\arg\min_{q(x)} \max \left\{ \int_{-t}^{t} p(x \mid y=1)\,dx, \int_{-1}^{-t} p(x \mid y=2)\,dx + \int_{t}^{1} p(x \mid y=2)\,dx \right\} \tag{29}$$

and thanks to the likelihood ratio $\gamma(x)$ being symmetrical around $x = 0$

$$= \arg\min_{q(x)} \max \left\{ \int_{-t}^{0} p(x \mid y=1)\,dx, \int_{-1}^{-t} p(x \mid y=2) \right\}. \tag{30}$$

If exists $q(x)$ such that

$$\int_{-t}^{0} p(x \mid y=1)\,dx = \int_{-1}^{-t} p(x \mid y=2)\,dx \tag{31}$$

then $-t, t$ are points where $q(x)$ changes and $q(x)$ is the optimal strategy.

$$\int_{-t}^{0} -x\,dx = \int_{-1}^{-t} 1 + x\,dx \tag{32}$$

$$-\left[\frac{x^2}{2}\right]_{-t}^{0} = \left[x + \frac{x^2}{2}\right]_{-1}^{-t} \tag{33}$$

$$\frac{t^2}{2} = \frac{t^2}{2} - t + 1 - \frac{1}{2} \tag{34}$$

$$t = \frac{1}{2} \tag{35}$$

$$X_1 = \left\langle -1, -\frac{1}{2} \right\rangle \cup \left\langle \frac{1}{2}, 1 \right\rangle \tag{36}$$

$$X_2 = \left\langle -\frac{1}{2}, \frac{1}{2} \right\rangle \tag{37}$$

Now, we compute the likelihood ratio threshold θ. Since we compute the threshold by substituting to the functions for $x \in (-1, 0)$, we substitute the x in both $p(x \mid y = 1)$ and $p(x \mid y = 2)$ for the value of $-t = -\frac{1}{2}$.

$$\theta = \frac{-x}{1+x} = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1. \tag{38}$$

Finally, we define the optimal strategy

$$q(x) = \begin{cases} 1, & \text{if } \gamma(x) \geq \theta \\ 2, & \text{else.} \end{cases} \tag{39}$$

Problem 3.1

a)

We define the probability model as

$$p(H) = \pi, \tag{40}$$

$$p(T) = 1 - \pi, \tag{41}$$

where $p(H)$ is the probability of head and $p(T)$ is the probability of tail. Now lets consider the given sequence $S = (H, H, H, T, T)$ of $N = 5$ draws of $k = 3$ heads. The probability (likelihood) of this sequence is

$$p(S|\pi) = \pi^k (1 - \pi)^{N-k}\,. \tag{42}$$

From the sequence we can derive the ML estimate of π as

$$\hat{\pi} = \arg\max_{\pi} p(S|\pi) = \arg\max_{\pi} \pi^k (1 - \pi)^{N-k}\,. \tag{43}$$

To find the $\arg\max$ we compute the derivative of the expression and set it equal to zero

$$\frac{\partial p}{\partial \pi} = k\pi^{k-1}(1 - \pi)^{N-k} - (N - k)\pi^k (1 - \pi)^{N-k-1} = 0\,. \tag{44}$$

Solving this equation gets

$$\hat{\pi} = \frac{k}{N}\,, \tag{45}$$

and in the particular case of our sequence S we get $\pi = 3/5$.

b)

Let $P(H)$ be

$$p(H) = \frac{1}{1+e^{-\eta}}\,, \tag{46}$$

then for $P(T)$ we have

$$p(T) = 1 - p(H) = \frac{e^{-\eta}}{1+e^{-\eta}}\,. \tag{47}$$

For some sequence S of N draws with k heads, we have the likelihood

$$p(S|\eta) = p(H)^k p(T)^{N-k} = \left(\frac{1}{1+e^{-\eta}}\right)^k \left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{N-k} \tag{48}$$

Now we need to find the ML estimate of η:

$$\hat{\eta} = \arg\max_{\eta} p(S|\eta) = \arg\max_{\eta} \left(\frac{1}{1+e^{-\eta}}\right)^k \left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{N-k} \tag{49}$$

To simplify the derivative we can use logarithm of the equation

$$\begin{aligned} \hat{\eta} &= \arg\max_{\eta} \log(p(S|\eta)) && (50)\\ &= \arg\max_{\eta} -k\log(1+e^{-\eta}) - (N-k)\log(1+e^{-\eta}) - (N-k)\eta && (51) \end{aligned}$$

Now we again use derivative to get the optimal $\hat{\eta}$

$$\frac{\partial \log p(S|\eta)}{\partial \eta} = \frac{ke^{-\eta}}{1+e^{-\eta}} + \frac{(N-k)e^{-\eta}}{1+e^{-\eta}} + (N-k) = \frac{Ne^{-\eta}}{1+e^{-\eta}} - (N-k) = 0\,, \tag{52}$$

which leads to the following

$$\begin{aligned} Ne^{-\eta} &= Ne^{-\eta} - ke^{-\eta} + N - k && (53)\\ e^{-\eta} &= \frac{N+k}{k} && (54)\\ \eta &= -\log\left(\frac{N-k}{k}\right) = \log\left(\frac{k}{N-k}\right)\,. && (55) \end{aligned}$$

So for our sequence (H, H, H, T, T), we get

$$\hat{\eta} = \log\frac{3}{2}. \tag{56}$$

Problem 3.2

a)

We need to define probability of some random sequence of draws. Draw with replacement means that when you draw a sock you put it back in the box and mix it with the others in the

box. When we draw N socks where R are red, we can write the probability of this sequence S as

$$p(S|r) = \binom{N}{R} r^R (1-r)^{(N-R)} \,. \tag{57}$$

We can see that the distribution is binomial, that is because we need to count every combination with exactly R red socks and the total amount of those combination is $\binom{N}{R}$.

b)

So the probability of drawing 2 red socks $(R = 2)$ will be

$$p(R = 2|r) = \binom{2}{2} r^2 = r^2 \,. \tag{58}$$

c)

The likelihood of drawing $R = 2$ red socks of $N = 10$ draws is

$$p(S|r) = \binom{N}{R} r^R (1-r)^{N-R} = \binom{10}{2} r^2 (1-r)^8. \tag{59}$$

d)

To find the ML estimate of r we will compute the derivative of the likelihood and set it to zero

$$\frac{\partial p(S|r)}{\partial r} = \binom{10}{2} (2r(1-r)^8 - 8r^2(1-r)^7) = 0 \tag{60}$$

We now divide the equation with $r(1-r)^7$ and get

$$\binom{10}{2} (2(1-r) - 8r) = 0 \tag{61}$$

$$r = \frac{2}{10} = 0.2 \,. \tag{62}$$

Problem 3.3

Since x_i are i.i.d., the log-likelihood of observing $x_1, \ldots x_N$ is given by the sum

$$l(\mu, \Sigma) = \sum_{i=1}^{N} \Big(\log \det(\Sigma)^{-\frac{1}{2}} - \tfrac{1}{2}(x_i - \mu)^\mathsf{T} \Sigma^{-1} (x_i - \mu) \Big) + \textit{const.} \tag{63}$$

a)

We find the optimal μ as a critical point of $l(\mu, \Sigma)$ in μ:

$$0 = \frac{dl}{d\mu} = \sum_{i=1}(x_i - \mu)^\mathsf{T}\Sigma^{-1}, \tag{64}$$

$$\sum_{i=1}\mu^\mathsf{T}\Sigma^{-1} = \sum_{i=1} x_i\Sigma^{-1}, \qquad \text{(open brackets)}$$

$$\Sigma^{-1}N\mu = \Sigma^{-1}\sum_{i=1} x_i, \qquad \text{(transpose)}$$

$$N\mu = \sum_{i=1} x_i. \qquad \text{(multiply with } \Sigma\text{)}$$

Therefore $\mu = \frac{1}{N}\sum_{i=1} x_i$ – the sample mean.

b)

Instead of finding optimal covariance matrix Σ, we will find optimal precision matrix $\Lambda = \Sigma^{-1}$. It is a one-to-one parameterization, which simplifies the differentiation. We have

$$l(\mu, \Lambda) = \frac{N}{2}\log\det(\Lambda) - \tfrac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^\mathsf{T}\Lambda(x_i - \mu)\Big) + \textit{const}. \tag{65a}$$

To differentiate $\det(\Lambda)$ we use the identity $\frac{d}{d\Lambda}\det(\Lambda) = \det(\Lambda)\Lambda^{-1}$. To differentiate the quadratic form $a^\mathsf{T}\Lambda a$ i Λ we can inspect in components that

$$\frac{d}{d\Lambda_{ij}}\sum_{i'} a_{i'}\sum_{j'}\Lambda_{i'j'}a_{j'} = a_i a_j \tag{65b}$$

and therefore $\frac{d}{d\Lambda}a^\mathsf{T}\Lambda a = aa^\mathsf{T}$, the outer product. We thus obtain

$$0 = \frac{dl}{d\Lambda} = N\frac{1}{\det(\Lambda)}\det(\Lambda)\Lambda^{-1} - \sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\mathsf{T} = 0, \tag{65c}$$

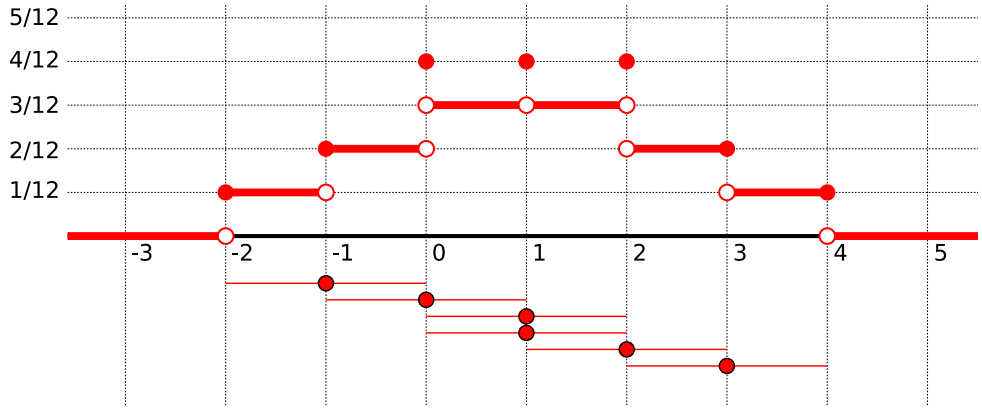
$$N\Lambda^{-1} = \sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\mathsf{T}, \tag{65d}$$

$$\Sigma = \Lambda^{-1} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\mathsf{T}, \tag{65e}$$

i.e. Σ is the sample covariance matrix of the data.

Problem 4.4

Top: the kernel density estimate, bottom: copies of the kernels placed at data points. Note that some isolated points jump up because the kernels are uniform over closed intervals. However, these points are not important for the resulting distribution because they have a zero measure.



Problem 5.2

a) Show that $\sigma(-z) = 1 - \sigma(z)$:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{66}$$

$$\sigma(-z) = \frac{1}{1+e^{z}} \cdot \frac{e^{-z}}{e^{-z}} = \frac{e^{-z}+1-1}{1+e^{-z}} = 1 - \frac{1}{1+e^{-z}} = 1 - \sigma(z) \tag{67}$$

b) Show that $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1-\sigma(z))$:

$$\begin{aligned}
\frac{\partial}{\partial z}\sigma(z) &= -\frac{1}{(1+e^{-z})^2} \cdot e^{-z} \cdot (-1) && (68)\\
&= e^{-z} \cdot \frac{1}{(1+e^{-z})^2} && (69)\\
&= \frac{e^{-z}}{1+e^{-z}} \cdot \frac{1}{1+e^{-z}} && (70)\\
&= \frac{e^{-z}}{e^{-z}} \cdot \frac{1}{1-e^{z}} \cdot \frac{1}{1-e^{-z}} && (71)\\
&= \sigma(-z)\sigma(z) && (72)\\
&= (1-\sigma(z))\sigma(z) && (73)
\end{aligned}$$

c) Compute $\frac{\partial}{\partial z}\log\sigma(z)$ using b):

$$\frac{\partial}{\partial z}\log\sigma(z) = \frac{1}{\sigma(z)}\frac{\partial\sigma(z)}{\partial z} = \frac{1}{\sigma(z)}(1-\sigma(z))\sigma(z) = 1-\sigma(z) = \sigma(-z). \tag{74}$$

d) Compute $\frac{\partial}{\partial \mathbf{x}}\log\sigma(\mathbf{w}^T\mathbf{x})$ using c):

Letting $z = \mathbf{w}^T\mathbf{x}$ and using the chain rule of derivative we get:

$$\frac{\partial}{\partial \mathbf{x}}\log(\sigma(z)) = \frac{\partial\log(\sigma(z))}{\partial z}\frac{\partial z}{\partial \mathbf{x}} = \sigma(-z)\frac{\partial\,\mathbf{w}^T\mathbf{x}}{\partial \mathbf{x}} = \sigma(-\mathbf{w}^T\mathbf{x})\mathbf{w}^\mathsf{T}. \tag{75}$$

e) Show that $-\log\sigma(z)$ is convex (using monotonicity of the first derivative or non-negativity of the second derivative):

We decided to show convexity using non-negativity of the second derivative. Using c) we get

first derivative:

$$\frac{\partial(-\log\sigma(z))}{\partial z} = -\sigma(-z) = -\frac{1}{1+e^z} \tag{76}$$

And using equation above, a) and b) we get

$$\frac{\partial^2(-\log\sigma(z))}{\partial z^2} = -\frac{\partial\sigma(-z)}{\partial z} = \frac{\partial(\sigma(z)-1)}{\partial z} = \sigma(z)(1-\sigma(z)) = \frac{e^{-z}}{(1+e^{-z})^2} \tag{77}$$

The second derivative is positive, because both the numerator and the denominator are positive for any real z. Therefore the function is convex.

Problem 6.1

a) The distance is $d = \frac{|a^\mathsf{T}x+b|}{\|a\|}$.

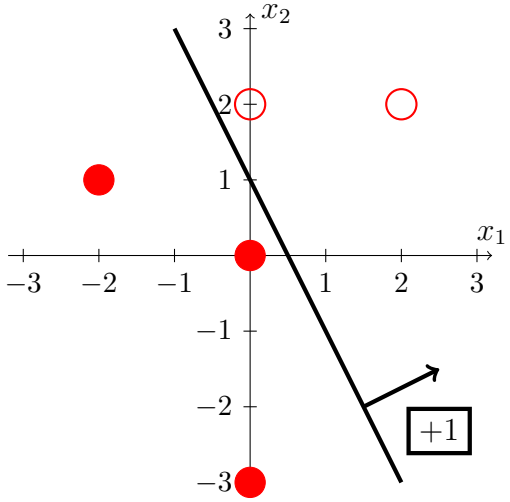
Problem 6.4

First, we augment the training data by appending 1 to each $\boldsymbol{x}_i$ and multiplying by the class k_i, resulting in $\bar{T} = \{(2,-1,-1),(0,0,-1),(0,2,1),(0,3,-1),(2,2,1)\}$. After this change from $\boldsymbol{x}_i$ to $\bar{\boldsymbol{x}}_i$, we will be looking for $\bar{\boldsymbol{w}} \in \mathbb{R}^3$ satisfying $\bar{\boldsymbol{w}}^\mathsf{T}\bar{\boldsymbol{x}}_i > 0$ for all i. The new $\bar{\boldsymbol{w}}$ vector is composed from the original $\boldsymbol{w}$ and the bias term b, in particular $\bar{\boldsymbol{w}} = (w_1, w_2, b)$.

The computation is carried out in the following table.

Iteration	Weight vector $\bar{\boldsymbol{w}}$	Misclassified example: $\bar{\boldsymbol{w}}^\mathsf{T}\bar{\boldsymbol{x}} \leq 0$
0	$(0,0,0)$	$(0,0,0)^\mathsf{T}(2,-1,-1) = 0 \leq 0$
1	$(0,0,0)+(2,-1,-1) = (2,-1,-1)$	$(2,-1,-1)^\mathsf{T}(0,2,1) = -3 \leq 0$
2	$(2,-1,-1)+(0,2,1) = (2,1,0)$	$(2,1,0)^\mathsf{T}(0,0,-1) = 0 \leq 0$
3	$(2,1,0)+(0,0,-1) = (2,1,-1)$	none

The Perceptron algorithm terminated after 3 steps with $\boldsymbol{w} = (2,1)$ and $b = -1$, corresponding to the visualization below. Note that picking misclassified examples differently in each step would result in different iterates with possibly different number of steps and the final solution.



Problem 7.2

a)

The unified form of the inequalities is

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \epsilon$$

for all data points $(\mathbf{x}_i, y_i)$.

b)

We show one possible solution, but alternative solutions are possible.

We denote the two hyper-planes as

$$\alpha_1 : \mathbf{w}^T\mathbf{x}_i + b = \epsilon \tag{78}$$
$$\alpha_2 : \mathbf{w}^T\mathbf{x}_i + b = -\epsilon\,. \tag{79}$$

If we pick two points $\mathbf{x} \in \alpha_1$ and $\mathbf{y} \in \alpha_2$ then the projection of the vector $(\mathbf{x} - \mathbf{y})$ to the hyper-plane normal $\mathbf{w}$ is computed as their dot product. If we compute the dot product with the normal vector normalised to unit length, the length of the resulting projection is the distance of the two hyper-planes. In particular

$$d = \left\langle (\mathbf{x} - \mathbf{y}), \frac{\mathbf{w}}{||\mathbf{w}||} \right\rangle = \frac{1}{||\mathbf{w}||}(\langle \mathbf{x}, \mathbf{w}\rangle - \langle \mathbf{y}, \mathbf{w}\rangle) \tag{80}$$

To simplify this, we need to express the formulas for $\mathbf{x}$ and $\mathbf{y}$. We will multiply both hyper-plane equations with vector $\mathbf{w}$ from left

$$\mathbf{w}\mathbf{w}^T\mathbf{x} + \mathbf{w}b = \mathbf{w}\epsilon \tag{81}$$
$$\mathbf{w}\mathbf{w}^T\mathbf{y} + \mathbf{w}b = -\mathbf{w}\epsilon\,. \tag{82}$$

Knowing that $\mathbf{w}^T\mathbf{w} = ||\mathbf{w}||^2$ we can express $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x} = \frac{\mathbf{w}}{||\mathbf{w}||^2}(\epsilon - b) \tag{83}$$
$$\mathbf{y} = -\frac{\mathbf{w}}{||\mathbf{w}||^2}(\epsilon + b) \tag{84}$$

Then we substitute these formulas back into the expression for d

$$d = \frac{1}{||\mathbf{w}||}\left(\left\langle \frac{\mathbf{w}(\epsilon - b)}{||\mathbf{w}||^2}, \mathbf{w} \right\rangle - \left\langle \frac{-\mathbf{w}(\epsilon + b)}{||\mathbf{w}||^2}, \mathbf{w} \right\rangle\right) \tag{85}$$
$$= \frac{\epsilon - b}{||\mathbf{w}||^3}\langle \mathbf{w}, \mathbf{w}\rangle + \frac{\epsilon + b}{||\mathbf{w}||^3}\langle \mathbf{w}, \mathbf{w}\rangle \tag{86}$$
$$= \frac{2\epsilon}{||\mathbf{w}||}\,. \tag{87}$$

The distance d is called margin.

c)

See the lecture slides. Start with the margin derived in b), solve the scale ambiguity (setting $\epsilon = 1$), change the maximisation into minimisation and go from a linear to a quadratic form.

Problem 7.6

To make the notation clear we will be using $\mathbf{x}, \mathbf{y}$ *instead of* $\mathbf{x}, \mathbf{x}'$.

a)

For given $\phi(\mathbf{x})$ the corresponding kernel function $k(\mathbf{x}, \mathbf{y})$ is equal to dot product $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ so

$$k(\mathbf{x}, \mathbf{y}) = 1 + x_1 y_1 + x_2 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 + x_1 y_1 x_2 y_2 \tag{88}$$

b)

Expanding the kernel function expression we get

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= (1 + x_1 y_1 + x_2 y_2)^2 & (89) \\ &= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 y_1 x_2 y_2 + + x_1^2 y_1^2 + x_2^2 y_2^2. & (90) \end{aligned}$$

Since all equation elements contain both $x_{1|2}$ and $y_{1|2}$ we can easily see that the mapping $\phi(\mathbf{x})$ corresponding to $k(\phi(\mathbf{x}), \phi(\mathbf{y})) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ is

$$\phi(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2)\,. \tag{91}$$

c)

We can derive the proper mapping the same way as we did in **b)**, first we need to expand the kernel formula for a given D

$$k(\mathbf{x}, \mathbf{y}) = \left(1 + \sum_{i=1}^{D} x_i y_i\right)^d \tag{92}$$

To expand this expression we need multinomial theorem, and after the expansion we will do the same thing as in section **b)**. But we can see that it is much faster to compute just the dot product $\langle \mathbf{x}, \mathbf{y} \rangle$, add one and compute the power of d.

Problem 13.1

First split by the *Nice weather* attribute. $IG(niceweather) = \frac{3\log(3)}{4} - 1 \approx 0.189$, while $IG(short) = IG(american) = 0$.