Learning.
Linear Methods for Regression and Classification.

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Decision strategy design

Using an observation $x \in X$ of an object of interest with a hidden state $k \in K$, we should design a decision strategy $q : X \to D$ which would be optimal with respect to certain criterion.

Bayesian decision theory requires complete statistical information $p_{XK}(x,k)$ of the object of interest to be known, and a suitable penalty function $W : K \times D \to \mathcal{R}$ must be provided.

Non-Bayesian decision theory studies tasks for which some of the above information is not available.

In practical applications, typically, none of the probabilities are known! The designer is only provided with the **training (multi)set** $T = \{(x_1,k_1),(x_2,k_2),\ldots,(x_l,k_l)\}$ of examples.

- It is simpler to provide good examples than to gain complete or partial statistical model, build general theories, or create explicit descriptions of concepts (hidden states).
- The training (multi)set is an unbiased *finite* sample from p_{XK} (at best); the found decision strategy is just an approximation of the Bayes strategy.

When do we need to use learning?

- When knowledge about the recognized object is insufficient to solve the PR task.
- Most often, we have insufficient knowledge about $p_{X|K}(x|k)$.

Types of feedback in learning

Supervised learning:

- A training multi-set of examples is available. Correct answers (hidden state, class, the quantity we want to predict) are *known* for all observations.
  - **Classification:** the answers (the output variable of the model) are nominal, i.e. the value specifies a class ID. (predict spam/ham based on email contents, predict $0/1/\ldots/9$ based on the image of the number, etc.)
  - **Regression:** the answers (the output variable of the model) are quantitative, often continuous (predict temperature in Prague based on date and time, predict height of a person based on weight and gender, etc.)

Unsupervised learning:

- A training multi-set of examples is available. Correct answers are *not known*, they must be sought in data itself $\Rightarrow$ data analysis.

Semisupervised learning:

- A training multi-set of examples is available. Correct answers are *known only for a subset* of the training set.

Reinforcement learning:

- A training multi-set of examples is *not available*. Correct answers, or rather rewards for good decisions in the past, *are given occasionally after decisions are taken*.

Learning as parameter estimation

1. **Assume** $p_{XK}(x,k) = p_{XK|\Theta}(x,k|\boldsymbol{\theta})$ has a particular form (e.g. Gaussian, mixture of Gaussians, piece-wise constant) with a small number of parameters Θ.
2. **Estimate** the values of parameters $\boldsymbol{\Theta}$ using the training set T.
3. **Solve** the classifier design problem as if the estimated $\hat{p}_{XK}(x,k) = p_{XK|\Theta}(x,k|\hat{\boldsymbol{\theta}})$ was the true (and unknown) $p_{XK}(x,k)$.

Pros and cons:

- If the true $p_{XK}(x,k)$ does not have the assumed form, the resulting strategy $q'(x)$ can be arbitrarilly bad, even if the training set size $|T|$ approaches infinity.
- Implementation is often straightforward, especially if the parameters Θ_k are assumed to be independent for each class (**naive bayes classifier**).

Learning as optimal strategy selection

- Choose a class Q of strategies $q_\Theta : X \to D$. The class Q is usually given as a set of parametrized strategies of the same kind.
- **Learning** then amounts to **selecting a particular strategy** q_{θ^*} **from the *a priori* known set** Q using the information provided as training set T.
  - Natural criterion for the selection of one particular strategy is the risk $R(q_\Theta)$, but it cannot be computed because $p_{XK}(x,k)$ is unknown.
  - The strategy $q_{\theta^*} \in Q$ is chosen by minimizing some other surrogate criterion on the training set which approximates $R(q_\Theta)$.
  - The choice of the surrogate criterion determines the *learning paradigm*.

Several surrogate criteria

All the following surrogate criteria can be computed using the training data T.

Learning as parameter estimation

- according to the **maximum likelihood**.
  - The likelihood of an instance of the parameters $\boldsymbol{\theta} = (\theta_k : k \in K)$ is the probability of T given $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = p(T|\boldsymbol{\theta}) = \prod_{(x_i,k_i)\in T} p_{XK|\Theta}(x_i, k_i|\boldsymbol{\theta}) = \prod_{(x_i,k_i)\in T} p_K(k_i) p_{X|K,\Theta}(x|k, \theta_k)$$

  - Learning then means to find $\boldsymbol{\theta}^*$ that maximizes the probability of T:

$$\boldsymbol{\theta}^* = (\theta_k^* : k \in K) = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

    which can be decomposed to

$$\theta_k^* = \arg\max_{\theta_k} \sum_{x\in X} \alpha(x,k) \log p_{X|K}(x|k, \theta_k),$$

    where $\alpha(x,k)$ is the frequency of the pair (x,k) in T (i.e. T is multiset).
  - The recognition is then performed according to $q_{\theta^*}(x) = q_\Theta(x, \theta^*)$.
- according to a **non-random training set**.
  - When random examples are not easy to obtain, e.g. in recognition of images.
  - T is carefully crafted by the designer:
    - it should cover the whole recognized domain
    - the examples should be typical ("quite probable") prototypes
  - Let $T(k), k \in K$, be a subset of the training set T with examples for state k. Then for all $k \in K$

$$\theta_k^* = \arg\max_{\theta_k} \min_{x\in T(k)} p_{X|K,\Theta}(x|k, \Theta_k)$$

  - Note that the $\boldsymbol{\theta}^*$ does not depend on the frequencies of (x,k) in T (i.e. T is a set).

Learning as optimal strategy selection

- by **minimization of the empirical risk**.
  - The set Q of parametrized strategies $q_\Theta : X \to D$, penalty function $W : K \times D \to \mathcal{R}$.
  - The quality of each strategy $q_\theta \in Q$ (i.e. the quality of each parameter set $\boldsymbol{\theta}$) could be described by the risk

$$R(\boldsymbol{\theta}) = R(q_{\boldsymbol{\theta}}) = \sum_{k\in K} \sum_{x\in X} p_{XK}(x,k) W(k, q_\Theta(x, \boldsymbol{\theta})),$$

    but p_{XK} is unknown.
  - We thus use the **empirical risk** R_{emp} (training set error):

$$R_{\text{emp}}(\boldsymbol{\theta}) = R_{\text{emp}}(q_{\boldsymbol{\theta}}) = \frac{1}{|T|} \sum_{(x_i,k_i)\in T} W(k_i, q_\Theta(x_i, \boldsymbol{\theta})).$$

  - Strategy $q_{\theta^*}(x) = q_\Theta(x, \boldsymbol{\theta}^*)$ is used where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} R_{\text{emp}}(\boldsymbol{\theta})$.
  - Examples: Perceptron, neural networks (backprop.), classification trees, ...
- by **minimization of the structural risk**.
  - Based on Vapnik-Chervonenkis theory
  - Examples: Optimal separating hyperplane, support vector machine (SVM)

Summary

Learning:

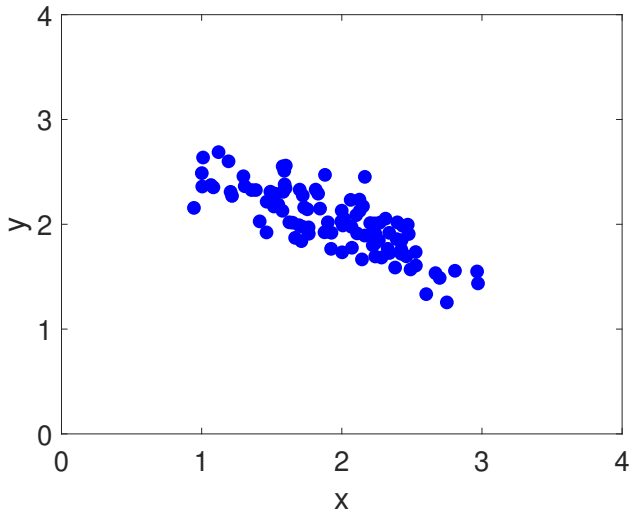
- Needed when we do not have sufficient statistical info for recognition without learning.
- There are several types of learning differing in the types of information the learning process can use.

Approaches to learning:

- Assume p_{XK} has a certain form and use T to estimate its parameters.
- Assume the right strategy is in a particular set and use T to choose it.
- There are several learning paradigms depending on the choice of criterion used instead of Bayesian risk.

Question: Line fitting

We would like to fit a line of the form $\hat{y} = w_0 + w_1 x$ to the following data:



The parameters of a line with a good fit will likely be

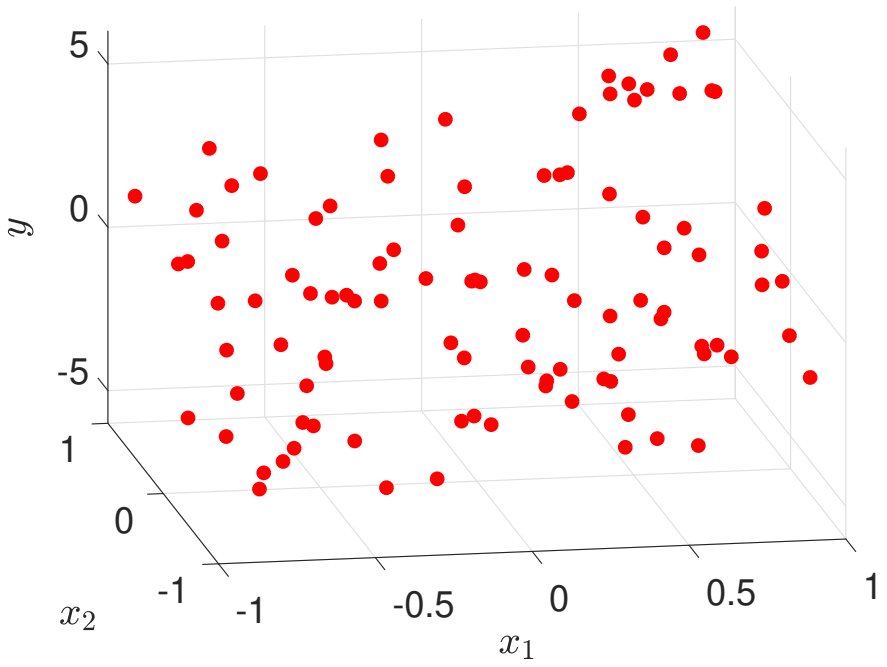
A $w_0 = -1, w_1 = -2$

B $w_0 = -\frac{1}{2}, w_1 = 1$

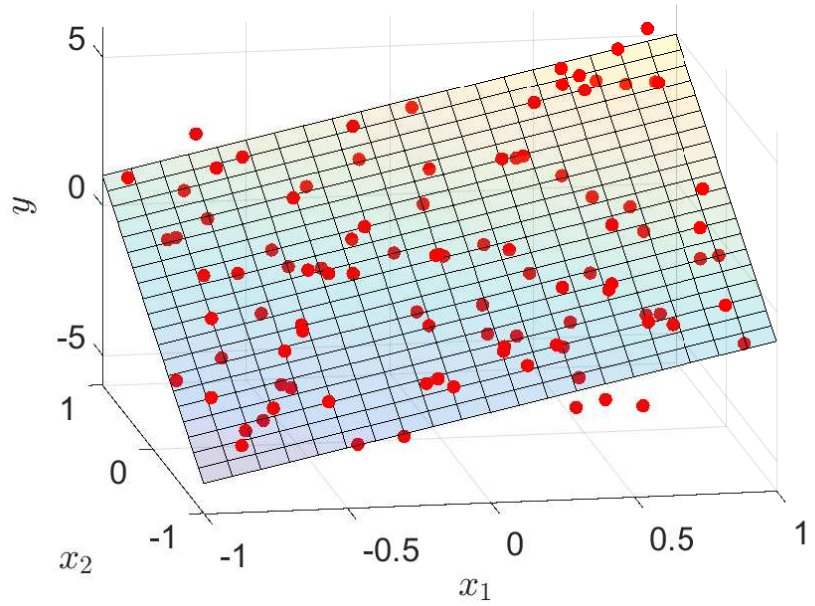
C $w_0 = 3, w_1 = -\frac{1}{2}$

D $w_0 = 2, w_1 = \frac{1}{3}$

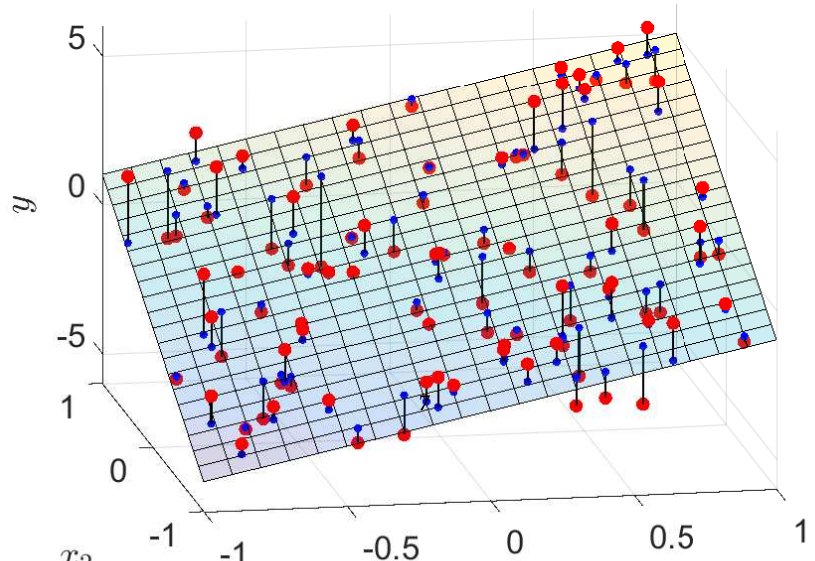
Linear regression: Illustration



Given a dataset of input vectors $x^{(i)}$ and the respective values of output variable $y^{(i)}$ ...



...we would like to find a linear model of this dataset ...



Linear regression

Regression task is a supervised learning task, i.e.

- a training (multi)set $T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(|T|)}, y^{(|T|)})\}$ is available, where
- the labels $y^{(i)}$ are *quantitative*, often *continuous* (as opposed to classification tasks where $y^{(i)}$ are nominal).
- Its purpose is to model the relationship between independent variables (inputs) $\boldsymbol{x} = (x_1, \ldots, x_D)$ and the dependent variable (output) y.

Linear regression is a particular regression model which assumes (and learns) linear relationship between the inputs and the output:

$$\widehat{y} = h(\boldsymbol{x}) = w_0 + w_1 x_1 + \ldots + w_D x_D = w_0 + \langle \boldsymbol{w}, \boldsymbol{x} \rangle = w_0 + \boldsymbol{x}\boldsymbol{w}^T,$$

where

- $\widehat{y}$ is the model *prediction* (*estimate* of the true value y),
- $h(\boldsymbol{x})$ is the linear model (a *hypothesis*),
- $w_0, \ldots, w_D$ are the coefficients of the linear function, w_0 is the *bias*, organized in a row vector $\boldsymbol{w}$,
- $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ is a *dot product* of vectors $\boldsymbol{w}$ and $\boldsymbol{x}$ (scalar product),
- which can be also computed as a matrix product $\boldsymbol{x}\boldsymbol{w}^T$ if $\boldsymbol{w}$ and $\boldsymbol{x}$ are row vectors.

Notation remarks

Homogeneous coordinates: If we add "1" as the first element of $\boldsymbol{x}$ so that $\boldsymbol{x} = (1, x_1, \ldots, x_D)$, then we can write the linear model in an even simpler form (without the explicit bias term):

$$\widehat{y} = h(\boldsymbol{x}) = w_0 \cdot 1 + w_1 x_1 + \ldots + w_D x_D = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \boldsymbol{x}\boldsymbol{w}^T.$$

Matrix notation: If we organize the data into matrix $\boldsymbol{X}$ and vector $\boldsymbol{y}$, such that

$$\boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{x}^{(1)} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}^{(|T|)} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(|T|)} \end{pmatrix},$$

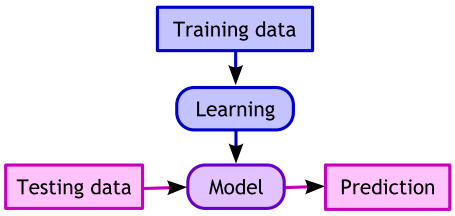
and similarly with $\widehat{\boldsymbol{y}}$, then we can write a batch computation of predictions for all data in $\boldsymbol{X}$ as

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}^T.$$

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) and
2. application (testing, making predictions).



The model h can be viewed as a function of 2 variables: $h(\boldsymbol{x}, \boldsymbol{w})$.

Model application: If the model is given ($\boldsymbol{w}$ is fixed), we can manipulate $\boldsymbol{x}$ to make predictions:

$$\hat{y} = h(\boldsymbol{x}, \boldsymbol{w}) = h_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: If the data is given (T is fixed), we can manipulate the model parameters $\boldsymbol{w}$ to fit the model to the data:

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} J(\boldsymbol{w}, T).$$

How to train the model?

Simple (univariate) linear regression

Simple (univariate) regression deals with cases where $\boldsymbol{x}^{(i)} = x^{(i)}$, i.e. the examples are described by a single feature (they are 1-dimensional).

Fitting a line to data:

- find parameters w_0, w_1 of a linear model $\hat{y} = w_0 + w_1 x$
- given a training (multi)set $T = \{(x^{(i)}, y^{(i)})\}_{i=1}^{|T|}$.

How to fit a line depending on the number of training examples $|T|$:

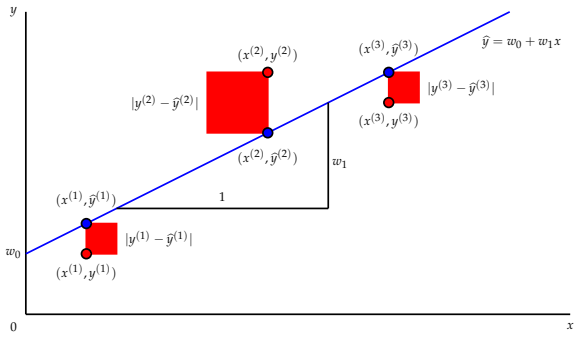
- Given a single example (1 equation, 2 parameters)
  $\Rightarrow$ infinitely many linear functions can be fitted.
- Given 2 examples (2 equations, 2 parameters)
  $\Rightarrow$ exactly 1 linear function can be fitted.
- Given 3 or more examples ($>$ 2 equations, 2 parameters)
  $\Rightarrow$ no line can be fitted with zero error
  $\Rightarrow$ a line which minimizes the "size" of error $y - \hat{y}$ can be fitted:

$$\boldsymbol{w}^* = (w_0^*, w_1^*) = \operatorname*{argmin}_{w_0, w_1} J(w_0, w_1, T).$$

The least squares method

The **least squares method (LSM)** suggests to choose such parameters $\boldsymbol{w}$ which minimize the *mean squared error* (MSE)

$$J_{MSE}(\boldsymbol{w}) = \frac{1}{|T|}\sum_{i=1}^{|T|}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \frac{1}{|T|}\sum_{i=1}^{|T|}\left(y^{(i)} - h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})\right)^2.$$

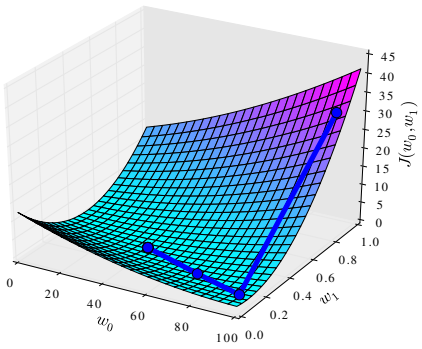


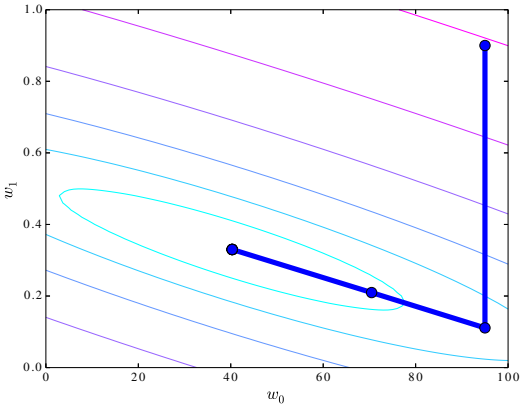
Explicit solution:

$$w_1 = \frac{\sum_{i=1}^{|T|}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{|T|}(x^{(i)} - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

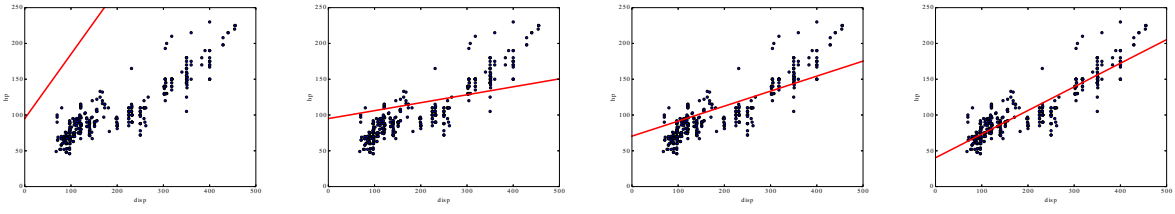
Universal fitting method: minimization of cost function J

The landscape of J in the space of parameters w_0 and w_1:





Gradually better linear models found by an optimization method (BFGS):



Gradient descent algorithm

Given a function $J(w_0, w_1)$ that should be minimized,

- start with a guess of w_0 and w_1 and
- change it, so that $J(w_0, w_1)$ decreases, i.e.
- update our current guess of w_0 and w_1 by taking a step in the direction opposite to the gradient:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \nabla J(w_0, w_1), \text{ i.e.}$$
$$w_d \leftarrow w_d - \alpha \frac{\partial}{\partial w_d} J(w_0, w_1),$$

where all w_is are updated simultaneously and α is a **learning rate** (step size).

For the cost function

$$J(w_0, w_1) = \frac{1}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - h_{\boldsymbol{w}}(x^{(i)}) \right)^2 = \frac{1}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - (w_0 + w_1 x^{(i)}) \right)^2,$$

the gradient can be computed as

$$\frac{\partial}{\partial w_0} J(w_0, w_1) = -\frac{2}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - h_{\boldsymbol{w}}(x^{(i)}) \right) = \frac{2}{|T|} \sum_{i=1}^{|T|} \left(h_{\boldsymbol{w}}(x^{(i)}) - y^{(i)} \right)$$
$$\frac{\partial}{\partial w_1} J(w_0, w_1) = -\frac{2}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - h_{\boldsymbol{w}}(x^{(i)}) \right) x^{(i)} = \frac{2}{|T|} \sum_{i=1}^{|T|} \left(h_{\boldsymbol{w}}(x^{(i)}) - y^{(i)} \right) x^{(i)}$$

Multivariate linear regression

Multivariate linear regression deals with cases where $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_D^{(i)})$, i.e. the examples are described by more than 1 feature (they are D-dimensional).

Model fitting:

- find parameters $\boldsymbol{w} = (w_1, \ldots, w_D)$ of a linear model $\widehat{y} = \boldsymbol{x}\boldsymbol{w}^T$
- given the training (multi)set $T = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{|T|}$.
- The model is a *hyperplane* in the $D+1$-dimensional space.

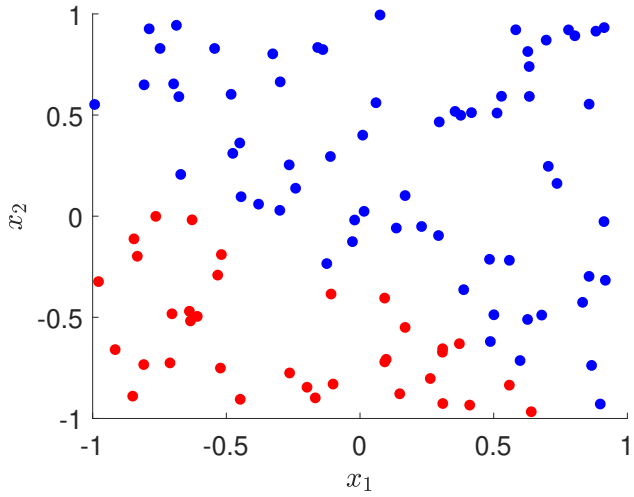
Fitting methods:

1. Numeric optimization of $J(\boldsymbol{w}, T)$:
   - Works as for simple regression, it only searches a space with more dimensions.
   - Sometimes one needs to tune some parameters of the optimization algorithm to work properly (learning rate in gradient descent, etc.).
   - May be slow (many iterations needed), but works even for very large D.
2. **Normal equation:**

$$\boldsymbol{w}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

   - Method to solve for the optimal $\boldsymbol{w}^*$ analytically!
   - No need to choose optimization algorithm parameters.
   - No iterations.
   - Needs to compute $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$, which is $O(D^3)$. Slow, or intractable, for large D.

Question



Intuitively, which of the training data points have the biggest influence on the decision whether a new, unlabeled data point shall be red or blue?

A Those which are closest to data points with the opposite color.

B Those which are farthest from the data points of the opposite color.

C Those which are near the middle of the points with the same color.

D None. All of the data points have the same importance.

Binary classification task (dichotomy)

Let's have the training dataset $T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(|T|)}, y^{(|T|)})$:

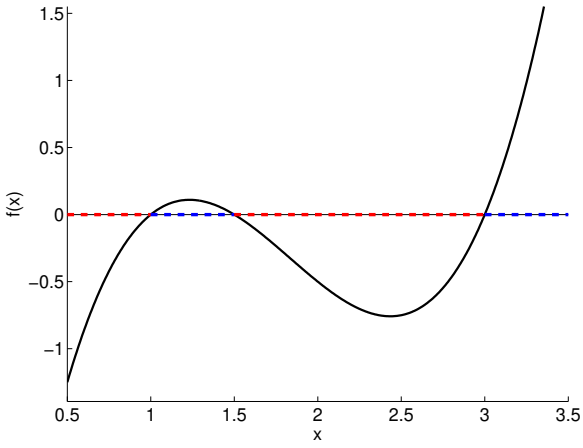
- each example described by a vector $\boldsymbol{x} = (x_1, \ldots, x_D)$,
- labeled with the correct class $y \in \{+1, -1\}$.

Discrimination function $f(x)$:

- It assigns a real number to each observation $\boldsymbol{x}$, may be linear or non-linear.
- For 2 classes, 1 discrimination function is enough.
- It is used to create a **decision rule** (which then assigns a class to an observation):

$$f(\boldsymbol{x}) > 0 \iff \widehat{y} = +1, \text{ and}$$
$$f(\boldsymbol{x}) < 0 \iff \widehat{y} = -1$$

i.e. $\widehat{y} = \text{sign}\,(f(\boldsymbol{x}))$.



- **Decision boundary:** $\{\boldsymbol{x} | f(\boldsymbol{x}) = 0\}$
- **Linear classification:** the decision boundaries must be linear.
- *Learning* then amounts to finding (suitable parameters of) function f.

Classification to more than two classes

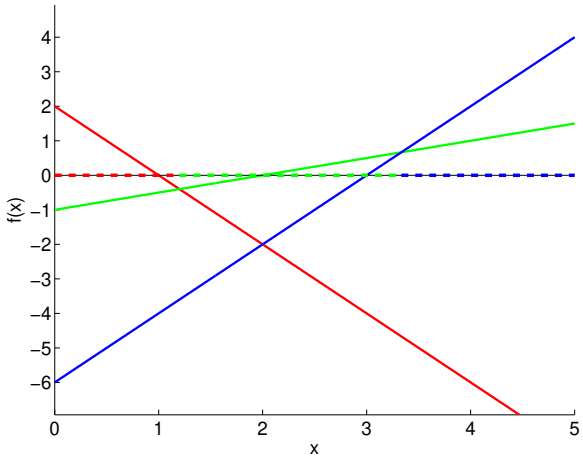
Let's have the training dataset $T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(|T|)}, y^{(|T|)})$:

- each example described by a vector $\boldsymbol{x} = (x_1, \ldots, x_D)$,
- labeled with the correct class $y \in \{1, \ldots, K\}$.

Discrimination functions $f_1(x), \ldots, f_K(x)$:

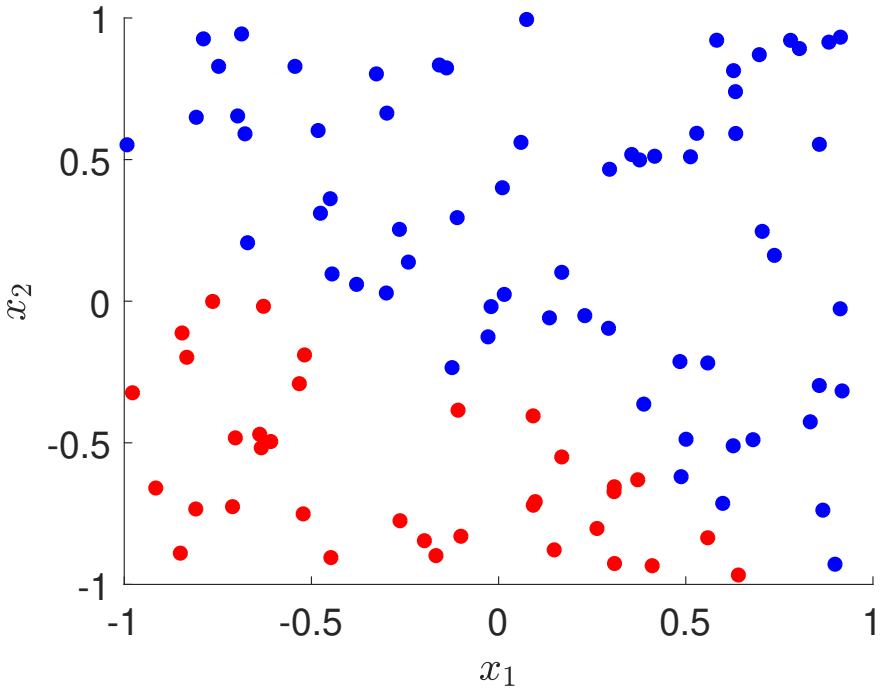
- They assign K real numbers to each observation $\boldsymbol{x}$, they may be linear or non-linear.
- Each function corresponds to a single class.
- They are used to create a **decision rule** (which assigns a class to an observation):

$$\widehat{y} = \arg\max_{k \in \langle 1,K \rangle} f_k(\boldsymbol{x}).$$

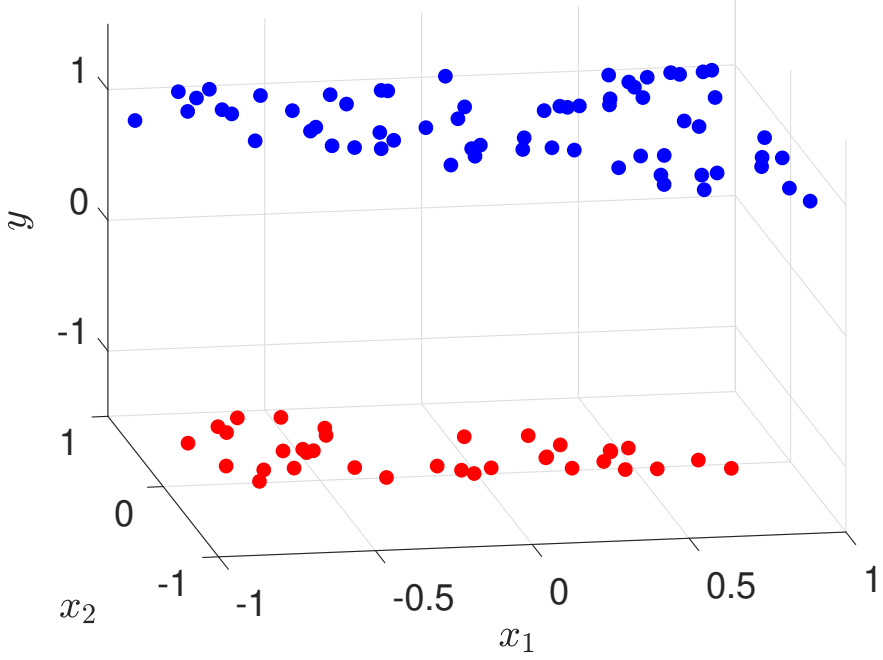


- **Decision boundaries:** $\{\boldsymbol{x} | f_i(\boldsymbol{x}) = f_j(\boldsymbol{x}) \wedge \neg\exists k : f_k(\boldsymbol{x}) > f_i(\boldsymbol{x})\}$
- **Linear classification:** the decision boundaries must be linear.
- *Learning* then amounts to finding (suitable parameters of) functions f_k.

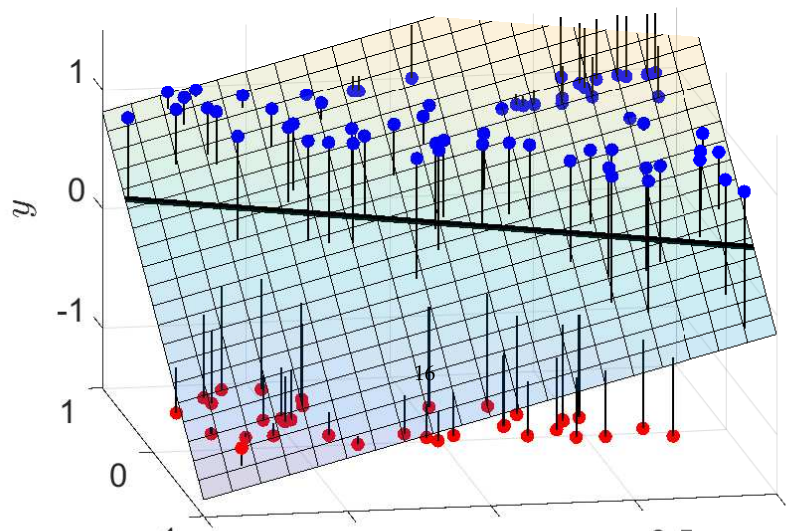
Naive approach: Illustration



Given a dataset of input vectors $\boldsymbol{x}^{(i)}$ and their classes $y^{(i)}$ ...



... we shall encode the class label as $y = -1$ and $y = 1$ ...



Naive approach

Problem: Learn a linear discrimination function f from data T.

Naive solution: fit linear regression model to the data!

- Use cost function

$$J_{MSE}(\boldsymbol{w}, T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - f(\boldsymbol{w}, \boldsymbol{x}^{(i)}) \right)^2,$$

- minimize it with respect to $\boldsymbol{w}$,
- and use $\widehat{y} = \text{sign}(f(\boldsymbol{x}))$.
- Issue: Points far away from the decision boundary have *huge effect* on the model!

Better solution: fit a linear discrimination function which minimizes the number of errors!

- Cost function:

$$J_{01}(\boldsymbol{w}, T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbb{I}(y^{(i)} \neq \widehat{y}^{(i)}),$$

  where $\mathbb{I}$ is the indicator function: $\mathbb{I}(a)$ returns 1 iff a is True, 0 otherwise.
- The cost function is non-smooth, contains plateaus, not easy to optimize, but there are algorithms which attempt to solve it, e.g. perceptron, Kozinec's algorithm, etc.

Perceptron

Perceptron algorithm

Perceptron [Ros62]:

- a simple model of a neuron
- a linear classifier (in this case, a classifier with a linear discrimination function)

Algorithm 1: Perceptron algorithm

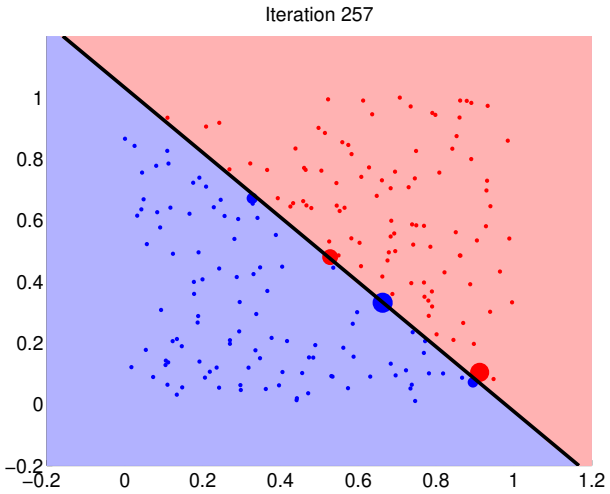
Input: Linearly separable training dataset: $\{\boldsymbol{x}^{(i)}, y^{(i)}\}, \boldsymbol{x}^{(i)} \in \mathcal{R}^{D+1}$ (homogeneous coordinates), $y^{(i)} \in \{+1, -1\}$
Output: Weight vector $\boldsymbol{w}$ such that $\boldsymbol{x}^{(i)}\boldsymbol{w}^T > 0$ iff $y^{(i)} = +1$ and $\boldsymbol{x}^{(i)}\boldsymbol{w}^T < 0$ iff $y^{(i)} = -1$

1 **begin**
2 Initialize the weight vector, e.g. $\boldsymbol{w} = \boldsymbol{0}$.
3 Invert all examples $\boldsymbol{x}$ belonging to class -1: $\boldsymbol{x}^{(i)} = -\boldsymbol{x}^{(i)}$ for all i, where $y^{(i)} = -1$.
4 Find an incorrectly classified training vector, i.e. find j such that $\boldsymbol{x}^{(j)}\boldsymbol{w}^T \leq 0$, e.g. the worst classified vector: $\boldsymbol{x}^{(j)} = \text{argmin}_{\boldsymbol{x}^{(i)}} (\boldsymbol{x}^{(i)}\boldsymbol{w}^T)$.
5 **if** *all examples classified correctly* **then**
6 Return the solution $\boldsymbol{w}$. Terminate.
7 **else**
8 Update the weight vector: $\boldsymbol{w} = \boldsymbol{w} + \boldsymbol{x}^{(j)}$.
9 Go to 4.

Instead of using the worst classified point, the algorithm may go over the training set (several times) and use all encountered wrongly classified points to update $\boldsymbol{w}$.

[Ros62] Frank Rosenblatt. *Principles of Neurodynamics: Perceptron and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C., 1962.

Demo: Perceptron



Features of the perceptron algorithm

Perceptron convergence theorem [Nov62]:

- Perceptron algorithm eventually finds a hyperplane that separates 2 classes of points in a finite number of steps, if such a hyperplane exists.
- If no separating hyperplane exists, the algorithm does not converge and will iterate forever.

Possible solutions:

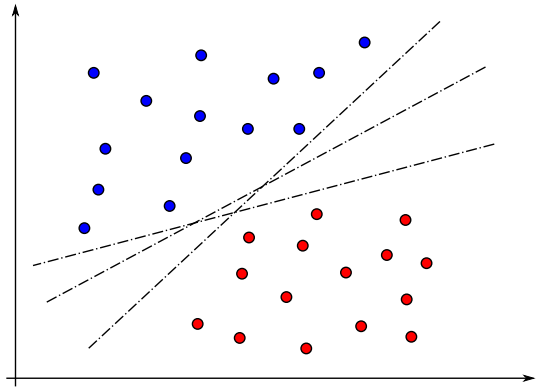
- Pocket algorithm — track the error the perceptron makes in each iteration and store the best weights found so far in a separate memory (pocket).
- Use a different learning algorithm, which finds an approximate solution, if the classes are not linearly separable.

[Nov62] Albert B. J. Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume 12, Brooklyn, New York, 1962.

The hyperplane found by perceptron

The perceptron algorithm

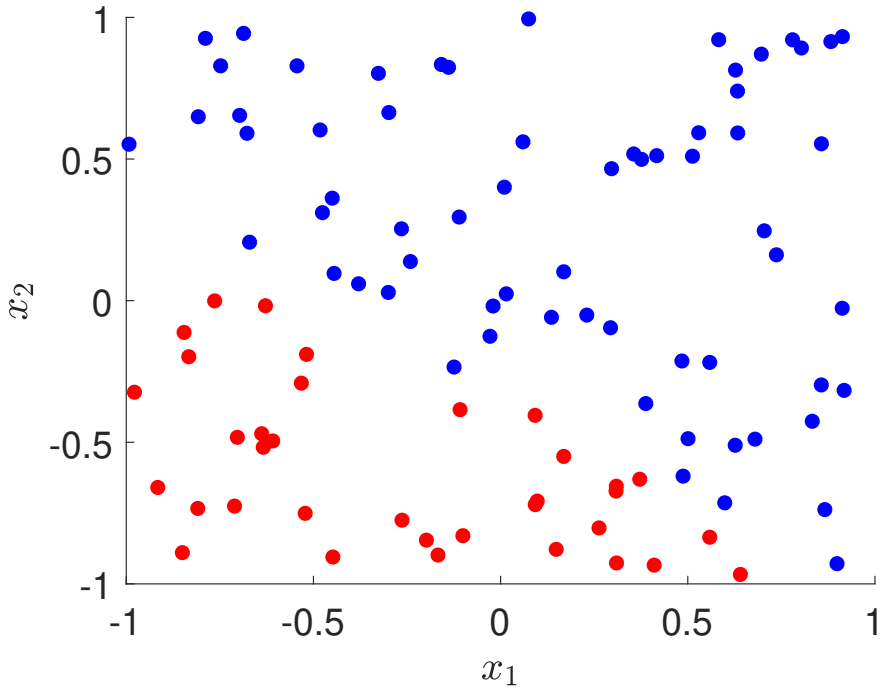
- finds a separating hyperplane, if it exists;
- but if a single separating hyperplane exists, then there are infinitely many (equally good?) separating hyperplanes.



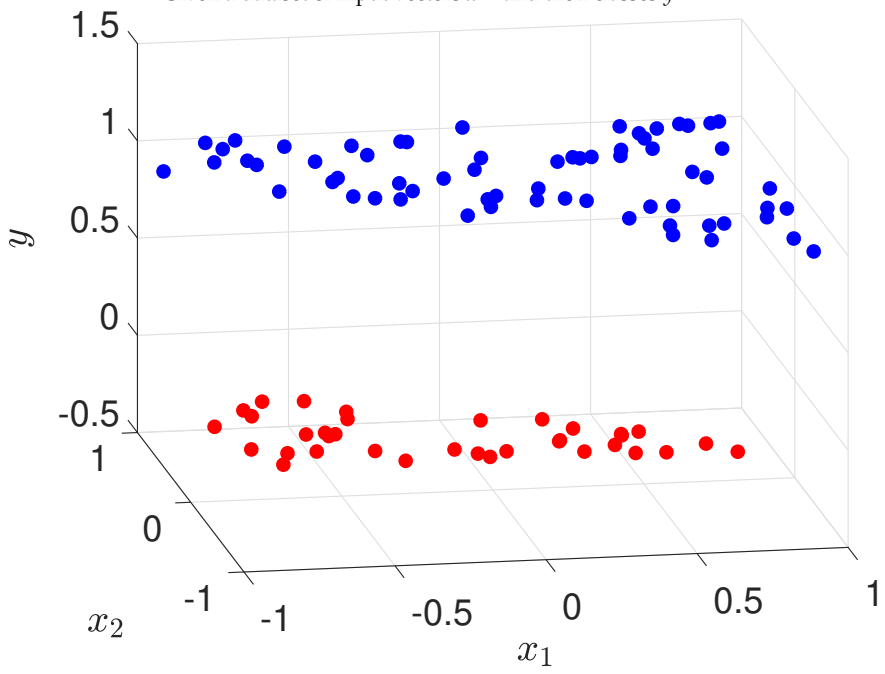
- and perceptron finds *any* of them!

Which separating hyperplane is the optimal one? What does "optimal" actually mean?

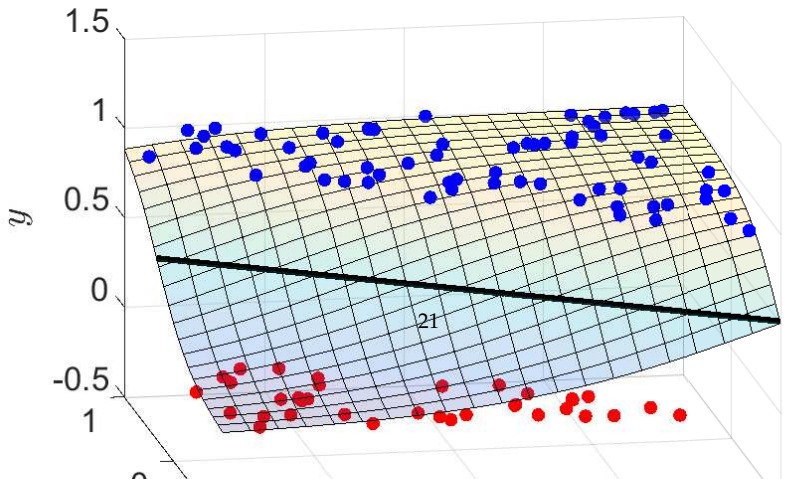
Logistic regression: Illustration



Given a dataset of input vectors $\boldsymbol{x}^{(i)}$ and their classes $y^{(i)}$ …



…we shall encode the class label as $y = 0$ and $y = 1$ …



Logistic regression model

Problem: Learn a binary **classifier** for the dataset $T = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}$, where $y^{(i)} \in \{0, 1\}$.[a]

To reiterate: when using linear regression, the examples far from the decision boundary have a huge impact on f. How to limit their influence?

Logistic regression uses a discrimination function which is a nonlinear transformation of the values of a linear function

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{x}\boldsymbol{w}^T) = \frac{1}{1 + e^{-\boldsymbol{x}\boldsymbol{w}^T}},$$

where $g(z) = \dfrac{1}{1 + e^{-z}}$ is the **sigmoid** function (a.k.a **logistic** function).

Interpretation of the model:

- $f_{\boldsymbol{w}}(\boldsymbol{x})$ is interpretted as an estimate of the probability that $\boldsymbol{x}$ belongs to class 1.
- The **decision boundary** is defined using a different level-set: $\{\boldsymbol{x} : f_{\boldsymbol{w}}(\boldsymbol{x}) = 0.5\}$.
- Logistic *regression* is a *classification model!*
- The discrimination function $f_{\boldsymbol{w}}(\boldsymbol{x})$ itself is not linear anymore; but the *decision boundary is still linear!*
- Thanks to the sigmoidal transformation, logistic regression is much less influenced by examples far from the decision boundary!

[a] Previously, we have used $y^{(i)} \in \{-1, +1\}$, but the values can be chosen arbitrarily, and $\{0, 1\}$ is convenient for logistic regression.

Cost function

To train the logistic regression model, one can use the J_{MSE} criterion:

$$J(\boldsymbol{w}, T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \left(y^{(i)} - f_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})\right)^2.$$

However, this results in a non-convex multimodal landscape which is hard to optimize.

Logistic regression uses a modified cost function (sometimes called *cross-entropy*):

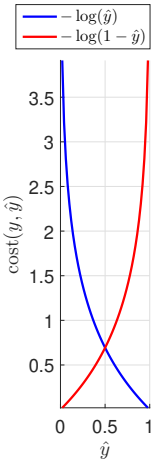
$$J(\boldsymbol{w}, T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \text{cost}(y^{(i)}, f_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})), \text{ where}$$

$$\text{cost}(y, \hat{y}) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1 - \hat{y}) & \text{if } y = 0 \end{cases},$$

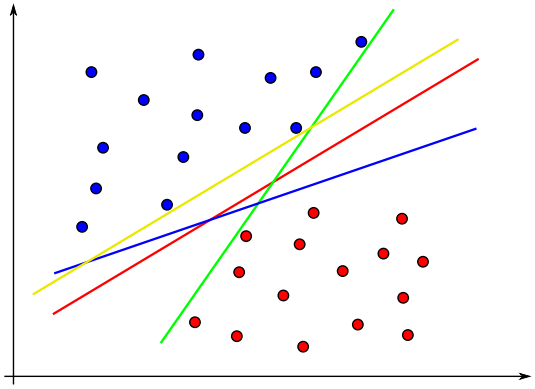
which can be rewritten in a single expression as

$$\text{cost}(y, \hat{y}) = -y \cdot \log(\hat{y}) - (1 - y) \cdot \log(1 - \hat{y}).$$

Such a cost function is simpler to optimize for numerical solvers.



Question



Which of the above decision boundaries would you choose? (And why?)

A Green line

B Yellow line

C Red line

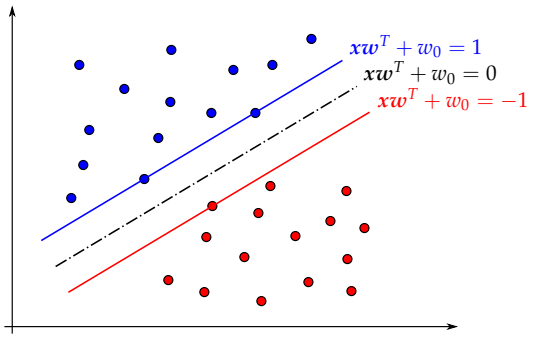
D Blue line

Optimal separating hyperplane (separable case)

Margin (cz:odstup):

- "The width of the band in which the decision boundary can move (in the direction of its normal vector) without touching any data point."

Maximum margin linear classifier



Plus 1 level: $\{x : \boldsymbol{x}\boldsymbol{w}^T + w_0 = 1\}$
Minus 1 level: $\{x : \boldsymbol{x}\boldsymbol{w}^T + w_0 = -1\}$
Decision boundary
(separating hyperplane): $\{x : \boldsymbol{x}\boldsymbol{w}^T + w_0 = 0\}$

Support vectors:

- Data points $\boldsymbol{x}$ lying at the plus 1 level or minus 1 level.
- Only these points influence the decision boundary!

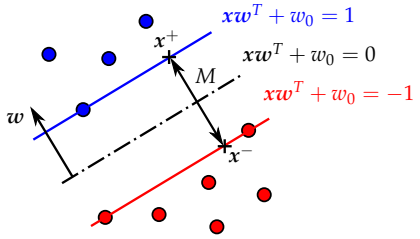
Why we would like to maximize the margin?

- Intuitively, it is safe.
- If we make a small error in estimating the boundary, the classification will likely stay correct.
- The model is invariant with respect to the training set changes, except the changes of support vectors.
- There are sound theoretical results that having a maximum margin classifier is good.
- Maximal margin works well in practice.

Margin size

How to compute the margin M given $\boldsymbol{w} = (w_1, \ldots, w_D)$, w_0 of certain sep. hyperplane?

- Let's choose two points $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$, lying in the plus 1 level and minus 1 level, respectively.
- Let's compute the margin M as their distance.



We know that:

$$\begin{aligned} \boldsymbol{x}^+\boldsymbol{w}^T + w_0 &= 1 \\ \boldsymbol{x}^-\boldsymbol{w}^T + w_0 &= -1 \\ \boldsymbol{x}^- + \lambda\boldsymbol{w} &= \boldsymbol{x}^+ \end{aligned}$$

And we can derive:

$$\begin{aligned} (\boldsymbol{x}^+ - \boldsymbol{x}^-)\boldsymbol{w}^T &= 2 \\ (\boldsymbol{x}^- + \lambda\boldsymbol{w} - \boldsymbol{x}^-)\boldsymbol{w}^T &= 2 \\ \lambda\boldsymbol{w}\boldsymbol{w}^T &= 2 \\ \lambda = \frac{2}{\boldsymbol{w}\boldsymbol{w}^T} &= \frac{2}{\|\boldsymbol{w}\|^2} \end{aligned}$$

Thus the margin size is

$$M = \|\boldsymbol{x}^+ - \boldsymbol{x}^-\| = \|\lambda\boldsymbol{w}\| = \lambda\|\boldsymbol{w}\| = \frac{2}{\|\boldsymbol{w}\|^2}\|\boldsymbol{w}\| = \frac{2}{\|\boldsymbol{w}\|}$$

Optimal separating hyperplane learning

We want to maximize margin $M = \frac{2}{\|\boldsymbol{w}\|}$ subject to the constraints ensuring correct classification of the training set T. This optimization problem can be formulated as a *quadratic programming* (QP) task.

- Primary QP task:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\boldsymbol{w}\boldsymbol{w}^T \text{ with respect to } w_0, \ldots, w_D \\ &\text{subject to } y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) \geq 1 \quad \forall i \in 1, \ldots, |T|. \end{aligned}$$

- Dual QP task:

$$\begin{aligned} &\text{maximize } \sum_{i=1}^{|T|}\alpha_i - \sum_{i=1}^{|T|}\sum_{j=1}^{|T|}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(j)}}^T \text{ with respect to } \alpha_1, \ldots, \alpha_{|T|} \\ &\text{subject to } \alpha_i \geq 0 \\ &\quad \text{and } \sum_{i=1}^{|T|}\alpha_i y^{(i)} = 0. \end{aligned}$$

- From the solution of the dual task, we can compute the solution of the primal task:

$$\boldsymbol{w} = \sum_{i=1}^{|T|}\alpha_i y^{(i)}\boldsymbol{x}^{(i)}, \qquad w_0 = y^{(k)} - \boldsymbol{x}^{(k)}\boldsymbol{w}^T,$$

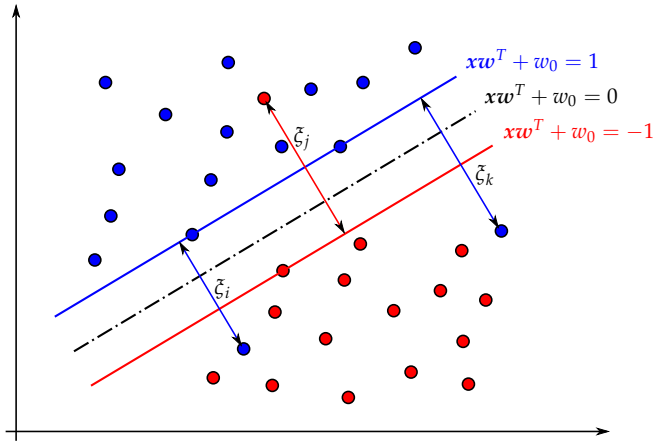
where $(\boldsymbol{x}^{(k)}, y^{(k)})$ is any *support vector*, i.e. $\alpha_k > 0$.

Non-separable case

Soft margin: Allows for incorrect classification of some data points.

Slack variables ξ_i: The shortest distances of data points to their "correct place":

- 0 for correctly classified data "outside the margin",
- positive for incorrectly classified data and data "inside the margin".



Optimal separating hyperplane learning for non-separable data

- Primary QP task with slack variables:

$$\text{minimize } \left(\frac{1}{2}\boldsymbol{w}\boldsymbol{w}^T + C\sum_{i=1}^{|T|}\xi_i\right) \text{ with respect to } w_0,\ldots,w_D,\xi_1,\ldots,\xi_{|T|}$$

$$\text{subject to } y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) \geq 1-\xi_i \quad \forall i \in 1,\ldots,|T|,$$

$$\text{and } \xi_i \geq 0 \ \ \forall i \in 1,\ldots,|T|.$$

- Dual QP task:

$$\text{maximize } \sum_{i=1}^{|T|}\alpha_i - \sum_{i=1}^{|T|}\sum_{j=1}^{|T|}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(j)}}^T \text{ with respect to } \alpha_1,\ldots,\alpha_{|T|},\mu_1,\ldots,\mu_{|T|},$$

$$\text{subject to } \alpha_i \geq 0,\ \mu_i \geq 0,\ \alpha_i + \mu_i = C,$$

$$\text{and } \sum_{i=1}^{|T|}\alpha_i y^{(i)} = 0.$$

- Variables α_i are more constrained than in the separable case, but the solution is the same:

$$\boldsymbol{w} = \sum_{i=1}^{|T|}\alpha_i y^{(i)}\boldsymbol{x}^{(i)}, \qquad w_0 = y^{(k)} - \boldsymbol{x}^{(k)}\boldsymbol{w}^T,$$

where $(\boldsymbol{x}^{(k)}, y^{(k)})$ is any *support vector*, i.e. $\alpha_k > 0$.

Lagrange function

Primary QP task:

$$(\boldsymbol{w}^*, w_0^*, \xi^*) = \arg\min_{\boldsymbol{w},w_0,\xi} \left(\frac{1}{2}\boldsymbol{w}\boldsymbol{w}^T + C\sum_{i=1}^{|T|}\xi_i \right)$$

with constraints $\forall i, i = 1, \ldots, |T|$

$$y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 + \xi_i \geq 0$$
$$\xi_i \geq 0$$

Method of Lagrange multipliers

- replaces the search for stationary points of function of D variables with K constraints by the search for stationary points of unconstrained function of $D + K$ variables;
- creates a new variable — *Langrange multiplier* — for each constraint and defines a new function, *Langrangian*, formed by the original function, constraints and multipliers.

$$L(\boldsymbol{w}, w_0, \xi_i, \alpha_i, \mu_i) = \frac{1}{2}|\boldsymbol{w}|^2 + C\sum_{i=1}^{|T|}\xi_i - \sum_{i=1}^{|T|}\alpha_i\{y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 + \xi_i\} - \sum_{i=1}^{|T|}\mu_i\xi_i$$

where

- $\alpha_i \geq 0$ are Lagrange multipliers for constraints ensuring the correct classification of points, and
- $\mu_i \geq 0$ are Lagrange multipliers for constraint on positivity of ξ_i.

The Lagrangian must be minimized w.r.t. the *primary variables* $\boldsymbol{w}$, w_0 *and* ξ_i and maximized w.r.t. the *dual variables* α_i *and* μ_i.

Dual QP Task

The dual QP task is obtained when we take the Lagrangian

$$L(\boldsymbol{w}, w_0, \xi_i, \alpha_i, \mu_i) = \frac{1}{2}|\boldsymbol{w}|^2 + C\sum_{i=1}^{|T|}\xi_i - \sum_{i=1}^{|T|}\alpha_i\{y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 + \xi_i\} - \sum_{i=1}^{|T|}\mu_i\xi_i$$

and we substitute for the primary variables $\boldsymbol{w}$, w_0 and ξ_i.

For a stationary point:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{|T|}\alpha_i y^{(i)}\boldsymbol{x}^{(i)} = 0 \quad \Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{|T|}\alpha_i y^{(i)}\boldsymbol{x}^{(i)}$$

$$\frac{\partial L}{\partial w_0} = -\sum_{i=1}^{|T|}\alpha_i y^{(i)} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{|T|}\alpha_i y^{(i)} = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad \Longrightarrow \quad C = \alpha_i + \mu_i$$

After substituting back to L and simplification we get the criterion of the dual task:

$$L_D = \frac{1}{2}\sum_{i=1}^{|T|}\sum_{j=1}^{|T|}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(j)}}^T + \sum_{i=1}^{|T|}\alpha_i\xi_i + \sum_{i=1}^{|T|}\mu_i\xi_i - \sum_{i=1}^{N}\sum_{j=1}^{|T|}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(j)}}^T$$
$$- \sum_{i=1}^{|T|}\alpha_i y^{(i)}w_0 + \sum_{i=1}^{|T|}\alpha_i - \sum_{i=1}^{|T|}\alpha_i\xi_i - \sum_{i=1}^{|T|}\mu_i\xi_i = \sum_{i=1}^{|T|}\alpha_i - \frac{1}{2}\sum_{i=1}^{|T|}\sum_{j=1}^{|T|}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(j)}}^T$$

Relations of the variables in Lagrangian

Lagrangian

$$L(\boldsymbol{w}, w_0, \xi_i, \alpha_i, \mu_i) = \frac{1}{2}|\boldsymbol{w}|^2 + C\sum_{i=1}^{|T|}\xi_i - \sum_{i=1}^{|T|}\alpha_i\{y^{(i)}(\boldsymbol{x}^{(i)}w^T + w_0) - 1 + \xi_i\} - \sum_{i=1}^{|T|}\mu_i\xi_i$$

shall be minimized w.r.t. the *primary variables* $\boldsymbol{w}$, w_0 *and* ξ_i and maximized w.r.t. the *dual variables* α_i *and* μ_i.

1. If a point $\boldsymbol{x}^{(i)}$ lies on an incorrect side of plus- or minus-plane:
   - $y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 < 0$, then $\xi_i > 0$ so that $y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 + \xi_i = 0$
   - $\xi_i > 0$ and L must be maximized w.r.t. μ_i, so that μ_i must be as small as possible, i.e. $\mu_i = 0$
   - $C = \alpha_i + \mu_i$ and $\mu_i = 0$, so that $\alpha_i = C$
2. If a point $\boldsymbol{x}^{(i)}$ lies on a correct side of plus- or minus-plane:
   - $y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 > 0$, so that $\xi_i = 0$
   - $y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 + \xi_i > 0$ and L must be maximized w.r.t. α_i, so that α_i must be as small as possible, i.e. $\alpha_i = 0$
   - $C = \alpha_i + \mu_i$ and $\alpha_i = 0$, so that $\mu_i = C$
3. If a point $\boldsymbol{x}^{(i)}$ lies directly on plus- or minus-plane:
   - $y^{(i)}(\boldsymbol{x}^{(i)}\boldsymbol{w}^T + w_0) - 1 = 0$, so that $\xi_i = 0$
   - $0 < \mu_i < C$
   - $0 < \alpha_i < C$

Optimal separating hyperplane: remarks

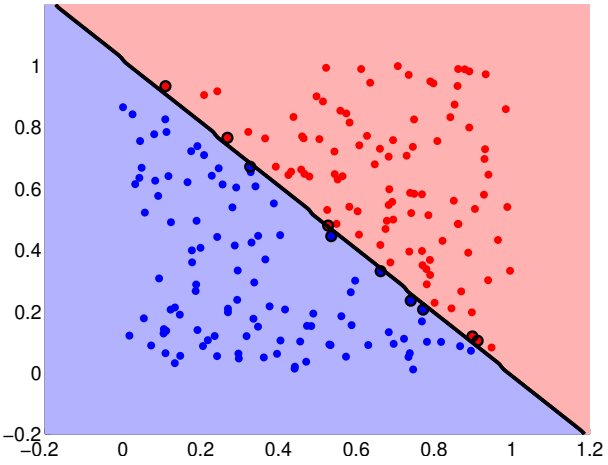
The importance of dual formulation:

- The QP task in dual formulation is easier to solve for QP solvers than the primal formulation.
- New, unseen examples can be classified using function

$$f(\boldsymbol{x}, \boldsymbol{w}, w_0) = \text{sign}(\boldsymbol{x}\boldsymbol{w}^T + w_0) = \text{sign}\left(\sum_{i=1}^{|T|}\alpha_i y^{(i)}\boldsymbol{x}^{(i)}\boldsymbol{x}^T + w_0\right),$$

  i.e. the discrimination function contains the examples $\boldsymbol{x}$ only in the form of dot products (which will be useful later).
- The examples with $\alpha_i > 0$ are *support vectors*, thus the sums may be carried out only over the support vectors.
- The dual formulation contains the data only in the form of dot products which allows for other tricks you will learn later.
- The primal task with soft margin has double the number of constraints, the task is more complex, but
- the results for the QP task with soft margin are of the same type as in the separable case.

Optimal separating hyperplane: demo



Summary 44 / 45

Competencies

After this lecture, a student shall be able to ...

1. define the types of learning (supervised, unsupervised, semisupervised, reinforcement) and describe conceptual differences between them;
2. define classification and regression types of problems, recognize them in practical situations;
3. describe 2 approaches to learning (as parameter estimation, as direct optimal strategy design) and give examples of surrogate criteria used in them.
4. define and recognize linear regression model (with scalar parameters, in scalar product form, in matrix form, non-homogenous and homogenous coordinates);
5. define the loss function suitable for fitting a regression model;
6. explain the least squares metod, draw an illustration;
7. compute coefficients of a simple (1D) linear regression by hand, write a computer program computing coefficients for multiple regression;
8. explain the concept of discrimination function for binary and multinomial classification;
9. define a loss function suitable for fitting a classification model;
10. describe a perceptron algorithm, perform a few iterations by hand;
11. explain the characteristics of the perceptron algorithm;
12. describe logistic regression, the interpretation of its outputs, and why we classify it as a linear model;
13. define loss functions suitable for fitting logistic regression;
14. define optimal separating hyperplane, explain in what sense it is optimal;
15. define what a margin is, what support vectors are, and explain their relation;
16. compute the margin given the parameters of separating hyperplane for which $\min_{i:y^{(i)}=+1}(x^{(i)}w^T + w_0) = 1$ and $\max_{i:y^{(i)}=-1}(x^{(i)}w^T + w_0) = -1$;
17. formulate the primary quadratic programming task which results in the optimal separating hyperplane (including the soft-margin version);
18. compute the parameters of optimal hyperplane given the set of support vectors and their weights.