Pattern Recognition and Machine Learning Course: Introduction. Bayesian Decision Theory.

lecturer: Jiří Matas, matas@cmp.felk.cvut.cz

authors: Václav Hlaváč, Jiří Matas, Ondřej Drbohlav

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

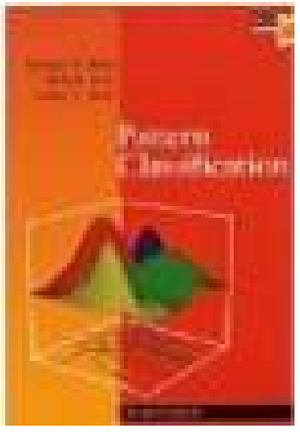
http://cmp.felk.cvut.cz

Version: 2nd October, 2017

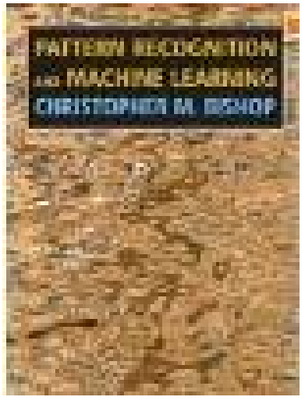
About the Pattern Recognition Course

- The selection of the topics in the course is mainstream. Besides the course material, a good wiki page is available for almost all topics covered in the course.
- We strongly recommend attendance of lectures. In PR&ML, many issues are intertwined and it is very difficult to understand the connections (e.g. understanding "why method X should be used instead of Y in case Z") just by reading about particular methods.
- Nevertheless, we do not introduce any hard "incentives" e.g. in the form of a written exam during a lecture. But correctly answering questions on lectures merits bonus points which can make up to 10% of the semester total.
- No single textbooks is ideal for Pattern Recognition and Machine Learning course. The field is still waiting for one.

Textbooks



Duda, Hart, Stork: Pattern Classification. Classical text, 2nd edition, "easy reading", about 5–10 available at the CMP library (G102, H. Pokorna will lend you a copy); some sections obsolete



Bishop: Pattern Recognition and Machine Learning. New, popular, but certain topics, in my opinion, could be presented in a clearer way

Schlesinger, Hlavac: Ten Lectures on Statistical and Structural Pattern Recognition. Advanced text, for those who want to know more than what is presented in the course; aims at maximum generality

Goodfellow, Bengio and Courville: Deep Learning http://www.deeplearningbook.org/

English/Czech Lectures

- Those of you who are fulfilling the requirement of OI to choose one course in English should attend the lecture in English, i.e. on Monday. It is acceptable to attend the Friday lectures a few times if you miss the one on Monday (ENG lectures precede the CZ ones.)
- You may attend both lectures (a couple of students did this last year to gain better understanding.)
- If English terminology is unclear, ask. As most of the terms will be used repeatedly, language problems will disappear over time.

Pattern Recognition

The course focuses on *statistical pattern recognition.*

We start with an example called *"Dilemma of a lazy short-sighted student"* which introduces most of the basic ingredients of a statistical decision problem.

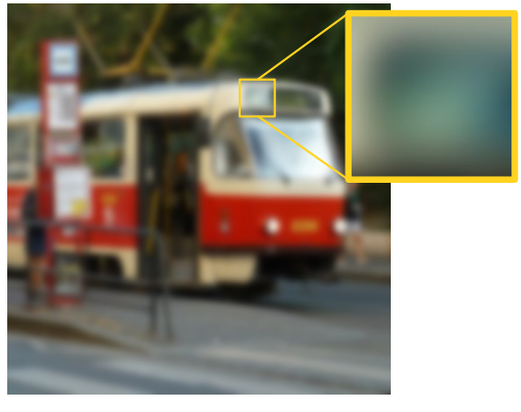
Example: A Lazy Short-Sighted Student Dilemma

A FEE student with **weak eyesight** and strong **dislike for running** is in a hurry. He needs to get to Albertov where he has arranged to play a poker game. He may get there on time, but only if he catches a tram going to Albertov immediately. He will have to pay 100 CZK fine to the poker club if he's late. As he exits the Building A at Karlovo namesti, he sees a tram at the stop:

?

He needs to decide: *(a) Run and catch the tram, or (b) not run and miss it?*

A Lazy Short-Sighted Student Dilemma

	new (box-style) tram, line 14	old (round-style) tram, line 22
Reality, tram number		
Student's observation		

The student cannot see the tram numbers, but can recognize the tram type (new, old). His decision can solely be based on the tram type.

A Lazy Short-Sighted Student Dilemma

The student:

- knows that trams 3, 6, 14, 22, 24 stop at Karlovo namesti
- knows that of those, trams 14 and 24 go to Albertov
- observes the tram type $x \in \{\text{old}, \text{new}\}$
- knows that the joint probability $p(x, k)$ of a tram of type $x \in \{\text{old}, \text{new}\}$ and line number $k \in \{3, 6, 14, 22, 24\}$ is

	3	6	14	22	24	$p(x)$
old	0.05	0.15	0.10	0.25	0.05	0.60
new	0.20	0.00	0.05	0.00	0.15	0.40
$p(k)$	0.25	0.15	0.15	0.25	0.20	

(1)

A Lazy Short-Sighted Student Dilemma

(copied from the previous slide) $p(x, k)$:

	3	6	14	22	24	$p(x)$
old type	0.05	0.15	0.10	0.25	0.05	0.60
new type	0.20	0.00	0.05	0.00	0.15	0.40
$p(k)$	0.25	0.15	0.15	0.25	0.20	

(1)

From this table, we see that:

$$p(\text{Albertov}|\text{old}) = p(14|\text{old}) + p(24|\text{old}) = 0.1/0.6 + 0.05/0.6 = 0.25 \quad (2)$$

$$p(\text{Albertov}|\text{new}) = p(14|\text{new}) + p(24|\text{new}) = 0.05/0.4 + 0.15/0.4 = 0.5 \quad (3)$$

This gives us an idea of how likely it is that a spotted tram goes in the desired direction, for both old and new types of trams. But the student prefers optimal decisions if possible.

A Lazy Short-Sighted Student Dilemma

We already know that missing a poker game means the loss of 100 CZK. If he misses the game *and* runs to the wrong tram (tram not going to Albertov), he considers this as an additional loss of 50 CZK, thus 150 CZK in total.

The loss (in CZK) can be summarized in the following table:

	he decides to run	he decides not to run
Tram goes to Albertov	0	100
Tram does not go to Albertov	150	100

Assume the probability of spotted tram going to Albertov is p. If he **runs**, the **expected** (average) loss is:

$$p \cdot \underset{\substack{\uparrow \\ \text{catches the} \\ \text{right tram}}}{0} \text{ CZK} + (1-p) \cdot \underset{\substack{\uparrow \\ \text{misses the game and} \\ \text{makes unnecessary run}}}{150} \text{ CZK} \tag{4}$$

Similarly, if he **does not run**, the expected loss is $p \cdot 100$ CZK $+(1-p) \cdot 100$ CZK $= 100$ CZK.

A Lazy Short-Sighted Student Dilemma

Thus, with $p(\text{Albertov}|\text{old}) = 0.25$ and $p(\text{Albertov}|\text{new}) = 0.5$ we have

Average loss (in CZK):

	he decides to run	he decides not to run
old	112.5	100
new	75	100

For each observation (old, new) he selects the decision which produces lower average (expected) loss.

A Lazy Short-Sighted Student Dilemma

(copied from the previous slide:)

Average loss (in CZK):

	he decides to run	he decides not to run
old	112.5	**100**
new	**75**	100

Therefore, the best strategy seems to be:

observation		
decision	do not run	run

Notes (1)

Recall the table $p(x, k)$ for $x \in \{\text{old}, \text{new}\}$ and $k \in \{3, 6, 14, 22, 24\}$:

	3	6	14	22	24	$p(x)$
old	0.05	0.15	0.10	0.25	0.05	0.60
new	0.20	0.00	0.05	0.00	0.15	0.40
$p(k)$	0.25	0.15	0.15	0.25	0.20	

(5)

The notation $p(x, k)$, $p(x)$, $p(k)$ is a shorthand that can lead to ambiguities. Here, the meanings of $p(\text{old})$, or $p(3)$ are clear.

But if trams were of type 1, 2 and 3, $p(3)$ would have been ambiguous. In that case, the notation $p(x = x')$, $p(x = x', k = k')$ could be used, e.g. $p(x = \text{old}, k = 14)$. We will use a $p_{XK}(x, k)$, $p_K(k)$ notation; $p_K(3)$ is the probability $p(k = 3)$.

The probabilities $p(x)$ and $p(k)$ are called *marginal*.

In pattern recognition literature, $p(k)$ is called *a priori probability*.

Notes (2)

In this example, there were only four strategies possible:

1. if you see an old tram, run, else don't run (and miss it)
2. if you see a new tram, run, else don't run
3. never run
4. always run

Question: Here, the set of observations is $X = \{\text{old type}, \text{new type}\}$ and there are two possible decisions, or actions: $\{\text{run}, \text{don't run}\}$. What is the number of strategies in the general case with D possible decisions and $|X|$ observations ?

Formulation of the Statistical PR Problem

Let us make a formal abstraction of the student dilemma. Let:

X be the set of observations. An observation (aka measurement, feature vector) $x \in X$ is what is known about an object.

K be the set of classes. A state $k \in K$ is what is not known about an object, it is unobservable (aka hidden parameter, hidden state, state-of-nature, class)

D be the set of possible *decisions* (actions).

p_{XK}: $X \times K \to \mathbb{R}$ be the joint probability that the object is in the state k and the observation x is made.

W: $K \times D \to \mathbb{R}$ be a *penalty (loss) function*, $W(k, d)$, $k \in K$, $d \in D$ is the penalty paid if the object is in a state k and the decision made is d. Defined for so-called Bayesian problems (will be dealt with soon).

q: $X \to D$ be a *decision function* (rule, strategy) assigning for each $x \in X$ the decision $q(x) \in D$. The quality of the strategy q can s measured by a number of ways, the most common being the expected (average) loss $R(q)$:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k)\, W(k, q(x))\,. \tag{6}$$

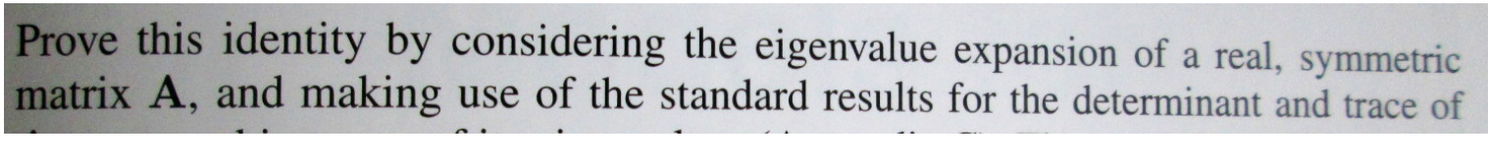
Statistical PR Problem, Examples (1)

Often, the sets of states K and decisions D coincide.
Such problem is then called *classification*.

Example 1: Vending machine

Classify coins according to their value. The set of measurements could be, say, weight, diameter and electrical resistance, thus $X \subset \mathbb{R}^3$. The set of classes is $K = \{1, 2, 5, 10, 20, 50\}$, and the set of decisions to make is $D \equiv K$.
Note: in many cases, the designer of the machine will soon discover the need to enlarge the set of decisions D by a "not a coin" class.

Example 2: Optical Character Recognition (OCR)


Prove this identity by considering the eigenvalue expansion of a real, symmetric matrix **A**, and making use of the standard results for the determinant and trace of

$\Rightarrow$ `Prove this identity by considering the eigenvalue expansion ...`

Here, an observation x is an image ($x \in X \subset \mathbb{R}^{1000000}$), $K = \{$non-character, a-z, A-Z, ...$\}$

Statistical PR Problem, Examples (2)

The observation x can be a number, symbol, function of one or two variables, a graph, algebraic structure, e.g.:

Application	Measurement	Decisions
license plate recognition	gray-level image	characters, numbers
fingerprint recognition	2D bitmap, gray-level image	personal identity
banknote verification	different sensors	{genuine, forgery}
EEG, ECG analysis	$\mathbf{x}(t)$	diagnosis
dictation machine	$x(t)$	words, sentences
speaker identification	$x(t)$	1 of N known identities
speaker verification	$x(t)$	{yes, no}
spam filter	mail content, sender, ...	{spam, ham}
stock value prediction	stock history, economic news, ...	value
recommender systems	purchase history of many users	product recommendation(s)

Examples of Statistical PR Problems: Notes (1)

- For many examples, most of the possible observations x will never appear, for most of them no x will be observed more than once.
- For most of the listed examples, there is therefore no hope of knowing $p(x, k)$.
- For some of the examples, try to estimate the cardinality of the space of observations X.
- For some of the examples, try to estimate the cardinality of the space of all possible strategies Q.

Examples of Statistical PR Problems: Notes (2)

The given formulation is very general. As seen in the example, the cardinalities of X and D (K) range from 2 to infinite.

For many applications, the formulation captures all important aspects. Nevertheless, other important aspect were ignored, e.g.:

- The choice of X, which was assumed given. In many applications, the choice of X is left to the designer.
- The cost and time of making a measurement was ignored. With a cheap camera, observations arrive instantly and at minimum cost (of powering the camera.)
  In medical applications, each measurement is costly (disposable material like vials, expensive hardware to take a scan, labor costs.)
- The time to decision, a strategy was characterized only by its loss.
- The measurements x were viewed as inputs. In many decision processes, e.g. seeing a doctor, values of initial measurements define what measurements will be made next.

Formulation of the Bayesian Decision Problem

Let the sets X, K and D, the joint probability $p_{XK}\colon X \times K \to \mathbb{R}$ and the penalty function $W\colon K \times D \to \mathbb{R}$ be given. For a strategy $q\colon X \to D$, the expectation of $W(k, q(x))$ is:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k)\, W(k, q(x))\,. \tag{7}$$

The quantity $R(q)$ is called the the Bayesian risk. Find the strategy q^* which minimizes Bayesian risk:

$$q^* = \operatorname*{argmin}_{q \in X \to D} R(q) \tag{8}$$

where the minimum is over all possible strategies. The minimizing strategy is called *Bayesian strategy.*

In the following slides, the identity

$$p_{XK}(x, k) = p_{Xk}(x|k) p_K(k) \tag{9}$$

will be used. Here, a handy notation used in the Schlesinger & Hlavac book is adopted: $p_{XK}(x, k)$ is a function of *two* variables x and k, $p_{Xk}(x|k)$ is a function of a single variable x (k is fixed), and $p_{xk}(x, k)$ is a single real number.

Finding the Bayesian Strategy (1)

The Bayesian risk $R(q^*)$ for the Bayesian strategy q^* is

$$R(q^*) = \min_{q \in X \to D} \sum_{x \in X} \sum_{k \in K} p_{XK}(x,k) W(k, q(x)) = \sum_{x \in X} \min_{q(x) \in D} \sum_{k \in K} p_{XK}(x,k) W(k, q(x)) \tag{10}$$

$$= \sum_{x \in X} p(x) \min_{q(x) \in D} \sum_{k \in K} p_{Kx}(k|x) W(k, q(x)) = \sum_{x \in X} p(x) \min_{d \in D} R(x,d)\,, \tag{11}$$

where

$$R(x,d) = \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d) \tag{12}$$

is the expectation of loss conditioned on x, called *partial risk*. From this it follows that minimization of the Bayesian Risk can be done by minizations of partial risk for each x independently. Thus, the optimal strategy $q^*(x)$ for each x can obtained as

$$q^*(x) = \operatorname*{argmin}_{d \in D} \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d)\,. \tag{13}$$

Application: Classification with 0-1 Loss Function

- The set of possible decisions D and of hidden states K coincide, $D = K$.
- The loss function assigns a unit penalty if $q(x) \neq k$, and no penalty otherwise, i.e.

$$W\big(k, q(x)\big) = \begin{cases} 0 & \text{if} \quad q(x) = k \\ 1 & \text{if} \quad q(x) \neq k \end{cases} \tag{14}$$

The partial risk for x is

$$R(x, d) = \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k, d) = \sum_{k \neq d} p_{Kx}(k \,|\, x) = 1 - p_{Kx}(d \,|\, x)\,, \tag{15}$$

and the optimal strategy for this x is then

$$q^*(x) = \operatorname*{argmin}_{d \in D} R(x, d) = \operatorname*{argmax}_{d \in D} p_{Kx}(d \,|\, x)\,. \tag{16}$$

Result: The Bayesian strategy for this problem is: For a given ovservation x, decide for the state d with the highest *a posteriori* probability $p_{Kx}(d \,|\, x)$.

Classification with 0-1 Loss Function: Example 1

Problem: Using a single measurement of sky cloudiness, decide whether it will rain. The cloudiness has four scales, ranging from 1 (no clouds) to 4 (very cloudy). There are two hidden states, 'rain' and 'no rain'. The joint probability $p(x, k)$ for $x \in \{1, 2, 3, 4\}$ and $k \in \{\text{rain}, \text{no rain}\}$ is as follows:

$p(x, k)$

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

There is no need to actually compute $p(k|x)$; it only matters whether or not

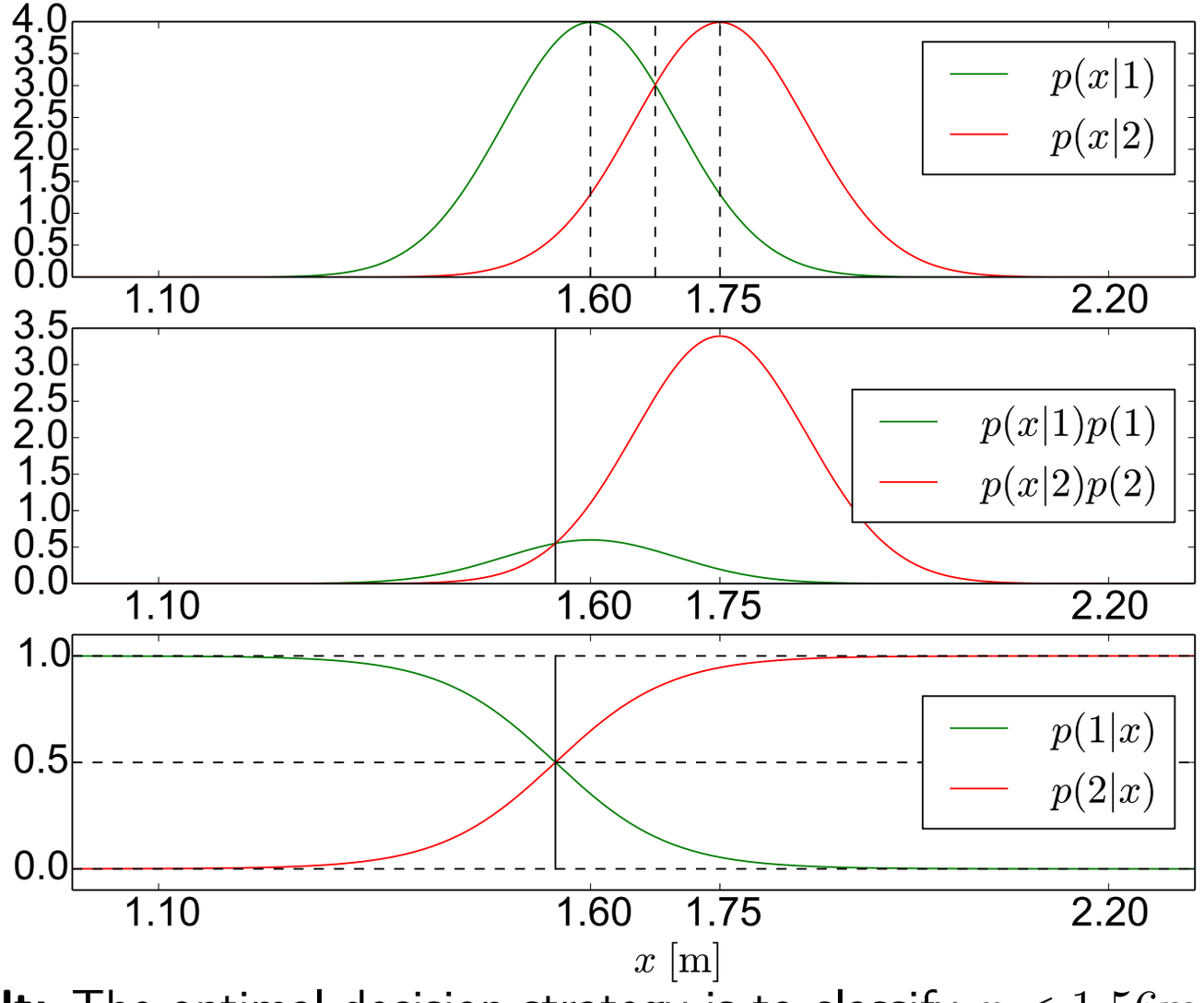
$$p(\text{rain}|x) > p(\text{no rain}|x) \Leftrightarrow p(\text{rain}, x) > p(\text{no rain}, x) \tag{17}$$

for a given observation x. Thus,

$$\begin{aligned} q^*(1) &= \text{no rain}\,, \\ q^*(2) &= \text{no rain}\,, \\ q^*(3) &= \text{rain}\,, \\ q^*(4) &= \text{rain}\,. \end{aligned} \tag{18}$$

Classification with 0-1 Loss Function: Example 2

Problem: Classify gender based on body height. Observation: $x =$ body height [m], class set $K = D = \{1, 2\}$ ($1 =$ woman, $2 =$ man.) The conditionals $p(x|1)$ and $p(x|2)$ are given below and the priors (at FEL) are $p(1) = 0.15$, $p(2) = 0.85$. Let the loss function be 0-1.



Result: The optimal decision strategy is to classify $x < 1.56\text{m}$ as women and $x > 1.56\text{m}$ as men.

Application: Bayesian Strategy with the Reject Option (1)

Consider an examination where for each question there are three possible answers: yes, no, `not known`. If your answer is correct, 1 point is added to your score. If your answer is wrong, 3 points are subtracted. If your answer is `not known`, your score is unchanged. What is the optimal Bayesian strategy if for each question you know the probabilities that $p(\text{yes})$ is the right answer?

Note that adding a fixed amount to all penalties and multiplying all penalties by a fixed amount does not change the optimal strategy. Adding 3 and multiplying by $1/4$ leads to 1 point for correct answer, $3/4$ for `not known` and 0 points of a wrong answer.

Any problem of this type can be transformed to an equivalent problem with penalty 0 for the correct answer, 1 for the wrong answer, and ϵ for `not known`. In realistic problems, $\epsilon \in (0, 1)$, since $\epsilon \geq 1$ means it is always better to guess than to say `not known`; $\epsilon \leq 0$ states that saying `not known` is preferred to giving the correct answer.

Let us solve the problem formally.

Application: Bayesian Strategy with Reject Option (2)

Let X and K be sets of observations and states, $p_{XK}\colon X \times K \to \mathbb{R}$ be a probability distribution and $D = K \cup \{\texttt{not known}\}$ be a set of decisions.

Let us define $W(k,d)$, $k \in K$, $d \in D$:

$$W(k,d) = \begin{cases} 0, & \text{if } d = k\,, \\ 1, & \text{if } d \neq k \text{ and } d \neq \texttt{not known}\,, \\ \varepsilon, & \text{if } d = \texttt{not known}\,. \end{cases}$$

Find the Bayesian strategy $q^*\colon X \to D$. The decision $q^*(x)$ corresponding to the observation x has to minimize the partial risk,

$$q^*(x) = \operatorname*{argmin}_{d \in D} \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d)\,.$$

Application: Bayesian Strategy with Reject Option (3)

Equivalent formulation of partial risk minimization:

$$q^*(x) = \begin{cases} \operatorname*{argmin}_{d \in K} R(x,d)\,, & \text{if } \min_{d \in K} R(x,d) < R(x, \texttt{not known})\,, \\ \texttt{not known}\,, & \text{if } \min_{d \in K} R(x,d) \geq R(x, \texttt{not known})\,. \end{cases}$$

For $\min_{d \in K} R(x,d)$, there holds (as before for the 0-1 loss function case):

$$\begin{aligned} \min_{d \in K} R(x,d) &= \min_{d \in K} \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d) && (19) \\ &= \min_{d \in K} \sum_{k \in K \setminus \{d\}} p_{Kx}(k \,|\, x) && (20) \\ &= \min_{d \in K} \left(\sum_{k \in K} p_{Kx}(k \,|\, x) - p_{Kx}(d \,|\, x) \right) && (21) \\ &= \min_{d \in K} \left(1 - p_{Kx}(d \,|\, x)\right) = 1 - \max_{d \in K} p_{Kx}(d \,|\, x)\,. && (22) \end{aligned}$$

Application: Bayesian Strategy with Reject Option (4)

For $R(x, \texttt{not known})$, there holds

$$\begin{aligned} R(x, \texttt{not known}) &= \sum_{k \in K} p_{K|X}(k \,|\, x)\; W(k, \texttt{not known}) \\ &= \sum_{k \in K} p_{K|X}(k \,|\, x)\; \varepsilon = \varepsilon \,. \end{aligned} \tag{23}$$

The decision rule becomes

$$q^*(x) = \begin{cases} \underset{k \in K}{\operatorname{argmax}}\; p_{K|X}(k \,|\, x)\,, & \text{if } 1 - \max\limits_{k \in K} p_{K|X}(k \,|\, x) < \varepsilon\,, \\ \texttt{not known}\,, & \text{if } 1 - \max\limits_{k \in K} p_{K|X}(k \,|\, x) \geq \varepsilon\,. \end{cases}$$

Application: Bayesian Strategy with Reject Option (5)

Strategy $q^*(x)$ can thus be described as follows:

First, find the state k which has the largest *a posteriori* probability.

If this probability is larger than $1 - \varepsilon$ then the optimal decision is k.

If its probability is not larger than $1 - \varepsilon$ then the optimal decision is `not known` .

Case of 2 Classes. Likelihood Ratio

- Let the number of classes be two; $K = \{1, 2\}$.
- Only conditional probabilities $p_{X|1}(x)$ and $p_{X|2}(x)$ are known.
- The *a priori* probabilities $p_K(1)$ and $p_K(2)$ and penalties $W(k, d)$, $k \in \{1, 2\}$, $d \in D$, are not known.
- In this situation the Bayesian strategy cannot be created.

Likelihood Ratio (1)

If the *a priori* probabilities $p_K(k)$ and the penalty $W(k,d)$ are known then the decision $q^*(x)$ about the observation x is

$$\begin{aligned} q^*(x) &= \operatorname*{argmin}_d \big(p_{XK}(x,1)\, W(1,d) + p_{XK}(x,2)\, W(2,d)\big) \\ &= \operatorname*{argmin}_d \big(p_{X|1}(x)\, p_K(1)\, W(1,d) + p_{X|2}(x)\, p_K(2)\, W(2,d)\big) \\ &= \operatorname*{argmin}_d \big(\frac{p_{X|1}(x)}{p_{X|2}(x)}\, p_K(1)\, W(1,d) + p_K(2)\, W(2,d)\big) \\ &= \operatorname*{argmin}_d \big(\gamma(x)\, c_1(d) + c_2(d)\big)\,. \end{aligned}$$

$\gamma(x)$ – likelihood ratio.

Likelihood Ratio (2) – linearity, convex subset of $\mathbb{R}$

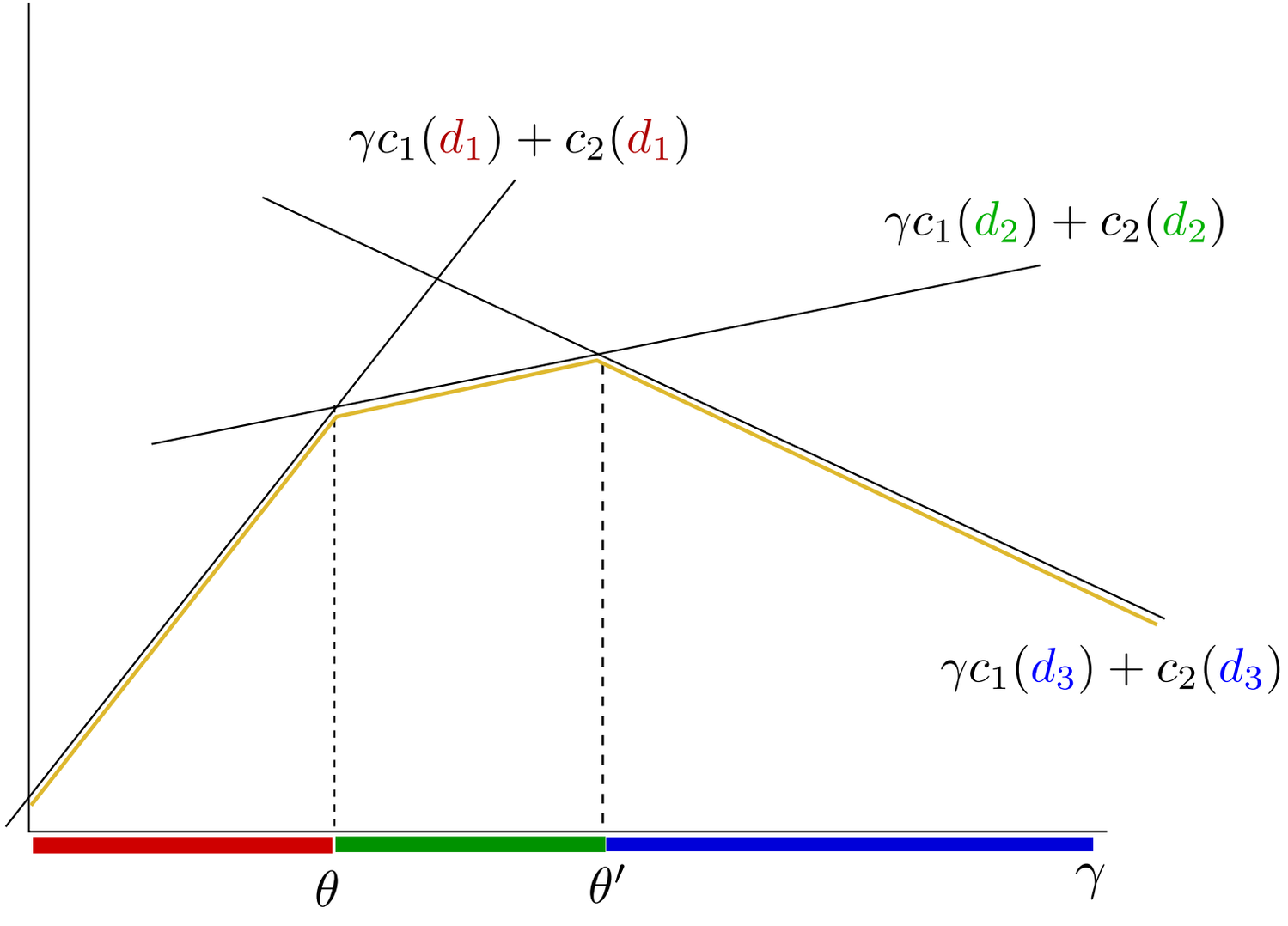
The subset of observations $X(d^*)$ for which the decision d^* should be made is the solution of the system of inequalities

$$\gamma(x)\, c_1(d^*) + c_2(d^*) \leq \gamma(x)\, c_1(d) + c_2(d)\,, \quad d \in D \setminus \{d^*\}\,.$$

- The system is linear with respect to the likelihood ratio $\gamma(x)$.
- The subset $X(d^*)$ corresponds to a convex subset of the values of the likelihood ratio $\gamma(x)$.
- As $\gamma(x)$ are real numbers, their convex subsets correspond to the numerical intervals.

Likelihood Ratio (3) – linearity, convex subset of $\mathbb{R}$

Example, 2 classes, 3 possible decisions
$\gamma : X \to \mathbb{R}$



Likelihood Ratio (4)

Any Bayesian strategy divides the real axis from 0 to ∞ into $|D|$ intervals $I(d)$, $d \in D$. The decision d is made for observation $x \in X$ when the likelihood ratio $\gamma = p_{X|1}(x)/p_{X|2}(x)$ belongs to the interval $I(d)$.

More particular case which is commonly known:

Two decisions only, $D = \{1, 2\}$. Bayesian strategy is characterised by a single threshold value θ. For an observation x the decision depends only on whether the likelihood ratio is larger or smaller than θ.

Example. 2 Classes, 3 Decisions

Object: a patient examined by the physician.

Observations X: some measurable parameters (temperature, . . .).

2 unobservable states $K = \{\text{healthy, sick}\}$.

3 decisions $D = \{\text{do not cure, weak medicine, strong medicine}\}$.

Penalty function $W : K \times D \to \mathbb{R}$

$W(k,d)$	do not cure	weak medicine	strong medicine
sick	10	2	0
healthy	0	5	10

Comments on the Bayesian Decision Problem.

Bayesian recognition is decision-making, where

- Decisions do not influence the state of nature (c.f. Game T., Control T.).
- A single decisions is made, issues of time are ignored in the model (unlike in Control Theory where decisions are typically taken continuously and in real-time)
- Cost of obtaining measurements is not modelled (unlike in Sequential Decision Theory).

The hidden parameter k (class information) is considered not observable. Common situations are:

- k could be observed, but at a high cost.
- k is a future state (e.g. of petrol price) and will be observed later.

It is interesting to ponder whether a state can ever be genuinely unobservable.

Generality of the Bayesian task formulation.

Two general properties of Bayesian strategies:

- Each Bayesian strategy corresponds to separation of the space of probabilities into convex subsets.
- Deterministic strategies are always better than randomized ones.

Bayesian Strategies are Deterministic

Instead of $q\colon X \to D$ consider stochastic strategy (probability distributions) $q_r(d \,|\, x)$.

THEOREM

Let X, K, D be finite sets, $p_{XK}\colon X \times K \to \mathbb{R}$ be a probability distribution, $W\colon K \times D \to \mathbb{R}$ be a penalty function. Let $q_r\colon D \times X \to \mathbb{R}$ be a stochastic strategy, i.e a strategy that selects decisions d with probability $q_r(d|x)$. The risk of the stochastic strategy is:

$$R_{\text{rand}} = \sum_{x \in X} \sum_{k \in K} p_{XK}(x,k) \sum_{d \in D} q_r(d \,|\, x)\, W(k,d)\,.$$

In such a case there exists the deterministic strategy $q\colon X \to D$ with the risk

$$R_{\text{det}} = \sum_{x \in X} \sum_{k \in K} p_{XK}(x,k)\, W\big(k, q(x)\big)$$

which is not greater than R_{rand}.

Note that $q_r(d|x)$ has the following properties for all x: $(i) \sum_{d \in D} q_r(d \,|\, x) = 1$ and $(ii) q_r(d \,|\, x) \geq 0$, $d \in D$.

PROOF #1 (Bayesian strategy are deterministic)

Comparing the risks associated with deterministic and stochastic strategies

$$R_{\text{rand}} = \sum_{x\in X}\sum_{k\in K} p_{XK}(x,k)\sum_{d\in D} q_r(d\,|\,x)\,W(k,d)\,, \quad R_{\text{det}} = \sum_{x\in X}\sum_{k\in K} p_{XK}(x,k)\,W\big(k,q(x)\big)$$

it is clear it is sufficient to prove that for every x

$$\sum_{k\in K} p_{XK}(x,k)\sum_{d\in D} q_r(d\,|\,x)\,W(k,d) \;\geq \sum_{k\in K} p_{XK}(x,k)\,W\big(k,q(x)\big)$$

Let us denote the losses associated with deterministic decision d as $\alpha_d = \sum_{k\in K} p_{XK}(x,k)\,W\big(k,d\big)$ and let the loss of the best deterministic strategy be denoted $\alpha_{d^*} = \min_{d\in D}\alpha_d$. Expressing the stochastic loss in terms of α_d we obtain:

$$\sum_{k\in K} p_{XK}(x,k)\sum_{d\in D} q_r(d\,|\,x)\,W(k,d) = \sum_{d\in D} q_r(d\,|\,x)\sum_{k\in K} p_{XK}(x,k)\,W(k,d) = \sum_{d\in D} q_r(d\,|\,x)\alpha_d$$

To prove the theorem, it is sufficient to show that $\sum_{d\in D} q_r(d\,|\,x)\alpha_d \geq \alpha_{d^*}$:

$$\forall d\in D : \alpha_d \geq \alpha_{d^*} \Rightarrow \sum_{d\in D} q_r(d\,|\,x)\alpha_d \geq \sum_{d\in D} q_r(d\,|\,x)\alpha_{d^*} = \alpha_{d^*}\sum_{d\in D} q_r(d\,|\,x) = \alpha_{d^*} \quad \square$$

.

PROOF #2 (Bayesian strategy are deterministic)

$$R_{\text{rand}} = \sum_{x \in X} \sum_{d \in D} q_r(d \,|\, x) \sum_{k \in K} p_{XK}(x, k)\, W(k, d)\,.$$

$\sum_{d \in D} q_r(d \,|\, x) = 1,\ x \in X, \quad q_r(d \,|\, x) \geq 0,\ d \in D,\ x \in X.$

$$R_{\text{rand}} \geq \sum_{x \in X} \min_{d \in D} \sum_{k \in K} p_{XK}(x, k)\, W(k, d) \quad \text{holds for all } x \in X,\ d \in D\,. \tag{24}$$

Let us denote by $q(x)$ any value d that satisfies the equality

$$\sum_{k \in K} p_{XK}(x, k)\, W\big(k, q(x)\big) = \min_{d \in D} \sum_{k \in K} p_{XK}(x, k)\, W(k, d)\,. \tag{25}$$

The function $q\colon X \to D$ defined in such a way is a deterministic strategy which is not worse than the stochastic strategy q_r. In fact, when we substitute Equation (25) into the inequality (24) then we obtain the inequality

$$R_{\text{rand}} \geq \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k)\, W\big(k, q(x)\big)\,.$$

The risk of the deterministic strategy q can be found on the right-hand side of the preceding inequality. It can be seen that $R_{\text{det}} \leq R_{\text{rand}}$ holds.

What's next? (1)

- The first part of the course is about solving statistical pattern recognition problems when the model $p_{XK}(x, k)$ is known.
- It is very rare that $p_{XK}(x, k)$ is known for a given application. Instead, it is almost always possible to obtain a set of representative samples T of (measurement, class) pairs.
  Example: Gender recognition. A person labels 1000 face images as male/female.
- One way to proceed is to find and estimate $p_{XK}(x, k)$ from T and proceed as if the estimate was equal to the true probability. A much more common approach is to obtain a strategy q ($=$ a classifier) with desirable properties directly from T.

What's next? (2)

The next lecture will deal with problems illustrated by a modified version of the **Student Dilemma**:

A student with a weak eyesight and a strong dislike for running is in a hurry. He needs to get to Albertov, where his girlfriend, a medical student is expecting him in 10 minutes. He might get there on time, but he needs to catch a tram immediately.

As he exits Building A at Karlovo namesti, he sees a tram at the stop. He cannot see the tram number as he is short-sighted, but he recognizes the tram is the rectangular shaped "new style" one, not the rounded "old style".

He knows, as before, the sets X, K, and the joint probability $p_{XK}(x, k)$ for all $x \in X, k \in K$.

He knows that his girlfriend tolerates him being late 20% of the time, and does not even comment. But she'd dump him if gets above that.

When should he run?

Interestingly, in this case, the student need not assign a cost to running or to the loosing his girlfriend (which might be rather difficult).

He needs a strategy that will tell him to run as rarely as possible, given the constraint: he must catch the tram 80% of time else he looses his girlfriend.