Pattern Recognition 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

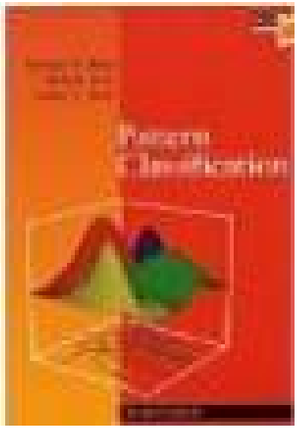
http://cmp.felk.cvut.cz

2nd October, 2015

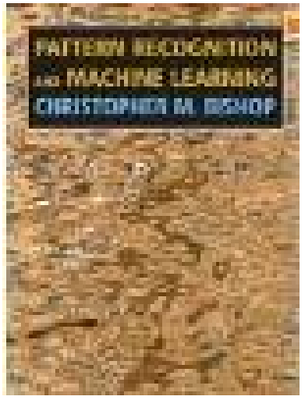
About the Pattern Recognition Course

- The selection of the topics in the course is mainstream. Besides course material, a good wiki page is available for almost all topic cover 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 "incentives" e.g. in the form of a written exam during a lecture.
- 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

English/Czech Lectures

- Those of you who are fulfilling the requirement of OI to chose 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.
- 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 of OI" which introduces most of the basic ingredients of a statistical decision problem.

Example: A lazy short-sighted OI 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 he has arranged to play a poker game. He might get there on time, but he needs to catch a tram immediately. The club rules stipulate he'll have to pay 100 CZK fine if he's late. 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 recognises the tram is the rectangular shaped "new style" one, not the rounded "old style".

Should he run?

Example: A Lazy Short-Sighted Student Dilemma

The student prefers well-justified, and if possible, optimal decisions. He travels to Albertov regularly and he knows:

- trams 18 and 24 go to Albertov
- the following trams stop at Karlovo namesti: 3, 6, 18, 22, 24
- Denote the set of tram types $X = \{\text{old}, \text{new}\}$ and the set of tram numbers $K = \{3, 6, 18, 22, 24\}$. The joint probability $p(x, k)$ of a tram of type $x \in X$ and line number $k \in K$ is:

	3	6	18	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)

Should he run?

Example: A Lazy Short-Sighted Student Dilemma

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

	3	6	18	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)

The notation $p(x, k)$, $p(x)$, $p(k)$ is a shorthand that can lead to ambiguities. But 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 be ambiguous. In that case, the notation $p(x = x')$, $p(x = x', k = k')$ will be used, e.g. $p(x = \text{old}, k = 18)$. Some textbooks 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*.

OI Student Dilemma

So should he run?

We still do not know enough to give a well-justified advice.

We know that by missing a tram 18 or 24 he'll loose 100 CZK.

Let us assume that the student is a perfect *homo economicus*. In his decisions, everything is converted to financial loss or gain. He values a needless run to be a loss of 50 CZK.

The advice will have a form of a strategy. In this example, there are 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 OI student dilemma. Let:

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

K is a set of hidden states. 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 a 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 expected (average) loss is the most common.

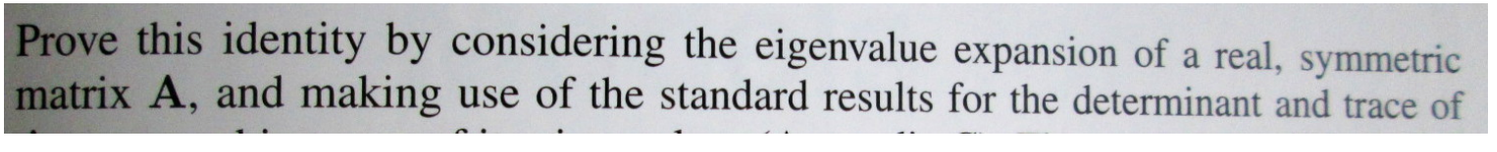
Statistical PR Problem, Examples (1)

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 hidden 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	$\bar{x}(t)$	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}

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 formulation given 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.

Examples of Statistical PR Problems: Notes (3)

- In some problems, the hidden state k cannot be observed in principle. Example: k is the value of the dollar against CZK tomorrow. x is the exchange rate for each day in the last year. Decisions are "sell USD now" or "buy USD now".
- Often the "hidden state" is potentially observable, but at a large cost. It is practical to equip notebooks with fingerprint readers and solve a statistical PR problem, with acceptable precision. A DNA analyzer would be error-free, but too costly.

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:

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

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

$$q^* = \operatorname*{argmin}_{q \in X \to D} R(q) \tag{3}$$

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

In the following slides, the identity

$$p_{XK}(x, k) = p_{Xk}(x|k) p_K(k) \tag{4}$$

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

$$R(q^*) = \min_{q \in X \to D} \sum_{x \in X} \sum_{k \in K} p_{XK}(x,k) W(k, q(x)) = \sum_{x \in X} \min_{q(x) \in D} \sum_{k \in K} p_{XK}(x,k) W(k, q(x)) \tag{5}$$

$$= \sum_{x \in X} p(x) \min_{q(x) \in D} \sum_{k \in K} p_{Kx}(k|x) W(k, q(x)) = \sum_{x \in X} p(x) \min_{d \in D} R(x,d)\,, \tag{6}$$

where

$$R(x,d) = \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d) \tag{7}$$

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

$$q^*(x) = \operatorname*{argmin}_{d \in D} \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k,d)\,. \tag{8}$$

Classification with 0-1 Loss Function (1)

- 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.

$$W\big(k, q(x)\big) = \begin{cases} 0 & \text{if} \quad q(x) = k \\ 1 & \text{if} \quad q(x) \neq k \end{cases} \tag{9}$$

The partial risk for x is

$$R(x, d) = \sum_{k \in K} p_{Kx}(k \,|\, x)\; W(k, d) = \sum_{k \neq d} p_{Kx}(k \,|\, x)\; W(k, d) = 1 - p_{Kx}(d \,|\, x)\,, \tag{10}$$

and the optimal strategy for this x is then

$$q^*(x) = \operatorname*{argmin}_{d \in D} R(x, d) = \operatorname*{argmax}_{d \in D} p_{Kx}(d|x)\,. \tag{11}$$

Result:

The Bayesian stragegy for this problem is the state d with the highest *a posteriori* probability $p_{Kx}(d \,|\, x)$.

Classification with 0-1 Loss Function (2)

The result shows that the *a posteriori* probability of each state k is to be calculated for the observation x and the optimal decision is in favour of the most probable state. The the maximum *a posteriori strategy* is the Bayesian strategy for the 0-1 loss function.

Dichotomy. In the situation with two possible decisions (and classes), the optimal decision can be expressed as a sign of discriminative function $g(x) = p_{kx}(1 \mid x) - p_{kx}(0 \mid x)$.

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.

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,

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

Bayesian Strategy with Reject Option (3)

Equivalent definition of partial risk

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

There holds for $\min_{d \in K} R(x,d)$

$$\begin{aligned} \min_{d \in K} R(x,d) &= \min_{d \in K} \sum_{k^* \in K} p_{K|X}(k^* \,|\, x)\; W(k^*, d) \\ &= \min_{k \in K} \sum_{k^* \in K \setminus \{k\}} p_{K|X}(k^* \,|\, x) \\ &= \min_{k \in K} \left(\sum_{k^* \in K} p_{K|X}(k^* \,|\, x) - p_{K|X}(k \,|\, x) \right) \\ &= \min_{k \in K} \big(1 - p_{K|X}(k \,|\, x)\big) = 1 - \max_{k \in K} p_{K|X}(k \,|\, x)\,. \end{aligned}$$

Bayesian Strategy with Reject Option (4)

There holds for $R(x, \texttt{not known})$

$$\begin{aligned} R(x, \texttt{not known}) &= \sum_{k \in K} p_{K|X}(k \,|\, x)\; W(k, \texttt{not known}) \\ &= \sum_{k \in K} p_{K|X}(k \,|\, x)\; \varepsilon = \varepsilon \,. \end{aligned} \tag{12}$$

The decision rule becomes

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

Bayesian Strategy with Reject Option (5)

Strategy $q^*(x)$ can 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` .

Convex subspaces. Special case: 2 hidden states.

- Hidden state assumes two values only, $K = \{1, 2\}$.
- 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)

Note:

There can be more than two decisions $d \in D$, $|D| > 2$ for only two states, $|K| = 2$.

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 Hidden States, 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)

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

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

$$R_{\text{rand}} \geq \sum_{x \in X} \min_{d \in D} \sum_{k \in K} p_{XK}(x, k)\, W(k, d) \quad \text{holds for all } x \in X,\ d \in D\,. \tag{13}$$

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

$$\sum_{k \in K} p_{XK}(x, k)\, W\big(k, q(x)\big) = \min_{d \in D} \sum_{k \in K} p_{XK}(x, k)\, W(k, d)\,. \tag{14}$$

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 (14) into the inequality (13) 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.

Importance of Linear Classifiers.

- Theoretical importance, decomposition of the probability space into convex cones.
- For some statistical models, the Bayesian or non-Bayesian strategy is implemented by linear discriminant function.
- Some non-linear discriminant functions can be implemented as linear after straightening the feature space.
- Capacity (VC dimension) of linear strategies in an n-dimensional space is $n + 2$. Thus, the learning task is correct, i.e., strategy tuned on finite training multiset does not differ much from correct strategy found for a statistical model.
- There are efficient algorithms to solve them.

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 man/woman.
- 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.

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.

What's next? (2)

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.