

BAYESIAN DECISION MAKING

Václav Hlaváč

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic
hlavac@fel.cvut.cz, <http://cmp.felk.cvut.cz>

MOTIVATION

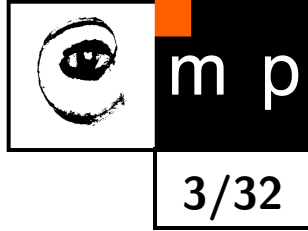
Why to study pattern recognition?

- ◆ constitutes a solid theoretical base.
- ◆ has regained strong research interest recently.
- ◆ is useful in variety of applications.

Monograph Schlesinger M.I., Hlaváč V.: *Ten lectures on statistical and structural pattern recognition*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 2002, 519 p.

Toolbox in Matlab written by V. Franc; the most comprehensive pattern recognition toolbox, publicly available.

BAYESIAN DECISION MAKING, CONCEPTS



Object (situation) is described by two parameters:

x is observable *feature* (observation).

k unobservable *hidden parameter* (state).

X is finite set of observations, $x \in X$.

K is finite set of hidden state, $k \in K$.

D is the finite set of possible *decisions*.

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

W : $K \times D \rightarrow \mathbb{R}$ is a *penalty function*, $W(k, d)$, $k \in K$, $d \in D$ is the penalty paid in for the object in the state k and the decision d made.

q : $X \rightarrow D$ is a *decision function* (rule, strategy) assigning for each $x \in X$ the decision $q(x) \in D$.

$R(q)$ is *risk*, i.e. the mathematical expectation of the penalty.

NOTATION REMARKS

- ◆ The joint probability $p_{XK}(x, k)$ is often expressed as $p_{XK}(x, k) = p_{Xk}(x|k) \cdot p_K(k)$.
- ◆ The standard notation for joint and conditional probabilities is ambiguous.
 - Are $p(x, k)$ and $p(x|k)$, respectively, a number, a function of a single variable or a function of two variables?
 - M. Schlesinger disambiguates with subscripts:
 $p_{XK}(x, k)$ is a *function of two variables*,
 $p_{Xk}(x|k)$ is a *function of a single variable x* ,
 and $p_{xk}(x, k)$ is a *single real number*.

DEFINITION OF THE BAYESIAN TASK

Bayesian task of statistical decision making seeks

for sets X , K and D , function $p_{XK}: X \times K \rightarrow \mathbb{R}$ and function $W: K \times D \rightarrow \mathbb{R}$

a strategy $q: X \rightarrow D$ which **minimizes the Bayesian** risk

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

The solution to the Bayesian task is the **Bayesian strategy** q minimizing the risk.

The formulation can be extended to infinite X , K and D by replacing summation with integration and probability with probability density.

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 $W : K \times D \rightarrow \mathbb{R}$.

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

COMMENTS ON THE BAYESIAN DECISION (1)

Bayesian recognition is decision-making, where

- ◆ Decisions do not influence the state of nature (c.f., game theory, control theory).
- ◆ A single decision is made, issues of time are ignored in the model (unlike in control theory, where decisions are typically taken continuously and in real-time).
- ◆ The cost of obtaining measurements is not modelled (unlike in sequential decision theory).

COMMENTS ON THE BAYESIAN DECISION (2)

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 the petrol price) and will be observed later.

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

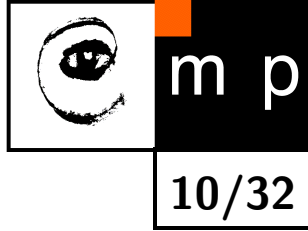
Classification is a special case of the decision-making problem where the set of decisions D and hidden states K coincide.

GENERALITY OF THE BAYESIAN FORMULATION

Observation x can be a number, symbol, function of two variables (e.g., an image), graph, algebraic structure, etc.

Application	Measurement	Decisions
value of a coin in a slot machine	$x \in \mathbb{R}^n$	value
optical character recognition	2D bitmap, intensity image	words, numbers
license plate recognition	gray-level image	characters, numbers
fingerprint recognition	2D bitmap, gray-level image	personal identity
speech recognition	$x(t)$	words
EEG, ECG analysis	$\bar{x}(t)$	diagnosis
forfeit detection	various	{yes, no}
speaker identification	$x(t)$	personal identity
speaker verification	$x(t)$	{yes, no}

TWO GENERAL PROPERTIES OF BAYESIAN STRATEGIES



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

TWO PARTICULAR BAYESIAN TASKS

1. Probability of the wrong estimate of the state.

In most cases, the pattern recognition task is to estimate the state of an object. This means that a set of decisions D and a set of states K are the same.

2. Decision with the reject option, i.e., not known.

The task is to estimate the state of an object with a high confidence or to reject the decision.

PROBABILITY OF THE WRONG ESTIMATE OF THE STATE (1)



The decision $q(x) = k$ means that an object is in the state k . The estimate $q(x)$ not always is equal to the actual state k^* . Thus the probability of the wrong decision $q(x) \neq k^*$ is required to be as small as possible.

Consider a **unit penalty** when the situation $q(x) \neq k^*$ occurs and it does not pay any penalty otherwise, i.e.,

$$W(k^*, q(x)) = \begin{cases} 0 & \text{if } q(x) = k^* , \\ 1 & \text{if } q(x) \neq k^* . \end{cases}$$

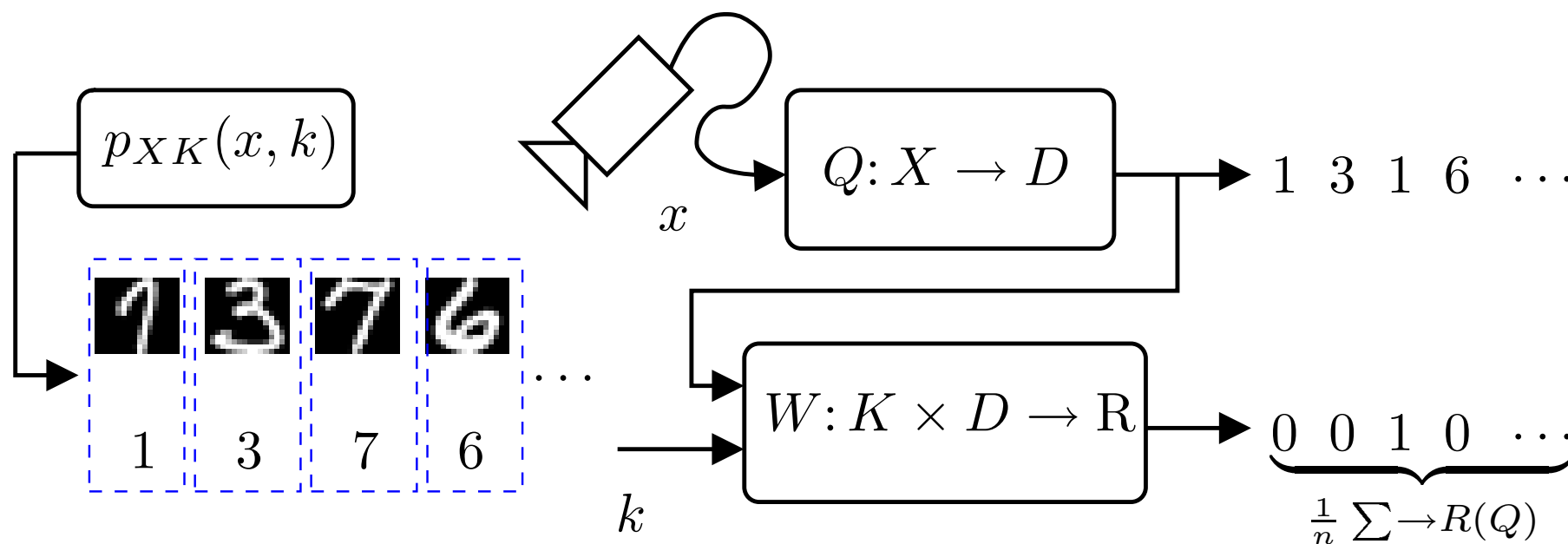
This means that the Bayesian risk

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

is thus the probability of the situation $q(x) \neq k^*$.

EXAMPLE, OPTICAL CHARACTER RECOGNITION

Illustration of the Bayesian setting:



X set of all possible intensity images.

K set of numerals $\{0, 1, \dots, 9\}$.

D equals to K , i.e., decision assigns images to classes.

W 0/1-loss function $W(k, q(x)) = \begin{cases} 0 & \text{if } k = q(x), \\ 1 & \text{if } k \neq q(x). \end{cases}$

PROBABILITY OF THE WRONG ESTIMATE OF THE STATE (2)

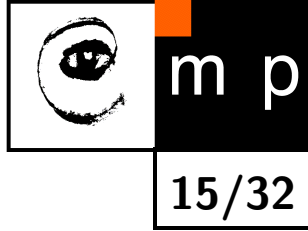


We have to determine the strategy $q: X \rightarrow K$ which minimizes the risk, i.e.,

$$\begin{aligned} q(x) &= \operatorname{argmin}_{k \in K} \sum_{k^* \in K} p_{KX}(x, k^*) W(k^*, k) \\ &= \operatorname{argmin}_{k \in K} \sum_{k^* \in K} p_{K|X}(k^* | x) W(k^*, k) \\ &= \operatorname{argmin}_{k \in K} \sum_{k^* \in K \setminus \{k\}} p_{K|X}(k^* | x) \\ &= \operatorname{argmin}_{k \in K} \left(\sum_{k^* \in K} p_{K|X}(k^* | x) - p_{K|X}(k | x) \right) \\ &= \operatorname{argmin}_{k \in K} (1 - p_{K|X}(k | x)) = \operatorname{argmax}_{k \in K} p_{K|X}(k | x). \end{aligned}$$

The result is that the *a posteriori* probability of each state k is to be calculated for the observation x and it is to be decided in favor of the most probable state.

BAYESIAN STRATEGY WITH REJECT OPTION (1)

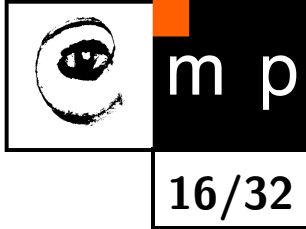


Let us introduce conditional mathematical expectation of the penalty by $R(x, d)$, called a **partial risk** (also conditional risk),

$$R(x, d) = \sum_{k \in K} p_{K|X}(k | x) W(k, d).$$

- ◆ Bayesian risk equals $R(q) = \sum_{x \in X} p_X(x) R(x, q(x))$.
- ◆ Decision $d = q(x)$ has to correspond to minimal $R(x, d)$.
- ◆ Sometimes this minimum will be quite large and the resulting decision should be **not known**.
- ◆ Decision **not known** is given if the observation x does not contain enough information to decide with a small risk.

BAYESIAN STRATEGY WITH REJECT OPTION (2)



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

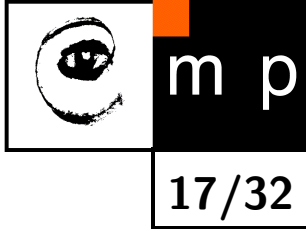
Let us determine penalties $W(k, d)$, $k \in K$, $d \in D$:

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

Find the Bayesian strategy $q: X \rightarrow D$. The decision $q(x)$ corresponding to the observation x has to minimise the partial risk,

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

BAYESIAN STRATEGY WITH REJECT OPTION (3)



Equivalent definition of the Bayesian strategy

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

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

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

BAYESIAN STRATEGY WITH REJECT OPTION (4)

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

$$\begin{aligned} R(x, \text{not known}) &= \sum_{k^* \in K} p_{K|X}(k^* | x) W(k^*, \text{not known}) \\ &= \sum_{k^* \in K} p_{K|X}(k^* | x) \varepsilon = \varepsilon . \end{aligned}$$

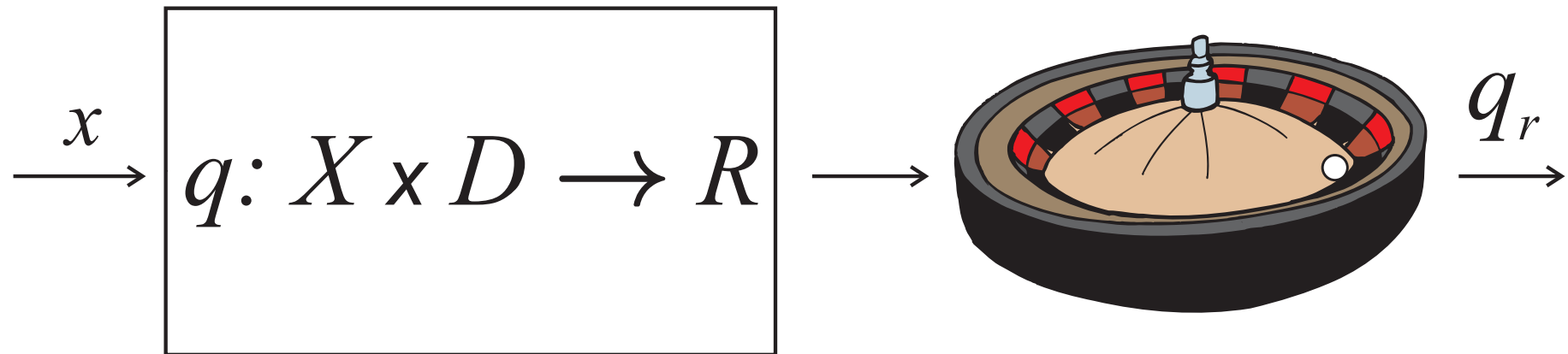
The decision rule becomes

$$q(x) = \begin{cases} \operatorname{argmax}_{k \in K} p_{K|X}(k | x), & \text{if } 1 - \max_{k \in K} p_{K|X}(k | x) < \varepsilon, \\ \text{not known}, & \text{if } 1 - \max_{k \in K} p_{K|X}(k | x) \geq \varepsilon. \end{cases}$$

Strategy $q(x)$ in words:

- ◆ The state k has to be found which has the largest *a posteriori* probability.
- ◆ If this probability is larger than $1 - \varepsilon$ then it is decided in favour of the state k .
- ◆ If its probability is not larger than $1 - \varepsilon$ then the decision not known is provided.

DOES IT MAKE SENSE TO RANDOMIZE BAYESIAN STRATEGIES ?



No!

A deterministic strategy is never worse than a randomized one.

BAYESIAN STRATEGIES ARE DETERMINISTIC

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

Theorem

Let X, K, D be finite sets, $p_{XK}: X \times K \rightarrow \mathbb{R}$ be a probability distribution, $W: K \times D \rightarrow \mathbb{R}$ be a penalty function. Let $q_r: D \times X \rightarrow \mathbb{R}$ be a stochastic strategy. Its risk is

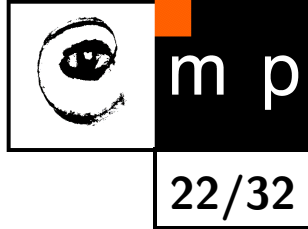
$$R_{\text{rand}} = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \sum_{d \in D} q_r(d|x) W(k, d).$$

In such a case, there exists the deterministic (Bayesian) strategy $q: X \rightarrow D$ with the risk

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

which is not greater than R_{rand} .

PROOF (Bayesian strategy, its deterministic character)



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

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

$$R_{\text{rand}} \geq \sum_{x \in X} \min_{d \in D} \sum_{k \in K} p_{XK}(x, k) W(k, d) \quad \text{holds for all } x \in X, d \in D. \quad (1)$$

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

$$\sum_{k \in K} p_{XK}(x, k) W(k, q(x)) = \min_{d \in D} \sum_{k \in K} p_{XK}(x, k) W(k, d). \quad (2)$$

The function $q: X \rightarrow D$ 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 (2) into the inequality (1) then we obtain the inequality

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

Right hand side gives the risk of the deterministic strategy q . $R_{\text{det}} \leq R_{\text{rand}}$ holds.

SPECIAL CASE, 2 STATES, LIKELIHOOD RATIO

- ◆ 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.
- ◆ The strategy cannot be arbitrary and should have certain properties.

LIKELIHOOD RATIO (2)

If the *a priori* probabilities $p_K(k)$ and the penalty $W(k, d)$ were known then the decision $q(x)$ about the observation x ought to be

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

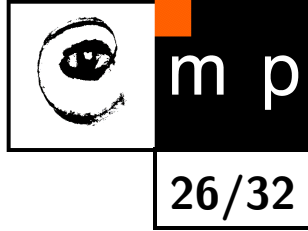
$$\gamma(x) = \frac{p_{X|1}(x)}{p_{X|2}(x)} - \text{likelihood ratio.}$$

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 (4)



Note:

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

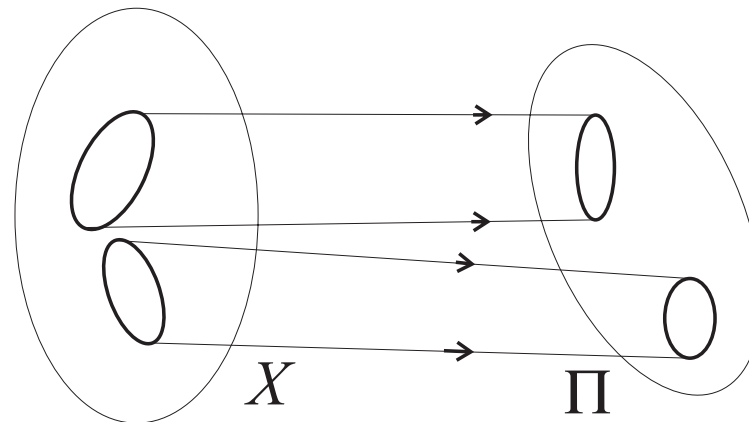
Any Bayesian strategy divides the real axis from 0 to ∞ into $|D|$ intervals $I(d)$, $d \in D$. The decision d is made for observation $x \in X$ when the likelihood ratio $\gamma = p_{X|1}(x)/p_{X|2}(x)$ belongs to the interval $I(d)$.

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 .

SPACE OF PROBABILITIES Π , AN IDEA

Consider a $|K|$ -dimensional linear space Π .

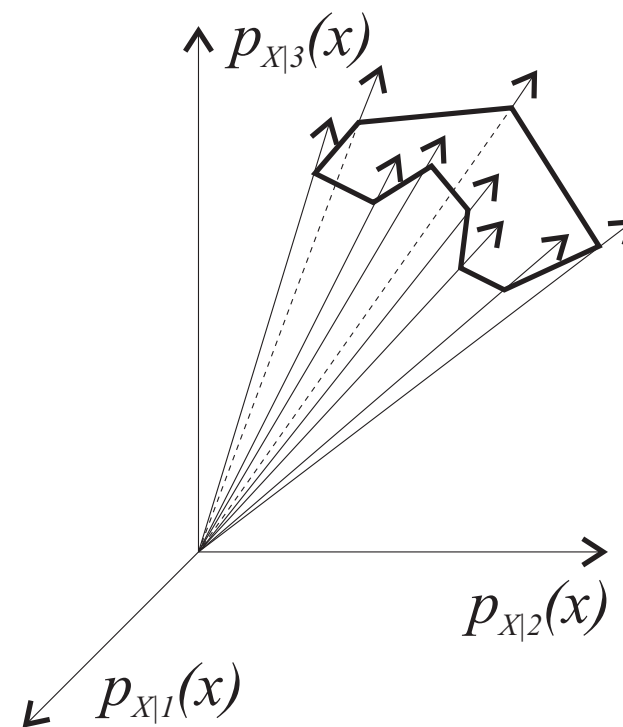
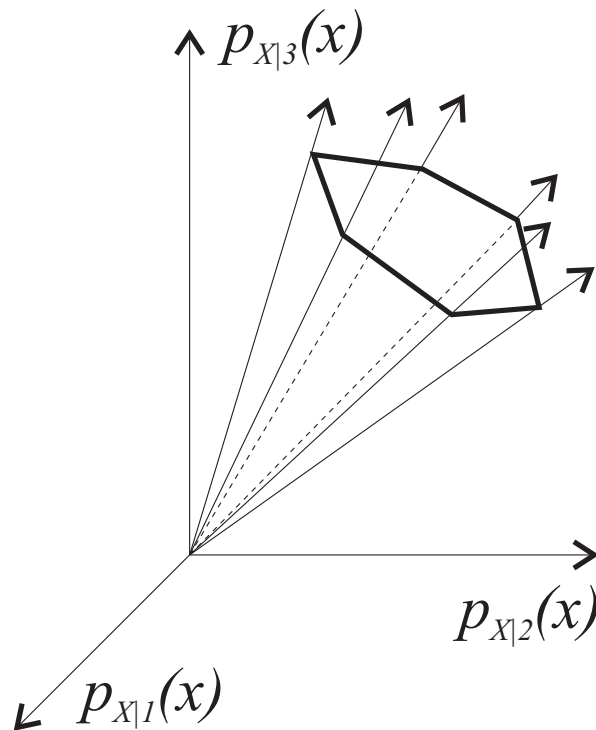


- ◆ The space of probabilities Π has coordinate axes given by probabilities $p_{X|1}(x)$, $p_{X|2}(x)$, \dots (in general $p_{X|k}(x)$, $k \in K$).
- ◆ The set of observations X is mapped into a positive hyperquadrant of Π . The observation $x \in X$ maps to the point $p_{X|k}(x)$, $k \in K$.
- ◆ An interesting question: Into what does the whole subset $X(d)$, $d \in D$, of the observation space corresponding to one decision map in the space of probabilities Π ?

CONE, CONVEX CONE

The subset $\Pi' \subset \Pi$ is called a **cone** if $\alpha \pi \in \Pi'$ for $\forall \pi \in \Pi'$ and for $\forall \alpha \in \mathbb{R}$, $\alpha > 0$.

If the subset Π' is a cone and, in addition, it is convex then it is called a **convex cone**.



GENERAL CASE FOR $K > 2$, CONVEX CONES

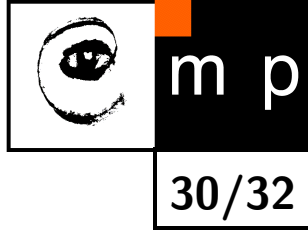
Theorem:

Let X, K, D be three finite sets and let $p_{XK}: X \times K \rightarrow \mathbb{R}$, $W: K \times D \rightarrow \mathbb{R}$ be two functions. Let $\pi: X \rightarrow \Pi$ be a mapping of the set X into a $|K|$ -dimensional linear space Π (**space of probabilities**); $\pi(x) \in \Pi$ is a point with coordinates $p_{X|k}(x)$, $k \in K$.

Let any decomposition of the positive hyperquadrant of the space Π into $|D|$ **convex cones** $\Pi(d)$, $d \in D$, define the strategy q for which $q(x) = d$ if and only if $\pi(x) \in \Pi(d)$. Then a decomposition $\Pi^*(d)$, $d \in D$, exists such that corresponding strategy q^* minimises a Bayesian risk

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

PROOF, Convex shape of classes in Π (1)



Let us create such cones. Enumerate decision $d \in D$ by numbers $n(d)$

$$\sum_{k \in K} p_{X|K}(x) p_K(k) W(k, d^*) \leq \sum_{k \in K} p_{X|K}(x) p_K(k) W(k, d), \quad n(d) < n(d^*),$$

$$\sum_{k \in K} p_{X|K}(x) p_K(k) W(k, d^*) < \sum_{k \in K} p_{X|K}(x) p_K(k) W(k, d), \quad n(d) > n(d^*).$$

PROOF, Convex shape of classes in Π (2)

Let us use coordinates in Π , $\pi_k = p_{X|k}(x)$. The point π with coordinates π_k , $k \in K$, has to be mapped into the set $\Pi(d^*)$, if

$$\sum_{k \in K} \pi_k p_K(k) W(k, d^*) \leq \sum_{k \in K} \pi_k p_K(k) W(k, d), \quad n(d) < n(d^*),$$

$$\sum_{k \in K} \pi_k p_K(k) W(k, d^*) < \sum_{k \in K} \pi_k p_K(k) W(k, d), \quad n(d) > n(d^*).$$

The set expressed in such a way is a cone, because if the point with coordinates π_k , $k \in K$, satisfies the inequalities then any point with coordinates $\alpha \pi_k$, $\alpha > 0$, satisfies the system too.

The system of inequalities is linear with respect to variables π_k , $k \in K$, and thus the set of its solutions $\Pi(d)$ is convex.

IMPORTANCE OF LINEAR CLASSIFIERS. WHY?

- ◆ **Theoretical importance**, decomposition of the space of probabilities 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 $n + 1$. 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.