Bayesian decision making

Václav Hlaváč

Czech Technical University in Prague
Czech Institute of Informatics, Robotics and Cybernetics
160 00 Prague 6, Jugoslávských partyzánů 1580/3, Czech Republic
http://people.ciirc.cvut.cz/hlavac, vaclav.hlavac@cvut.cz
also Center for Machine Perception, http://cmp.felk.cvut.cz

Courtesy: M.I. Schlesinger

Outline of the talk:

- Bayesian task formulation.
- Two general properties of the Bayesian task.
- Probability of the wrong estimate.
- Reject option.
- Non-random strategy is the best.
- Linear separability in space of class-conditioned probabilities; convex cones.

Notation remarks: Joint and conditional probabilities

- The joint probability $p_{XY}(x, y)$ can be expressed as $p_{XY}(x, y) = p_{X|y}(x|y) \cdot p_Y(y)$.
  $p_Y(y)$ is called the a priori probability of the hidden state y or simply as a prior of y.
- The standard notation for the joint probability $p(x, y)$ and for the conditional probability $p(x|y)$ is ambiguous.
  - Are $p(x, y)$ and $p(x|y)$ numbers, functions of a single variable or functions of two variables?
  - Let us disambiguate the notation using subscripts:
    $p_{XY}(x, y)$ be a *function of two variables*,
    $p_{X|y}(x|y)$ be a *function of a single variable* x,
    and $p_{xy}(x, y)$ be a *single real number*.

Bayesian decision making, concepts

Object (situation) is described by two parameters:

- x is an observable feature (observation).
- y is an unobservable hidden parameter (state).
- X is a finite set of observations, $x \in X$.
- Y is a finite set of hidden states, $y \in Y$.
- D is a finite set of possible decisions $d \in D$.

- p_{XY}: $X \times Y \to \mathbb{R}$ is the statistical model (the joint probability) that the object is in the state y and the observation x is made.
- W: $Y \times D \to \mathbb{R}$ is the *penalty function*, $W(y, d)$, $y \in Y$, $d \in D$ is the penalty paid in for the object in the state y and the decision d made.
- q: $X \to D$ is the *decision function* (rule, strategy) assigning to each $x \in X$ the decision $q(x) \in D$.
- $R(q)$ is the risk, i.e. the mathematical expectation of the penalty.

Definition of the Bayesian decision making task

- Given: sets X, Y and D, a joint probability $p_{XY}\colon X \times Y \to \mathbb{R}$ and function $W\colon Y \times D \to \mathbb{R}$
- Task: The Bayesian task of the statistical decision making task seeks a strategy $q\colon X \to D$ which minimizes the Bayesian risk

$$R(q) = \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y)\, W(y, q(x))\,.$$

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

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

Expressing Bayesian risk as a weighted sum of partial risks

$$R(q^*) = \min_{q \in X \to D} \sum_{x \in X} \sum_{y \in y} p_{XY}(x, y)\, W(y, q(x))$$

$$R(q^*) = \sum_{x \in X} \min_{q(x) \in D} \sum_{y \in Y} p_{XY}(x, y)\, W(y, q(x))$$

$$R(q^*) = \sum_{x \in X} \min_{q(x) \in D} p(x) \sum_{y \in Y} p_{Y|X}(y|x)\, W(y, q(x))$$ (*posterior conditional probability in a yellow box*)

$$R(q^*) = \sum_{x \in X} p(x)\, R(x, d^*)$$

where

$R(x, d^*) = \sum_{y \in Y} p_{Y|X}(y \mid x)\, W(y, d^*)$ is the partial risk,

i.e., the posterior conditional (conditioned on x) mathematical expectation of the penalty;
$R(x, d^*) \leq R(x, d), d \in D$, i.e., $q^*(x) = d^*$.

The problem of finding the optimal strategy $q(x)$ can be thus solved "observation-wise", i.e. by finding the optimal decision for each observation x (a point in the feature space).

Comments on the Bayesian decision making (1)

In the Bayesian decision making (aka recognition):

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

Comments on the Bayesian decision making (2)

The hidden parameter y (the class information) is considered not observable.

Common situations are:

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

It is interesting to ponder whether a state can ever be genuinely unobservable (cf. Schrödinger's cat).

Classification is a special case of the decision-making problem, in which the set of decisions D and hidden states Y coincide.

Example: two hidden states, three decisions

- Object: a patient examined by a physician.
- Observations (some measurable parameters): $X = \{\text{temperature, blood pressure, } \ldots\}$.
- Two unobservable states: $Y = \{\text{healthy, sick}\}$.
- Three decisions: $D = \{\text{do not cure, weak medicine, strong medicine}\}$.
- The penalty function: $W : Y \times D \to \mathbb{R}$.

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

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	$x \epsilon \mathbb{R}^n$	value
optical character recognition	2D bitmap, gray-level 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. A deterministic strategy is never worse than a randomized one.
2. Each Bayesian strategy corresponds to a separation of the space of class-conditioned probabilities *(to be explained later)* into convex subsets.

Before proving the above mentioned general properties, let us explain two practically useful special cases of Bayesian decision making tasks.

Two particularly useful special Bayesian tasks

1. Minimising the classification error (= minimising the probability of the wrong state estimate) is likely the most commonly solved patter recognition task.

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

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

   The task estimates the state of an object with a prescribed (high) confidence or rejects to decide.

Probability of the wrong estimate of the state (1)

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

A unit penalty is paid when the wrong decision $q(x) \neq y^*$ occurs. No penalty is paid otherwise,

$$W\big(y^*, q(x)\big) = \begin{cases} 0 & \text{if} \quad q(x) = y^*, \\ 1 & \text{if} \quad q(x) \neq y^*. \end{cases}$$ is called a 0/1-loss function.

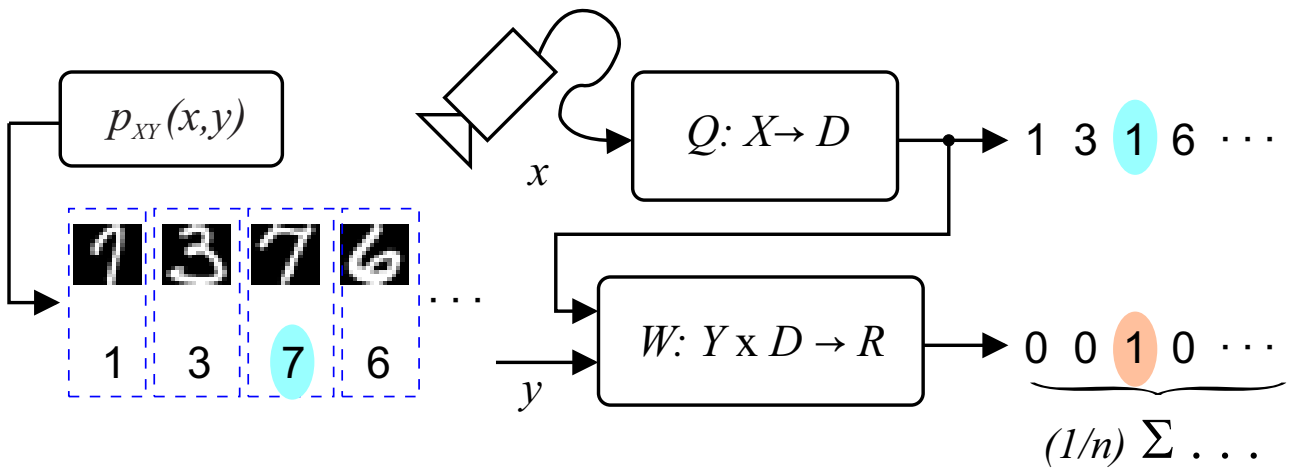
Bayesian risk

$$R(q) = \sum_{x \in X} \sum_{y^* \in Y} p_{XY}(x, y^*)\, W\big(y^*, q(x)\big)$$

becomes the probability of the wrong estimate of the state $q(x) \neq y^*$.

Example: optical character recognition

Illustration of the Bayesian setting:



- X – set of all possible intensity images.
- Y – set of numerals $\{0, 1, \ldots, 9\}$.
- D – equals to Y, i.e., decision assigns images to classes.
- W – $0/1$-loss function $= W(y, q(x)) = \begin{cases} 0 & \text{if} \quad y = q(x)\,, \\ 1 & \text{if} \quad y \neq q(x)\,. \end{cases}$

Probability of the wrong estimate of the state (2)

We have to determine the strategy $q\colon X \to Y$, which minimizes the risk, i.e.,

$$
\begin{aligned}
q(x) &= \operatorname*{argmin}_{y\in Y} \sum_{y^*\in Y} p_{XY}(x\,, y^*)\; W(y^*, y) = \operatorname*{argmin}_{y\in Y} \sum_{y^*\in y} p(x)\; p_{Y|X}(y^* \mid x)\; W(y^*, y) \\
&= \operatorname*{argmin}_{y\in Y} \sum_{y^*\in y} p_{Y|X}(y^* \mid x)\; W(y^*, y) = \operatorname*{argmin}_{y\in Y} \sum_{y^*\in Y\setminus\{y\}} p_{Y|X}(y^* \mid y) \\
&= \operatorname*{argmin}_{y\in Y} \left(\sum_{y^*\in Y} p_{Y|X}(y^* \mid x) - p_{Y|X}(y \mid x) \right) \\
&= \operatorname*{argmin}_{y\in Y} \left(1 - p_{Y|X}(y \mid x)\right) = \operatorname*{argmax}_{y\in Y}\; p_{Y|X}(y \mid x)\,.
\end{aligned}
$$

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

Motivation: Bayesian strategy with the reject option

- Consider an examination. There are three possible answers for each question: yes, no, `not known`. If student's answer is correct, one point is added to her/his score. If the answer is wrong, three points are subtracted. If the answer is `not known`, student score remains unchanged.

  What is the optimal Bayesian strategy if the student knows the probabilities for each question that $p(yes)$ 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, $\varepsilon \in (0, 1)$, since $\varepsilon > 1$ would mean it is always better to guess than to say `not known`. $\varepsilon < 0$ would mean that saying `not known` is preferred to giving the correct answer.

Bayesian strategy with the reject option (1)

Recall (cf. slide 5) the conditional mathematical expectation of the penalty $R(x, d)$, called a partial risk (also a posterior conditional risk) given the observation x,

$$R(x, d) = \sum_{y \in Y} p_{Y|X}(y \,|\, x)\; W(y, d)\,.$$

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

Bayesian strategy with the reject option (2)

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

Let us set penalties $W(y, d)$, $y \in Y$, $d \in D$:

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

Task: Find the Bayesian strategy $q\colon X \to D$ such that the decision $q(x)$ corresponding to the observation x has to minimize the partial risk,

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

Bayesian strategy with the reject option (3)

The equivalent definition of the Bayesian strategy

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

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

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

Bayesian strategy with the reject option (4)

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

$$R(x, \texttt{not known}) = \sum_{y^* \in Y} p_{Y|X}(y^* \,|\, x)\, W(y^*, \texttt{not known})$$
$$= \sum_{y^* \in Y} p_{Y|X}(y^* \,|\, x)\, \varepsilon = \varepsilon\,.$$

The decision rule becomes

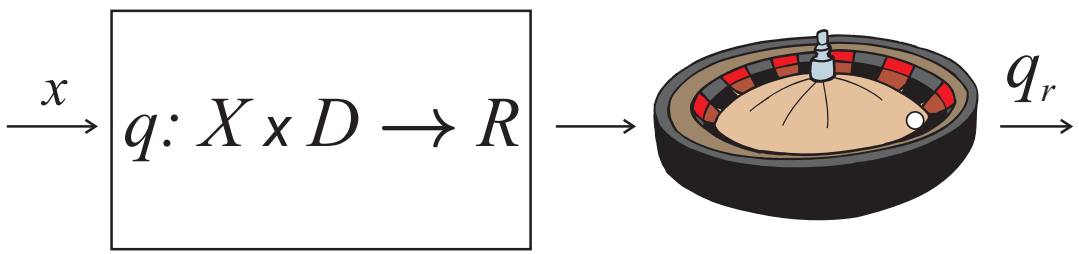
$$q(x) = \begin{cases} \operatorname*{argmax}_{y \in Y} p_{Y|X}(y \,|\, x)\,, & \text{if } 1 - \max_{y \in Y} p_{Y|X}(y \,|\, x) < \varepsilon\,, \\ \texttt{not known}\,, & \text{if } 1 - \max_{y \in Y} p_{Y|X}(y \,|\, x) \geq \varepsilon\,. \end{cases}$$

Bayesian strategy with the reject option (5)

Bayesian strategy with the reject option $q(x)$ in words:

- The state y has to be found which has the largest *posterior* probability.
- If this posterior probability is larger than $1 - \varepsilon$ then it is decided in favor of the state y.
- If this posterior probability is not larger than $1 - \varepsilon$ then the decision `not known` is provided.

Q: Does it make sense to randomize Bayesian strategies?



Answer: No!

A deterministic strategy is never worse than a randomized one.

Bayesian strategies are deterministic

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

Theorem

Let X, Y, D be finite sets, $p_{XY}\colon X \times Y \to \mathbb{R}$ be a probability distribution, $W\colon Y \times D \to \mathbb{R}$ be a penalty function. Let $q_r\colon D \times X \to \mathbb{R}$ be a stochastic strategy. Its risk is

$$R_{\text{rand}} = \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y) \sum_{d \in D} q_r(d \,|\, x)\, W(y, d)\,.$$

In such a case, there exist a deterministic (Bayesian) strategy $q\colon X \to D$ with the risk

$$R_{\text{det}} = \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y)\, W\big(y, q(x)\big)$$

which is not greater than R_{rand}.

Proof (Bayesian strategy is deterministic)

$$R_{\text{rand}} = \sum_{x \in X} \sum_{d \in D} q_r(d \,|\, x) \sum_{y \in Y} p_{XY}(x, y)\, W(y, d)\,.$$

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

$$R_{\text{rand}} \geq \sum_{x \in X} \min_{d \in D} \sum_{y \in Y} p_{XY}(x, y)\, W(y, d) \quad \text{holds for all } x \in X,\ d \in D\,. \tag{1}$$

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

$$\sum_{y \in Y} p_{XY}(x, y)\, W\big(y, q(x)\big) = \min_{d \in D} \sum_{y \in Y} p_{XY}(x, y)\, W(y, d)\,. \tag{2}$$

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 (2) into the inequality (1) then we obtain the inequality

$$R_{\text{rand}} \geq \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y)\, W\big(y, q(x)\big)\,.$$

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

A special case: two states only, likelihood ratio

- Hidden state assumes two values only, $Y = \{1, 2\}$.
- Only conditional probabilities $p_{X|1}(x)$ and $p_{X|2}(x)$ are known.
- The *a priori* probabilities $p_Y(1)$ and $p_Y(2)$ and penalties $W(y, d)$, $y \in \{1, 2\}$, $d \in D$, are not known.
- In this situation, the Bayesian strategy cannot be created.
- Nevertheless, the strategy cannot be an arbitrary one and should comply to certain constraints.

Likelihood ratio (2)

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

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

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

Likelihood ratio (3) – 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 numerical intervals.

Likelihood ratio (4)

Note:

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

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 $I(d)$.

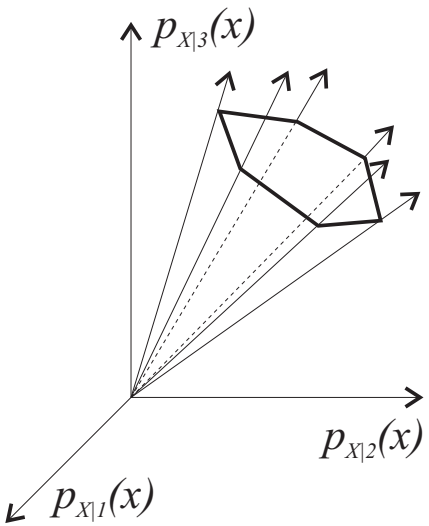
A more particular case, which is commonly known:

Two decisions only, $D = \{1, 2\}$. Bayesian strategy is characterized 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.

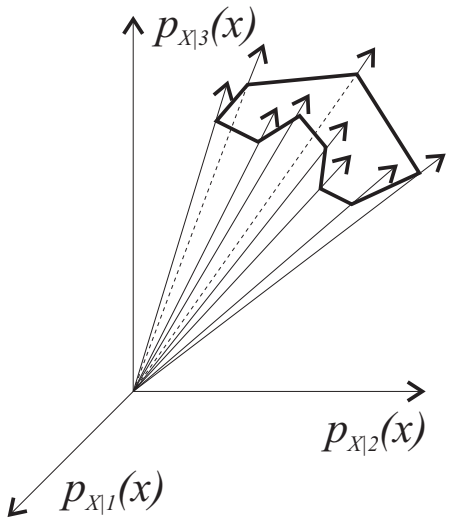
Cone, convex cone

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.



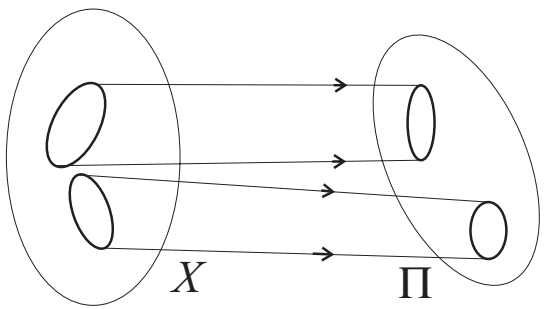
convex cone



non-convex cone

Space of class conditioned probabilities π, an idea

- Consider a $|Y|$-dimensional linear space Π which we call the space of class-conditioned probabilities.



- The space of class-conditioned probabilities Π has coordinate axes given by probabilities $p_{X|1}(x)$, $p_{X|2}(x)$, ... (in general $p_{X|y}(x)$, $y \in Y$).
- The set of observations X is mapped into a positive hyperquadrant of Π. The observation $y \in Y$ maps to the point $p_{X|y}(x)$, $y \in Y$.
- An interesting question: Where does the whole subset $X(d)$, $d \in D$, of the observation space corresponding to individual decisions maps in the space of class conditioned probabilities Π?

A general case for y convex cones, $y > 2$

Theorem:

Let X, Y, D be finite sets and let $p_{XY} : X \times Y \to \mathbb{R}$, $W : Y \times D \to \mathbb{R}$ be two functions. Let $\pi : X \to \Pi$ be a mapping of the set X into a $|Y|$-dimensional linear space Π (space of class-conditioned probabilities); $\pi(x) \in \Pi$ is a point with coordinates $p_{X|y}(x)$, $y \in Y$.

Any decomposition of the positive hyperquadrant of the space Π into $|D|$ convex cones $\Pi(d)$, $d \in D$, defines a 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^* minimizes a Bayesian risk

$$\sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y)\, W\left(y, q(x)\right) \ .$$

Proof: Convex shape of classes in Π (1)

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

$$\sum_{y \in Y} p_{X|Y}(x)\, p_Y(y)\, W(y, d^*) \leq \sum_{y \in Y} p_{X|Y}(x)\, p_Y(y)\, W(y, d)\,, \qquad n(d) < n(d^*)\,,$$

$$\sum_{y \in Y} p_{X|Y}(x)\, p_Y(y)\, W(y, d^*) < \sum_{y \in Y} p_{X|Y}(x)\, p_Y(y)\, W(y, d)\,, \qquad n(d) > n(d^*)\,.$$

Both equations express strategies d^* with the minimal risk. The equations assure that the optimal strategies are at the end of the numbering series.

Proof: Convex shape of classes in Π (2)

Let us use coordinates in Π, $\pi_y = p_{Y|y}(x)$. The point π with coordinates π_y, $y \in Y$, has to be mapped into the set $\Pi(d^*)$, if

$$\sum_{y \in Y} \pi_y \, p_Y(y) \, W(y, d^*) \leq \sum_{y \in Y} \pi_y \, p_Y(y) \, W(y, d) \,, \qquad n(d) < n(d^*) \,,$$

$$\sum_{y \in Y} \pi_y \, p_Y(y) \, W(y, d^*) < \sum_{y \in Y} \pi_y \, p_Y(y) \, W(y, d) \,, \qquad n(d) > n(d^*) \,.$$

- The set expressed in such a way is a cone, because if the point with coordinates π_y, $y \in Y$, satisfies the inequalities then any point with coordinates $\alpha \, \pi_y$, $\alpha > 0$, satisfies the system too.
- The system of inequalities is linear with respect to variables π_y, $y \in Y$, and thus the set of its solutions $\Pi(d)$ is convex.

Importance of linear classifiers. Why?

- Theoretical importance, decomposition of the space of class-conditioned probabilities into convex cones. $\Rightarrow$ Linear classifiers are important.
- Efficient algorithms exist to implement linear classification tasks and learn them from training data empirically.
- For some statistical models, the Bayesian or a few non-Bayesian strategies are implemented by linear discriminant functions.
- Some non-linear discriminant functions can be implemented as linear ones after straightening the feature space (globally or locally by kernel functions).
- 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 a finite training multiset does not differ much from the correct strategy found for a statistical model.