Bayesian Decision Theory

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

authors: Václav Hlaváč, Jiri Matas

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

Details in Chapter 1 of
Schlesinger M.I., Hlaváč V.: *Ten lectures on statistical and structural pattern recognition.* Kluwer Academic Publisher, Dordrecht, The Netherlands, 2002, 519 p.

Bayesian Decision Making. Basic Concepts.

An *object* (situation) is described by two parameters:

x which is observable; called observation, measurement, or "feature vector".

k which is unobservable; called hidden parameter, state, state-of-nature or "class".

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

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

D is a finite set of possible *decisions* (actions).

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

W: $K \times D \to \mathbb{R}$ is a *penalty (loss) 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 $\to D$ is a *decision function* (rule, strategy) assigning for each $x \in X$ the decision $q(x) \in D$.

Formulation of the Bayesian Decision Problem.

Given the sets X, K and D, the joint probability $p_{XK}\colon X \times K \to \mathbb{R}$ and the penalty function $W\colon K \times D \to \mathbb{R}$, find the strategy $q\colon X \to D$ which minimises the expectation of $W(k, q(x))$:

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

The quantity $R(q)$ is called the the Bayesian risk. The solution to the Bayesian problem is the Bayesian strategy q^* minimizing the Bayesian risk.

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

Notes:

• The probability $p_{XK}(x, k)$ is often expressed as $p_{XK}(x, k) = p_{Xk}(x|k) * p_K(k)$

• 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? 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.

Expressing the Bayes risk via the partial risk.

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

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

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

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

where

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

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

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.

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 task formulation.

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

Application	Measurement	Decisions
value of a coin in a slot machine	$x \epsilon R^n$	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:

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

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)

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

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

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

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

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

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

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

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.

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 $W(k,d)$ were known then the decision $q(x)$ about the observation x ought to be

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

$\gamma(x)$ – 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

Space of probabilities Π

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

The space Π has coordinate axes given by probabilities $p_{X|1}(x)$, $p_{X|2}(x)$, ... (in general $p_{X|k}(x)$, $k \in K$).

The set of observations X is mapped into positive hyperquadrant of Π. The observation $x \in X$ maps to the point $p_{X|k}(x)$, $k \in K$.

The interesting question how the whole subset $X(d)$, $d \in D$, of the observation space corresponding to one decision maps to 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.

Convex Cones: the general case $K > 2$

Theorem:

Let X, K, D be three finite sets and let $p_{XK}\colon X \times K \to \mathbb{R}$, $W\colon K \times D \to \mathbb{R}$ be two functions. Let $\pi\colon X \to \Pi$ 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

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

PROOF, Convex shape of classes in Π (1)

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

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

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

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

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

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

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.

- 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.

Two special cases of Bayesian Problems.

1. Minimisation of the probability of the incorrect estimation of the hidden state (i.e. minimisation of classification error) is one the most common recognition problems. We show that it is a special case of Bayes risk minimisation.

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

Minimisation of the probability of incorrect decision. (1)

Consider the following problem:

- The object is in an unknown state k.
- The set of possible decisions D and of hidden states K coincide, $D = K$.
- The cost function assigns a unit penalty when $q(x) \neq k$ occurs and no penalty otherwise, i.e.

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

The Bayesian risk

$$\begin{aligned} R(q) &= \sum_{x \in X} \sum_{k \in K} p_{XK}(x,k)\, W\big(k, q(x)\big) = \sum_{x \in X} p_X(x) \sum_{k \neq q(x)} p_{Kx}(k|x) \\ &= \sum_{x \in X} p_X(x)(1 - p_{xk}(q(x)|x) \end{aligned}$$

is then equal the probability of the situation $q(x) \neq k$ (probability of classification error) or 1 – probability of correct decision.

Minimisation of the probability of incorrect decision. (2)

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

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

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 $g(x) = p_{k|x}(1\,|\,x) - p_{k|x}(0\,|\,x)$.

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 minimises 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 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 $R(x, \texttt{not known})$

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

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` .