Pattern Recognition. Bayesian and non-Bayesian Tasks.

Petr Pošík

Czech Technical University in Prague

Faculty of Electrical Engineering

Dept. of Cybernetics

This lecture is based on the book
Ten Lectures on Statistical and Structural Pattern Recognition
by Michail I. Schlesinger and Václav Hlaváč (Kluwer, 2002).
(V české verzi kniha vyšla pod názvem
Deset přednášek z teorie statistického a strukturálního rozpoznávání
ve vydavatelství ČVUT v roce 1999.)

Pattern Recognition

Examples of pattern recognition tasks

- ✔ Given a gene expression profile of a human tissue, decide whether the tissue is cancerous or not.
- ✔ Given an ECG signal, decide whether the person is threatened with a heart attack.
- ✔ Given results of several medical tests, decide if the patient needs no cure, if she get a weak drug, or a strong drug.
- ✔ Given a speech recording, decide what word was pronounced.
- ✔ Given a picture, decide if the picture contains a human face.
- ✔ Given a picture, decide if it was taken in spring, summer, autumn, or in winter.
- ✔ Given several (2D) pictures of the same scene taken from different angles, decide the (3D) position of an object in the space.
- ✔ Given a sentence with missing characters, decide what the missing characters are.
- ✔ . . .

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision function (or strategy) $q : X \rightarrow D$

- ✔ D is a set of possible decisions. (Very often $D = K$.)
- ✔ q is a function that assigns a decision $q(x) \in D$ to each $x \in X$.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \to \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision function (or strategy) $q : X \to D$

- ✔ D is a set of possible decisions. (Very often $D = K$.)
- ✔ q is a function that assigns a decision $q(x) \in D$ to each $x \in X$.

Penalty function (or loss function) $W : K \times D \to \mathcal{R}$ (real numbers)

- ✔ $W(k, d)$ is a penalty for decision d if the object is in state k.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \to \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision function (or strategy) $q : X \to D$

- ✔ D is a set of possible decisions. (Very often $D = K$.)
- ✔ q is a function that assigns a decision $q(x) \in D$ to each $x \in X$.

Penalty function (or loss function) $W : K \times D \to \mathcal{R}$ (real numbers)

- ✔ $W(k, d)$ is a penalty for decision d if the object is in state k.

The quality of the decision strategy q is given by

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \to \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision function (or strategy) $q : X \to D$

- ✔ D is a set of possible decisions. (Very often $D = K$.)
- ✔ q is a function that assigns a decision $q(x) \in D$ to each $x \in X$.

Penalty function (or loss function) $W : K \times D \to \mathcal{R}$ (real numbers)

- ✔ $W(k, d)$ is a penalty for decision d if the object is in state k.

The quality of the decision strategy q is given by

- ✔ the **Bayesian risk** $R(q)$ of strategy q, i.e. by the mathematical expectation of the penalty which must be paid when using the strategy q:

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- ✔ observation $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- ✔ hidden state $k \in K$.
- ✔ k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- ✔ $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- ✔ $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision function (or strategy) $q : X \rightarrow D$

- ✔ D is a set of possible decisions. (Very often $D = K$.)
- ✔ q is a function that assigns a decision $q(x) \in D$ to each $x \in X$.

Penalty function (or loss function) $W : K \times D \rightarrow \mathcal{R}$ (real numbers)

- ✔ $W(k, d)$ is a penalty for decision d if the object is in state k.

The quality of the decision strategy q is given by

- ✔ the **Bayesian risk** $R(q)$ of strategy q, i.e. by the mathematical expectation of the penalty which must be paid when using the strategy q:

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \cdot W(k, q(x)) \tag{1}$$

Two types of pattern recognition

The description of the concepts is very general—so far we did not specify what the items of the X, K, and D sets actually are, how they are represented.

Two types of pattern recognition

The description of the concepts is very general—so far we did not specify what the items of the X, K, and D sets actually are, how they are represented.

1. **Statistical pattern recognition**
   - ✔ Objects are represented as points in a vector space.
   - ✔ The point (vector) $\mathbf{x}$ contains the individual observations (in a numerical form) as its coordinates.
2. **Structural pattern recognition**
   - ✔ The object observations contain a structure which is represented and used for recognition.
   - ✔ A typical example of the representation of a structure is *a grammar*.

Bayesian Decision Theory

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \to \langle 0, 1 \rangle$ and $W : K \times D \to \mathcal{R}$, find a strategy $q : X \to D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \to \langle 0, 1 \rangle$ and $W : K \times D \to \mathcal{R}$, find a strategy $q : X \to D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

- ✔ *Probability space*? Any observation x is mapped to a point in a $|K|$-dimensional linear space (delimited by the positive coordinates) with the coordinates $(p_{X|1}(x|1), p_{X|2}(x|2), \ldots, p_{X|k}(x|k))$.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1\rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

- ✔ *Probability space*? Any observation x is mapped to a point in a $|K|$-dimensional linear space (delimited by the positive coordinates) with the coordinates $(p_{X|1}(x|1), p_{X|2}(x|2), \ldots, p_{X|k}(x|k))$.
- ✔ *Cone*? Let S be a linear space. Any subspace $C \subset S$ is a *cone* if for each $x \in C$ also $\alpha x \in C$ for any real number $\alpha > 0$.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \to \langle 0, 1 \rangle$ and $W : K \times D \to \mathcal{R}$, find a strategy $q : X \to D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

- ✔ *Probability space*? Any observation x is mapped to a point in a $|K|$-dimensional linear space (delimited by the positive coordinates) with the coordinates $(p_{X|1}(x|1), p_{X|2}(x|2), \ldots, p_{X|k}(x|k))$.
- ✔ *Cone*? Let S be a linear space. Any subspace $C \subset S$ is a *cone* if for each $x \in C$ also $\alpha x \in C$ for any real number $\alpha > 0$.
- ✔ *Convex cone*? For any 2 points $x_1 \in C$ and $x_2 \in C$, and for any point x lying on the line between x_1 and x_2, also $x \in C$.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \to \langle 0, 1 \rangle$ and $W : K \times D \to \mathcal{R}$, find a strategy $q : X \to D$ which minimizes the Bayesian risk of the strategy q

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

The optimal strategy q is then called *the Bayesian strategy*.

Bayesian strategy is deterministic.

- ✔ q provides the same decision $d = q(x)$ for the same x, despite k may be different.
- ✔ What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- ✔ The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

- ✔ *Probability space*? Any observation x is mapped to a point in a $|K|$-dimensional linear space (delimited by the positive coordinates) with the coordinates $(p_{X|1}(x|1), p_{X|2}(x|2), \ldots, p_{X|k}(x|k))$.
- ✔ *Cone*? Let S be a linear space. Any subspace $C \subset S$ is a *cone* if for each $x \in C$ also $\alpha x \in C$ for any real number $\alpha > 0$.
- ✔ *Convex cone*? For any 2 points $x_1 \in C$ and $x_2 \in C$, and for any point x lying on the line between x_1 and x_2, also $x \in C$.
- ✔ The individual $C(d)$ are *linearly separable*!!!

Two special cases of the Bayesian decision task

Probability of error when estimating k

- ✔ The task is to decide the object state k, i.e. $D = K$.
- ✔ The goal is to minimize $Pr(q(x) \neq k)$.
- ✔ $Pr(q(x) \neq k) = R(q)$ if

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

- ✔ In this case:

$$\begin{aligned} q(x) &= \arg\min_{d \in D} \sum_{k \in K} p_{XK}(x, k) W(k, d) = \\ &= \arg\max_{d \in D} p_{K|X}(k|X), \end{aligned} \tag{2}$$

i.e. compute a posteriori probabilities of all states k given the observation x, and decide for the most probable state.

Two special cases of the Bayesian decision task

Probability of error when estimating k

- The task is to decide the object state k, i.e. $D = K$.
- The goal is to minimize $Pr(q(x) \neq k)$.
- $Pr(q(x) \neq k) = R(q)$ if

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

- In this case:

$$\begin{aligned} q(x) &= \arg\min_{d \in D} \sum_{k \in K} p_{XK}(x,k) W(k,d) = \\ &= \arg\max_{d \in D} p_{K|X}(k|X), \end{aligned} \tag{2}$$

i.e. compute a posteriori probabilities of all states k given the observation x, and decide for the most probable state.

Bayesian strategy with the `dontknow` decision

- Let's define partial risk for certain observation x and decision d as $R(x,d) = \sum_{k \in K} p_{K|X}(k|x) W(k,d)$.
- For each observation x, we shall provide the decision d minimizing $R(x,d)$.
- But even this optimal $R(x,d)$ may not be sufficiently low, i.e. x does not convey sufficient information for a low-risk decision.

- Let's use $D = K \cup \{\texttt{dontknow}\}$ and define

$$W(k,d) = \begin{cases} 0 & \text{if } d = k, \\ 1 & \text{if } d \neq k \text{ and } d \neq \texttt{dontnow} \\ \epsilon & \text{if } d = \texttt{dontknow}. \end{cases}$$

- In this case:

$$q(x) = \begin{cases} \arg\max_{k \in K} p_{K|X}(k|X) \\ \qquad \text{if } \max_{k \in K} p_{K|X}(k|X) > 1 - \epsilon, \\ \texttt{dontknow} \\ \qquad \text{if } \max_{k \in K} p_{K|X}(k|X) \leq 1 - \epsilon. \end{cases}$$

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The a priori probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The a priori probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- ✔ Important: $W(k, d) \in R$
- ✔ We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- ✔ How do your compare the following penalties:
  - ✘ games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The a priori probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- ✔ Important: $W(k, d) \in R$
- ✔ We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- ✔ How do your compare the following penalties:
  - ✘ games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee
  - ✘ system diagnostics, health diagnosis:
    false alarm (costs you some money) vs. overlooked danger (may cost you a human life)

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The a priori probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- ✔ Important: $W(k,d) \in R$
- ✔ We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- ✔ How do your compare the following penalties:
  - ✘ games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee
  - ✘ system diagnostics, health diagnosis:
    false alarm (costs you some money) vs. overlooked danger (may cost you a human life)
  - ✘ judicial error:
    to convict an innocent (huge harm for 1 innocent person) vs. to free a killer (potential harm to many innocent persons)

Limitations of the Bayesian approach (cont.)

A priori probabilities of states:

- ✔ Probabilities $p_K(k)$
  - ✘ may be unknown (then we can determine them by further study), or
  - ✘ may not exist at all (if the state k is not random).
- ✔ E.g. we observe a plane x and we want to decide if it is an enemy aircraft or not.
  - ✘ $p_{X|K}(x|k)$ may be quite complex, but known (it at least exists).
  - ✘ $p_K(k)$, however, do not exist—the frequency of enemy plane observation does not converge to any number.

Limitations of the Bayesian approach (cont.)

A priori probabilities of states:

- ✔ Probabilities $p_K(k)$
  - ✘ may be unknown (then we can determine them by further study), or
  - ✘ may not exist at all (if the state k is not random).
- ✔ E.g. we observe a plane x and we want to decide if it is an enemy aircraft or not.
  - ✘ $p_{X|K}(x|k)$ may be quite complex, but known (it at least exists).
  - ✘ $p_K(k)$, however, do not exist—the frequency of enemy plane observation does not converge to any number.

Conditional probabilities of observations:

- ✔ Again, probabilities $p_{X|K}(x|k)$ may not be known or may not exist.
- ✔ E.g. if we want to decide what characters are on paper cards written by several persons, the observation x of the state k is influenced by an unobservable non-random intervention—by the writer z.
  - ✘ We can only talk about $p_{X|K,Z}(x|k,z)$, not about $p_{X|K}(x|k)$.
  - ✘ If Z was random and if we knew $p_Z(z)$, than we could compute also $p_{X|K}(x|k)$.

Non-Bayesian Decision Theory

Non-Bayesian decision tasks

When?

- ✔ Tasks where W, p_K, or $p_{X|K}$ are not known.
- ✔ Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- ✔ In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

Non-Bayesian decision tasks

When?

- ✔ Tasks where W, p_K, or $p_{X|K}$ are not known.
- ✔ Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- ✔ In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

There are several special cases of practically useful non-Bayesian formulations for which the solution is known:

- ✔ The strategies that solve these non-Bayesian tasks are of the same form as Bayesian strategies—**they divide the probability space to a set of convex cones**.
- ✔ These non-Bayesian tasks can be formulated as linear programs and **solved by linear programming methods**.

Non-Bayesian decision tasks

When?

- ✔ Tasks where W, p_K, or $p_{X|K}$ are not known.
- ✔ Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- ✔ In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

There are several special cases of practically useful non-Bayesian formulations for which the solution is known:

- ✔ The strategies that solve these non-Bayesian tasks are of the same form as Bayesian strategies—**they divide the probability space to a set of convex cones**.
- ✔ These non-Bayesian tasks can be formulated as linear programs and **solved by linear programming methods**.

There are many other non-Bayesian tasks for which the solution is not known yet.

Neyman-Pearson task

Situation:

- ✔ Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- ✔ The probability distribution $p_{X|K}(x|k)$ exists and is known.
- ✔ Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- ✔ The set X is to be divided to 2 subsets X_1 and X_2, $X = X_1 \cup X_2$.
- ✔ In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Neyman-Pearson task

Situation:

- Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- The probability distribution $p_{X|K}(x|k)$ exists and is known.
- Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- The set X is to be divided to 2 subsets X_1 and X_2, $X = X_1 \cup X_2$.
- In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Each strategy q is characterized by 2 numbers:

- Probability of false positive (false alarm):

$$\omega(1) = \sum_{x \in X_2} p_{X|K}(x|1)$$

- Probability of false negative (overlooked danger):

$$\omega(2) = \sum_{x \in X_1} p_{X|K}(x|2)$$

Neyman-Pearson task

Situation:

- ✔ Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- ✔ The probability distribution $p_{X|K}(x|k)$ exists and is known.
- ✔ Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- ✔ The set X is to be divided to 2 subsets X_1 and X_2, $X = X_1 \cup X_2$.
- ✔ In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Each strategy q is characterized by 2 numbers:

- ✔ Probability of false positive (false alarm):

$$\omega(1) = \sum_{x \in X_2} p_{X|K}(x|1)$$

- ✔ Probability of false negative (overlooked danger):

$$\omega(2) = \sum_{x \in X_1} p_{X|K}(x|2)$$

- Neyman-Pearson task formulation:
Find a strategy q, i.e. a decomposition of X into X_1 and X_2, such that

- ✔ the probability of overlooked danger (FN) is not larger than a predefined value ϵ, i.e.

$$\omega(2) = \sum_{x \in X_1} p_{X|K}(x|2) \le \epsilon,$$

- ✔ and the probability of false alarm (FP) is minimal, i.e.

$$\text{minimize } \omega(1) = \sum_{x \in X_2} p_{X|K}(x|1),$$

- ✔ under the additional conditions

$$X_1 \cap X_2 = \emptyset, X_1 \cup X_2 = X.$$

Solution: The optimal strategy q^* separates X_1 and X_2 according the *likelihood ratio*:

$$q^*(x) = \begin{cases} 1 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} > \theta, \\ 2 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} < \theta. \end{cases}$$

Minimax task

Situation:

- ✔ Observation $x \in X$, states $k \in K$.
- ✔ $q : X \to K$ — given the observation x, the strategy decides the object state k.
- ✔ The set X is to be divided to $|K|$ subsets $X_1, \ldots, X_{|K|}$, $X = X_1 \cup \cdots \cup X_{|K|}$.
- ✔ Again, $p_K(k)$ and $W(k, d)$ are not required.

Minimax task

Situation:

- ✔ Observation $x \in X$, states $k \in K$.
- ✔ $q : X \to K$ — given the observation x, the strategy decides the object state k.
- ✔ The set X is to be divided to $|K|$ subsets $X_1, \ldots, X_{|K|}$, $X = X_1 \cup \cdots \cup X_{|K|}$.
- ✔ Again, $p_K(k)$ and $W(k, d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{x \notin X_k} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task

Situation:

- ✔ Observation $x \in X$, states $k \in K$.
- ✔ $q : X \rightarrow K$ — given the observation x, the strategy decides the object state k.
- ✔ The set X is to be divided to $|K|$ subsets $X_1, \ldots, X_{|K|}$, $X = X_1 \cup \cdots \cup X_{|K|}$.
- ✔ Again, $p_K(k)$ and $W(k,d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{x \notin X_k} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task formulation:
Find a strategy q^* which minimizes

$$\max_{k \in K} \omega(k)$$

Minimax task

Situation:

- ✔ Observation $x \in X$, states $k \in K$.
- ✔ $q : X \to K$ — given the observation x, the strategy decides the object state k.
- ✔ The set X is to be divided to $|K|$ subsets $X_1, \ldots, X_{|K|}$, $X = X_1 \cup \cdots \cup X_{|K|}$.
- ✔ Again, $p_K(k)$ and $W(k, d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{x \notin X_k} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task formulation:
Find a strategy q^* which minimizes

$$\max_{k \in K} \omega(k)$$

Solution:

- ✔ The solution is of the same form as the Bayesian strategies.
- ✔ The solution for the $|K| = 2$ case is similar to the Neyman-Pearson task, with the exception that in minimax task the probability of FN cannot be controlled explicitly.

Wald task

Motivation:

- ✔ The Neyman-Pearson task is asymmetric: the prob. of FN is controlled explicitly, while the probability of FP is minimized (but can be quite high).
- ✔ Can we find a strategy for which *both* the error probabilities would not exceed a predefined ϵ? No, the demands often cannot be accomplished in the same time.

Motivation:

- ✔ The Neyman-Pearson task is asymmetric: the prob. of FN is controlled explicitly, while the probability of FP is minimized (but can be quite high).
- ✔ Can we find a strategy for which *both* the error probabilities would not exceed a predefined ϵ? No, the demands often cannot be accomplished in the same time.

Wald's relaxation:

- ✔ Decompose X into X_1, X_2, and X_0 corresponding to $D = K \cup \{\texttt{dontknow}\}$.
- ✔ Strategy of this form is characterized by 4 numbers:
  - ✘ the conditional prob. of a wrong decision about the state k,

$$\omega(1) = \sum_{x \in X_2} p_{X|K}(x|1) \quad \text{and} \quad \omega(2) = \sum_{x \in X_1} p_{X|K}(x|2),$$

  - ✘ the conditional prob. of the `dontknow` decision when the object state is k,

$$\chi(1) = \sum_{x \in X_0} p_{X|K}(x|1) \quad \text{and} \quad \chi(2) = \sum_{x \in X_0} p_{X|K}(x|2).$$

- ✔ The requirements $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$ are no longer contradictory for an arbitrarily small $\epsilon > 0$, since the strategy $X_0 = X, X_1 = \emptyset, X_2 = \emptyset$ is plausible.
- ✔ Each strategy fulfilling $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$ is then characterized by how often the strategy refuses to decide, i.e. by the number $\max(\chi(1), \chi(2))$.

Wald task (cont.)

Wald task formulation:
Find a strategy q^* which minimizes

$$\max(\chi(1), \chi(2))$$

subject to conditions $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$.

Wald task (cont.)

Wald task formulation:
Find a strategy q^* which minimizes

$$\max(\chi(1), \chi(2))$$

subject to conditions $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$.

Solution: The optimal decision is based on the likelihood ratio and 2 thresholds $\theta_1 > \theta_2$:

$$q^*(x) = \begin{cases} 1 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} > \theta_1, \\ 2 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} < \theta_2, \\ \texttt{dontknow} & \text{otherwise.} \end{cases}$$

Wald task formulation:
Find a strategy q^* which minimizes

$$\max(\chi(1), \chi(2))$$

subject to conditions $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$.

Solution: The optimal decision is based on the likelihood ratio and 2 thresholds $\theta_1 > \theta_2$:

$$q^*(x) = \begin{cases} 1 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} > \theta_1, \\ 2 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} < \theta_2, \\ \texttt{dontknow} & \text{otherwise.} \end{cases}$$

In [SH02], also the generalization for $|K| > 2$ is given.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Previous non-Bayesian tasks did not require

- ✔ the a priori probabilities of the states $p_K(k)$, and
- ✔ the penalty function $W(k, d)$ to be known.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Previous non-Bayesian tasks did not require

- ✔ the a priori probabilities of the states $p_K(k)$, and
- ✔ the penalty function $W(k, d)$ to be known.

In Linnik tasks,

- ✔ the conditional probabilities $p_{X|K}(x|k)$ do not exist,
- ✔ the a priori probabilities $p_K(k)$ may exist (it depends on the fact if the state k is a random variable or not),
- ✔ but the conditional probabilities $p_{X|K,Z}(x|k, z)$ do exist, i.e. the random observation x depends not only on the (random or non-random) object state k, but also on a non-random intervention z.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Previous non-Bayesian tasks did not require

- ✔ the a priori probabilities of the states $p_K(k)$, and
- ✔ the penalty function $W(k, d)$ to be known.

In Linnik tasks,

- ✔ the conditional probabilities $p_{X|K}(x|k)$ do not exist,
- ✔ the a priori probabilities $p_K(k)$ may exist (it depends on the fact if the state k is a random variable or not),
- ✔ but the conditional probabilities $p_{X|K,Z}(x|k, z)$ do exist, i.e. the random observation x depends not only on the (random or non-random) object state k, but also on a non-random intervention z.

Goal:

- ✔ find a strategy that minimizes the probability of incorrect decision in case of the worst intervention z.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Previous non-Bayesian tasks did not require

- ✔ the a priori probabilities of the states $p_K(k)$, and
- ✔ the penalty function $W(k, d)$ to be known.

In Linnik tasks,

- ✔ the conditional probabilities $p_{X|K}(x|k)$ do not exist,
- ✔ the a priori probabilities $p_K(k)$ may exist (it depends on the fact if the state k is a random variable or not),
- ✔ but the conditional probabilities $p_{X|K,Z}(x|k, z)$ do exist, i.e. the random observation x depends not only on the (random or non-random) object state k, but also on a non-random intervention z.

Goal:

- ✔ find a strategy that minimizes the probability of incorrect decision in case of the worst intervention z.

See examples in [SH02].

Summary of PR

- ✔ The aim of PR is to design decision strategies (classifiers) which—given an observation x of an object with a hidden state k—provide a decision d such that this decision making process is optimal with respect to some criteria.
- ✔ If the statistical properties of (x, k) are completely known, and if we are able to design a suitable penalty function $W(k, d)$, we should solve the task in the *Bayesian framework* and search for the *Bayesian strategy* which optimizes the *Bayesian risk* of the strategy.
  - ✘ The minimization of the probability of an error is a special case, the resulting Bayesian strategy decides for the state with the *maximum a posteriori probability*.
- ✔ If the statistical properties are known only partially, or are not known at all, or if a reasonable penalty function cannot be constructed, we face a *non-Bayesian task*.
  - ✘ Several practically important special cases of non-Bayesian tasks are well-analyzed and solved (Neyman-Pearson, minimax, Wald, . . .).
  - ✘ There are plenty of non-Bayesian tasks we can say nothing about.

Reference

[SH02] Michail I. Schlesinger and Václav Hlaváč. *Ten Lectures on Statistical and Structural Pattern Recognition*. Kluwer Academic Publishers, Dodrecht, The Netherlands, 2002.