Bayes Classifier

Thierry Chateau

Clermont Auvergne University

2017

Plan

Bayesian Classifier

- Bayesian decision is very popular in pattern recognition and machine learning
- It is a probabilistic based model
- The problem is expressed using probabilities (input and output).
- Under such hypothesis, this theory is optimal.
- But ...

Toy example

- Example of a company that transforms tree trunks into wooden planks. Inputs trees of this factory are from two varieties
- Let define the state (class) of a plank as the category of tree that is used: (class ω_1 for category 1) or (class ω_2 for category 2).

Toy example

Prior probalility

- We assume the proportion of planks produced are known: 75% of trees from category 1 and 25% of trees from category 2.
- **Question**: With no measure, how to decide the class associated to the next plunk that will be produce?
- **Answer** : We will bet on category 1 (minimization of the error probability)
- Finally, we use a important informations: (**prior probabilities**):
  - $p(\omega_1) = 0.75$
  - $p(\omega_2) = 0.25$

Toy example

Prior probabilities

- When no prior is known, the same probability for each class is chosen.
- When it is possible, prior can learn with statistics.

Règle de Bayes

- Let $\{\omega_1, \omega_2, ...\omega_c\}$ be a set of c classes and $\mathbf{x}$ a feature vector (measures).
- For each class ω_i we assume that we known:
  - $P(\omega_i)$: Prior probability for each class,
  - $p(\mathbf{x}|\omega_i)$: the probability density function of the features given the class (likelihood function)

Bayes rule

- The Bayes rule computes the posterior probability using the following rule:

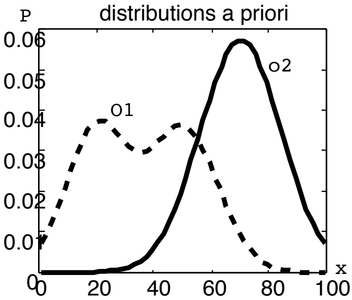
$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

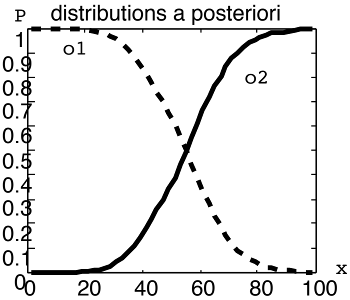
with :

$$p(\mathbf{x}) = \sum_i (p(\mathbf{x}|\omega_i).P(\omega_i))$$

Illustration of Bayes rule

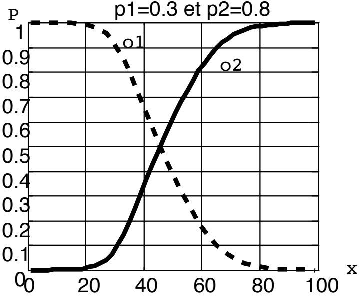
2 classes





$P(\omega_1) = P(\omega_2) = 0.5$

2 classes



1. Left: $P(\omega_1) = 0.3$, $P(\omega_2) = 0.7$
2. Right: $P(\omega_1) = 0.7$, $P(\omega_2) = 0.3$

Error probability

Let $\mathbf{x}$ a feature vector and $\delta(\mathbf{x}) = \omega_i$ a decision. The error probability associated to this decision is:

$$P(\text{error}|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

The global error probability associated to the system is :

$$P(\text{errorglob}|\mathbf{x}) = \int_{-\infty}^{\infty} P(\text{error}|\mathbf{x}).P(\mathbf{x})dx$$

Optimal decision

The optimal decision (that minimize the error probability) is computed by:

$$\delta(\mathbf{x}) = \omega_i$$

such as:

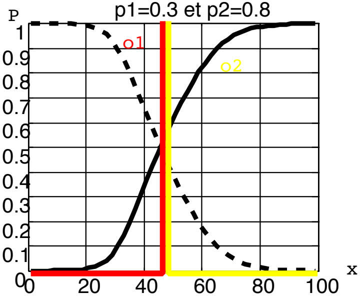
$$P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x}) \forall j$$

which is equivalent to:

$$p(\mathbf{x}|\omega_i).P(\omega_i) \geq p(\mathbf{x}|\omega_j).P(\omega_j) \forall j$$

Regions of decision

For two classes (Régions of decision, boundary of decision)



1. left $P(\omega_1) = 0.3$, $P(\omega_2) = 0.7$
2. right $P(\omega_1) = 0.7$, $P(\omega_2) = 0.3$

Cost and Risk

- Let $\{\delta_1, \delta_2, ..\delta_d\}$, be the set of the possible decisions: δ_i is associated to $\delta(\mathbf{x}) = \omega_i$
- Let $\lambda(\delta_i|\omega_i)$ a cost of the decision δ_i when the object belongs to the classω_i (correct classification)
- the error probabiliy seen below is the special case:

$$\lambda(\delta_i, \omega_i) = \left\{ \begin{array}{l} 0 \;\; i = j \\ 1 \;\; i \neq j \end{array} \right. \tag{1}$$

Risq

- The risk associated to the decision δ_i (conditional risk) is:

$$R(\delta_i|\mathbf{x}) = \sum_j \lambda(\delta_i|\omega_j)P(\omega_j|\mathbf{x})$$

- the global risk is computed by:

$$R = \int_{R^n} R(\delta(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- Minimizing the global risk is obtained by choosing, for each value of $\mathbf{x}$, the decision which minimize the conditional risk.

Example for two classes

- Let's define $\lambda_{ij} = \lambda(\delta_i|\omega_j)$ (the cost of the predicted decision δ_i While the true class is ω_j)

$$R(\delta_1|\mathbf{x}) = \lambda_{11} p(\omega_1|\mathbf{x}) + \lambda_{12} p(\omega_2|\mathbf{x})$$

$$R(\delta_2|\mathbf{x}) = \lambda_{21} p(\omega_1|\mathbf{x}) + \lambda_{22} p(\omega_2|\mathbf{x})$$

  with : $\lambda_{11} < \lambda_{12}$ and $\lambda_{21} < \lambda_{22}$ (because the right decision must be the one with the lowest cost):

$$\omega_1 \text{ if } R(\delta_1|\mathbf{x}) < R(\delta_2|\mathbf{x})$$

Example of two classes

- therefore:

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x})$$

- and

$$(\lambda_{21} - \lambda_{11})P(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22} p(\mathbf{x}|\omega_1)P(\omega_1)$$

- finally, we decide ω_1 if:

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

- this ratio is called **likelihood ratio**

Classification by minimizing the error

- A symetric cost function is defined by: $\lambda_{ij} = 0$ if $i = j$ and $\lambda_{ij} = 1$ si $i \neq j$.
- risk:

$$R(\delta_i|\mathbf{x}) = \sum_j \lambda(\delta_i|\omega_j) P(\omega_j|\mathbf{x})$$

$$R(\delta_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

- Minimizing the risk Minimiser is done by maximizing posterior probabilities.

Discriminative functions

- To define a decision rule, we use a discriminative function:

$$g_i(x)\ , i = 1, ..s$$

  (s=number of classes)
- The s discriminative functions are such as a unknown feature vector $\mathbf{x}$ is classify in the class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})\ \forall j \neq i$

Discriminative functions

Several discriminative functions could be defined:

- $g_i(\mathbf{x}) = -R(\delta_i|\mathbf{x})$
- $g_i(\mathbf{x}) = p(\omega_i|\mathbf{x})$
- $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$
- $f(g_i(\mathbf{x}))$ if f is a monotonous increasing function g_i is a discriminative function.

Discriminative functions

Relevant discriminative functions:

- $g_i(\mathbf{x}) = p(\omega_i|\mathbf{x})$
- $g_i(\mathbf{x}) = \dfrac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_i(p(\mathbf{x}|\omega_i).P(\omega_i))}$
- $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$
- $g_i(\mathbf{x}) = log(p(\mathbf{x}|\omega_i)) + log(p(\omega_i))$