AARHUS
UNIVERSITET

Optimization
and
Data Analytics

Alexandros Iosifidis
@
Aarhus University, Department of Engineering

Probability-based learning

Let us assume that for a problem, we can define K possible outcomes

$$\mathcal{C} = \{c_1, \ldots, c_K\}$$

These outcomes may correspond to e.g. K classes in a classification problem, or to K clusters in a clustering problem.

Probability-based learning

Let us assume that for a problem, we can define K possible outcomes

$$\mathcal{C} = \{c_1, \ldots, c_K\}$$

These outcomes may correspond to e.g. K classes in a classification problem, or to K clusters in a clustering problem.

We use $P(c_k)$ to denote the (a priori) probability of each possible outcome.

Since $P(c_k)$ are probabilities, we have

$$\sum_{k=1}^{K} P(c_k) = 1$$

Probability-based learning

Example:

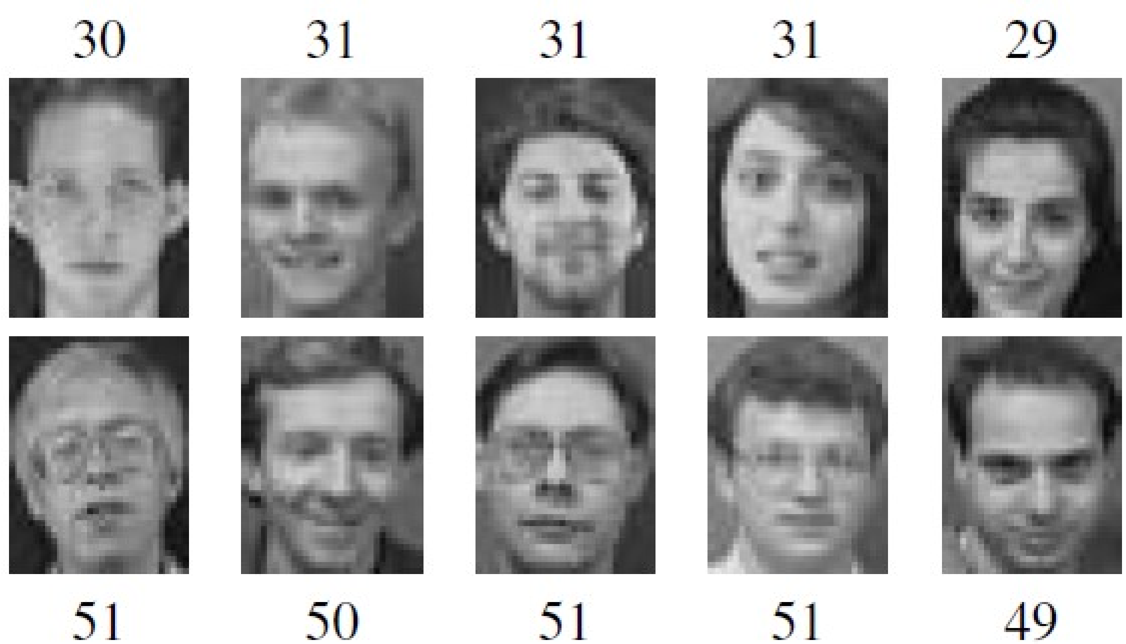
Let us define two classes C = {‘young’, ‘old’}

Probability-based learning

Example:

Let us define two classes C = {‘young’, ‘old’}

Given the following samples, what is $P(c_k)$ for k = 1, 2?



Probability-based learning

Example:

Let us define two classes C = {'young', 'old'}

Given the following samples, what is $P(c_k)$ for k = 1, 2?

$P(c_1) = 5 / 10 = 0.5$ or 50%

$P(c_2) = 5 / 10 = 0.5$ or 50%

Can we use the a priori probabilities as decision rules for new samples?

Probability-based learning

Example:

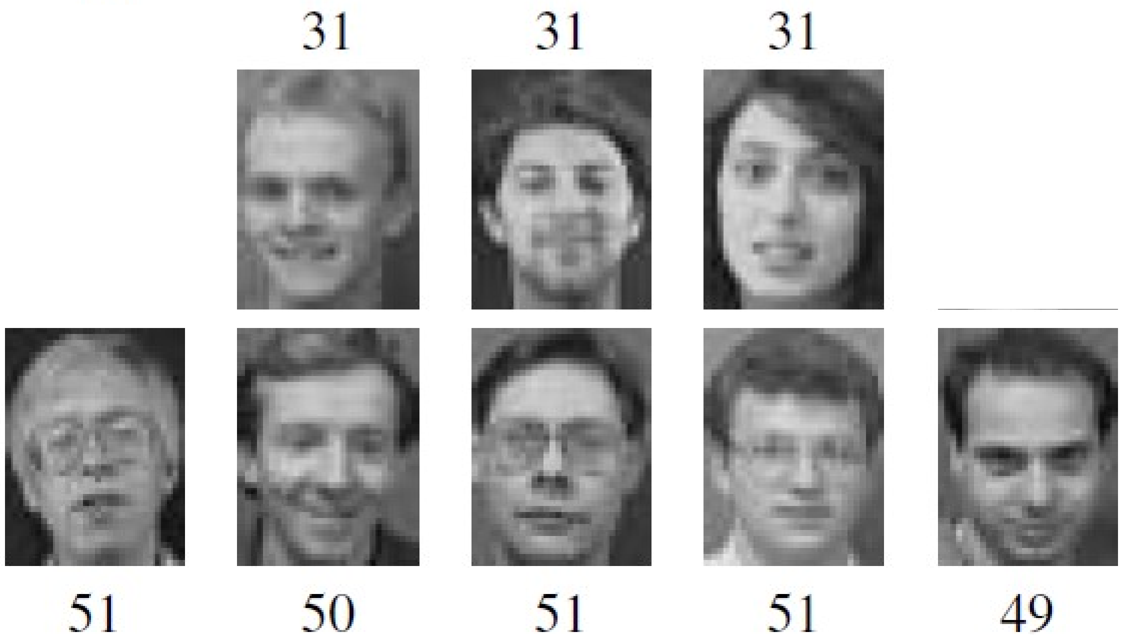
Let us define two classes C = {‘young’, ‘old’}

Can we use the a priori probabilities for classification now?

$P(c_1) = 3 / 8 = 0.375$ or 37.5%

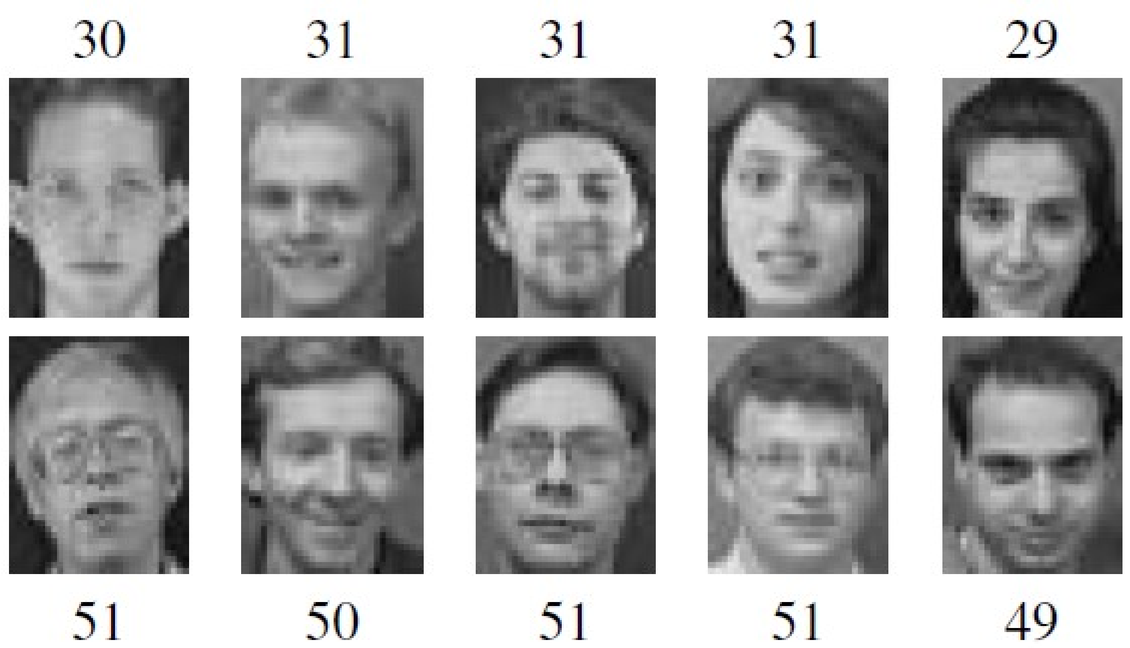
$P(c_2) = 5 / 8 = 0.625$ or 62.5%

The classifier does not use any input information so it may classify all



Probability-based learning

Usually, classification rules are defined based on measurable variables x or a set of variables **x**. In our case, such a measurable variable can be the age of each person.



If x=35 then we getinto fu

Probability-based learning

Usually, classification rules are defined based on measurable variables x or a set of variables **x**. In our case, such a measurable variable can be the age of each person.

We define the conditional probability of class c_k given x as $P(c_k \mid x)$

For example, in our example, $P(c_1 \mid x=30)$ expresses the probability of class ‘young’ given an observation x = 30.

Probability-based learning

Usually, classification rules are defined based on measurable variables x or a set of variables **x**. In our case, such a measurable variable can be the age of each person.

We define the conditional probability of class c_k given x as $P(c_k \mid x)$

In our example, $P(c_1 \mid x=30)$ expresses the probability of class 'young' given an observation x = 30.

In a similar way we can define the class-conditional probability $p(x \mid c_k)$.

In our example, $P(x=30 \mid c_1)$ expresses the probability of observing an age value of x = 30, given that the sample belongs to class 'young'.

Probability-based learning

We also define the joint probability of c_k and x as

$$p(c_k, x) = P(c_k|x)p(x) = p(x|c_k)P(c_k)$$

where

$$p(x) = \sum_{k=1}^{K} p(x|c_k)P(c_k)$$

Can you calculate the conditional probability $p(c_1|x=31)$?

Probability-based learning

We also define the joint probability of c_k and x as

$$p(c_k, x) = P(c_k|x)p(x) = p(x|c_k)P(c_k)$$

The above can lead to the Bayes' formula

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

Probability-based learning

Given $P(c_k \mid x)$ we can define the probability of error as follows

$$P(error|x) = \begin{cases} P(c_1|x), & \text{if } x \text{ is classified to } c_2 \\ P(c_2|x), & \text{if } x \text{ is classified to } c_1 \end{cases}$$

Probability-based learning

Given $P(c_k | x)$ we can define the probability of error as follows

$$P(error|x) = \begin{cases} P(c_1|x), & \text{if } x \text{ is classified to } c_2 \\ P(c_2|x), & \text{if } x \text{ is classified to } c_1 \end{cases}$$

which is given by

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

Probability-based learning

Given $P(c_k \,|x)$ we can define the probability of error as follows

$$P(error|x) = \begin{cases} P(c_1|x), & \text{if } x \text{ is classified to } c_2 \\ P(c_2|x), & \text{if } x \text{ is classified to } c_1 \end{cases}$$

Thus, we can define the following decision rule (Bayes' rule)

$$\text{Decide } c_1 \text{ if } \; P(c_1|x) > P(c_2|x), \;\; \text{otherwise decide } c_2$$

Probability-based learning

Given $P(c_k|x)$ we can define the probability of error as follows

$$P(error|x) = \begin{cases} P(c_1|x), & \text{if } x \text{ is classified to } c_2 \\ P(c_2|x), & \text{if } x \text{ is classified to } c_1 \end{cases}$$

Thus, we can define the following decision rule (Bayes' rule)

$$\text{Decide } c_1 \text{ if } \; P(c_1|x) > P(c_2|x), \;\; \text{otherwise decide } c_2$$

Which is the decision function?

Probability-based learning

Given $P(c_k \mid x)$ we can define the probability of error as follows

$$P(error|x) = \begin{cases} P(c_1|x), & \text{if } x \text{ is classified to } c_2 \\ P(c_2|x), & \text{if } x \text{ is classified to } c_1 \end{cases}$$

Thus, we can define the following decision rule (Bayes' rule)

$$\text{Decide } c_1 \text{ if } P(c_1|x) > P(c_2|x), \text{ otherwise decide } c_2$$

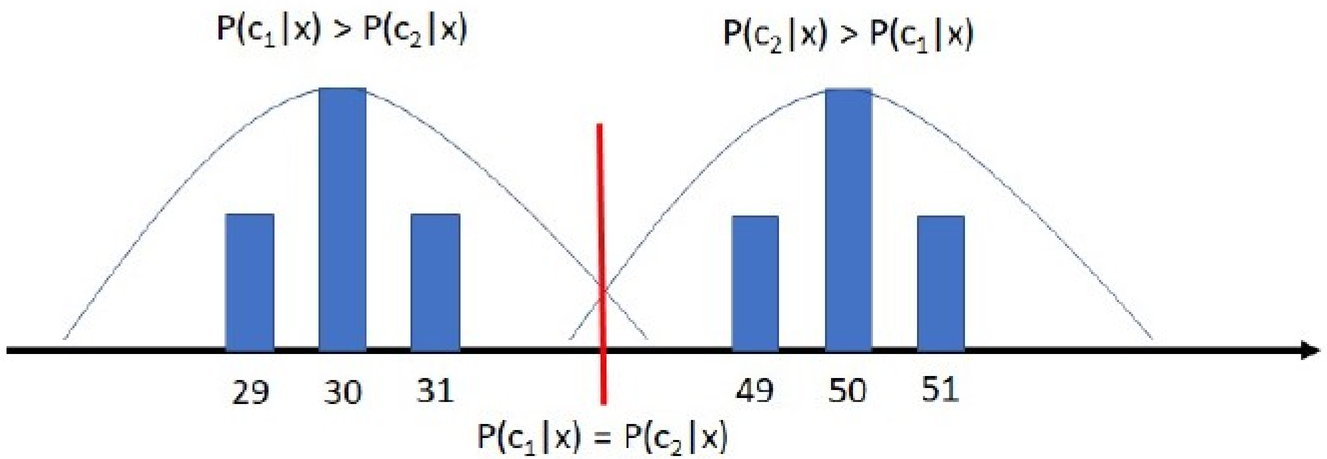
Which is the decision function? → It is obtained by finding x for which

$$P(c_1|x) = P(c_2|x)$$

Probability-based learning

Which is the decision function? → It is obtained by finding x for which

$$P(c_1|x) = P(c_2|x)$$



Probability-based learning

The Bayes' rule can be extended to more than K = 2 classes

$$\text{Decide } c_l \text{ for which } P(c_l|x) > P(c_i|x),\ i \neq l.$$

In the case where $P(c_k) = 1 / K$, $k=1,\ldots,K$, we have

$$P(c_k|x) = \alpha\, p(x|c_k).$$

Thus the decision rule can be defined on $p(x|c_k)$. Because $p(x|c_k)$ is also called likelihood of c_k with respect to x, in this case Bayes' decision rule is also called as Maximum Likelihood Classification.

Some classification errors are more important than other. Fx. if we classify that there is no tumor but there is actually a tumor (misclassifcation), this is more important classifying there is a tumor if there is not not

Risk-based decision functions

Let us consider the general case where the observations are more than one and are stored to a vector **x**. Then, the Bayes' formula is

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

where

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|c_k)P(c_k)$$

Risk-based decision functions

Suppose that given the observation **x**, we take the action of classifying the new sample to class i. We call this action α_i.

Risk-based decision functions

Suppose that given the observation **x**, we take the action of classifying the new sample to class i. We call this action α_i.

Let us also define a loss function $\lambda(\alpha_i \mid c_k)$, which expresses the loss incurred by taking action α_i, given that the correct class is c_k.

The values of lambda depends on pair of predicted class and the true class.

Why do we need such a loss function?

Risk-based decision functions

Suppose that given the observation **x**, we take the action of classifying the new sample to class i. We call this action α_i.

Let us also define a loss function $\lambda(\alpha_i \mid c_k)$, which expresses the loss incurred by taking action α_i, given that the correct class is c_k.

Why do we need such a loss function?

What was the loss function in the previous example? 1/K

The impact of misclassifying cancer patient as healthy is very high.

Risk-based decision functions

Suppose that given the observation **x**, we take the action of classifying the new sample to class i. We call this action α_i.

Let us also define a loss function $\lambda(\alpha_i \mid c_k)$, which expresses the loss incurred by taking action α_i, given that the correct class is c_k.

Then, we define the risk of α_i given the observation **x** as

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda(\alpha_i|c_k) P(c_k|\mathbf{x})$$

Risk-based decision functions

Suppose that given the observation **x**, we take the action of classifying the new sample to class i. We call this action α_i.

Let us also define a loss function $\lambda(\alpha_i \mid c_k)$, which expresses the loss incurred by taking action α_i, given that the correct class is c_k.

The loss function (i.e., the loss value) for misclassification is given by the user.

Then, we define the risk of α_i given the observation **x** as

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda(\alpha_i|c_k) P(c_k|\mathbf{x})$$

After calculating the risk for each action, we can take the action with the smallest risk.

Risk-based decision functions

Then, we define the risk of α_i given the observation **x** as

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda(\alpha_i|c_k)P(c_k|\mathbf{x})$$

For the two-class case, we have

$$\begin{aligned} R(\alpha_1|\mathbf{x}) &= \lambda(\alpha_1|c_1)P(c_1|\mathbf{x}) + \lambda(\alpha_1|c_2)P(c_2|\mathbf{x}) \\ R(\alpha_2|\mathbf{x}) &= \lambda(\alpha_2|c_1)P(c_1|\mathbf{x}) + \lambda(\alpha_2|c_2)P(c_2|\mathbf{x}) \end{aligned}$$

We classify $\mathbf{x}$ to c_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$, or:

$$\left(\lambda(\alpha_2|c_1) - \lambda(\alpha_1|c_1)\right) P(c_1|\mathbf{x}) > \left(\lambda(\alpha_1|c_2) - \lambda(\alpha_2|c_2)\right) P(c_2|\mathbf{x})$$

Risk-based decision functions

For the two-class case, we have

$$\begin{aligned} R(\alpha_1|\mathbf{x}) &= \lambda(\alpha_1|c_1)P(c_1|\mathbf{x}) + \lambda(\alpha_1|c_2)P(c_2|\mathbf{x}) \\ R(\alpha_2|\mathbf{x}) &= \lambda(\alpha_2|c_1)P(c_1|\mathbf{x}) + \lambda(\alpha_2|c_2)P(c_2|\mathbf{x}) \end{aligned}$$

We classify $\mathbf{x}$ to c_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$, or:

$$(\lambda(\alpha_2|c_1) - \lambda(\alpha_1|c_1))\, P(c_1|\mathbf{x}) > (\lambda(\alpha_1|c_2) - \lambda(\alpha_2|c_2))\, P(c_2|\mathbf{x})$$

By substituting the posterior probabilities using the Bayes formula, we get

$$(\lambda(\alpha_2|c_1) - \lambda(\alpha_1|c_1))\, p(\mathbf{x}|c_1)P(c_1) > (\lambda(\alpha_1|c_2) - \lambda(\alpha_2|c_2))\, p(\mathbf{x}|c_2)P(c_2)$$

or (taking the form of the *likelihood ratio*):

$$\frac{p(\mathbf{x}|c_1)}{p(\mathbf{x}|c_2)} > \frac{(\lambda(\alpha_1|c_2) - \lambda(\alpha_2|c_2))\, P(c_2)}{(\lambda(\alpha_2|c_1) - \lambda(\alpha_1|c_1))\, P(c_1)}.$$