Elements of Bayesian Decision Theory

A simple practical problem

A medical test of a disease presents 1% false positives. The disease strikes 2 on 10000 of the population.

People are tested at random, regardless of whether they are suspected of having the disease.

A patient's test is positive. What is the probability of the patient having the disease?

Solution

A thought experiment: test 10000 people.

2 will test positive because they have the disease

0.01*9998≈100 will test positive because the test will give a false positive result (1%)

Hence, only 2 of the 102 who test positive do have the disease => probability of having the disease if the test is positive is

$$2/102 \approx 0.02$$

Solution in a formula

$$P(sick|positive) = \frac{2}{2 + 0.01 * 9998} =$$

$$= \frac{p(positive|sick)P(sick)}{p(positive|sick)P(sick) + p(positive|\neg sick)P(\neg sick)}$$

Some definitions

P(sick), P(¬ sick) – prior probabilities

p(positive|sick), p(positive|¬ sick) – class conditional probabilities (likelihoods)

p(positive|sick)*P(sick) + p(positive|¬ sick)*P(¬ sick) - evidence

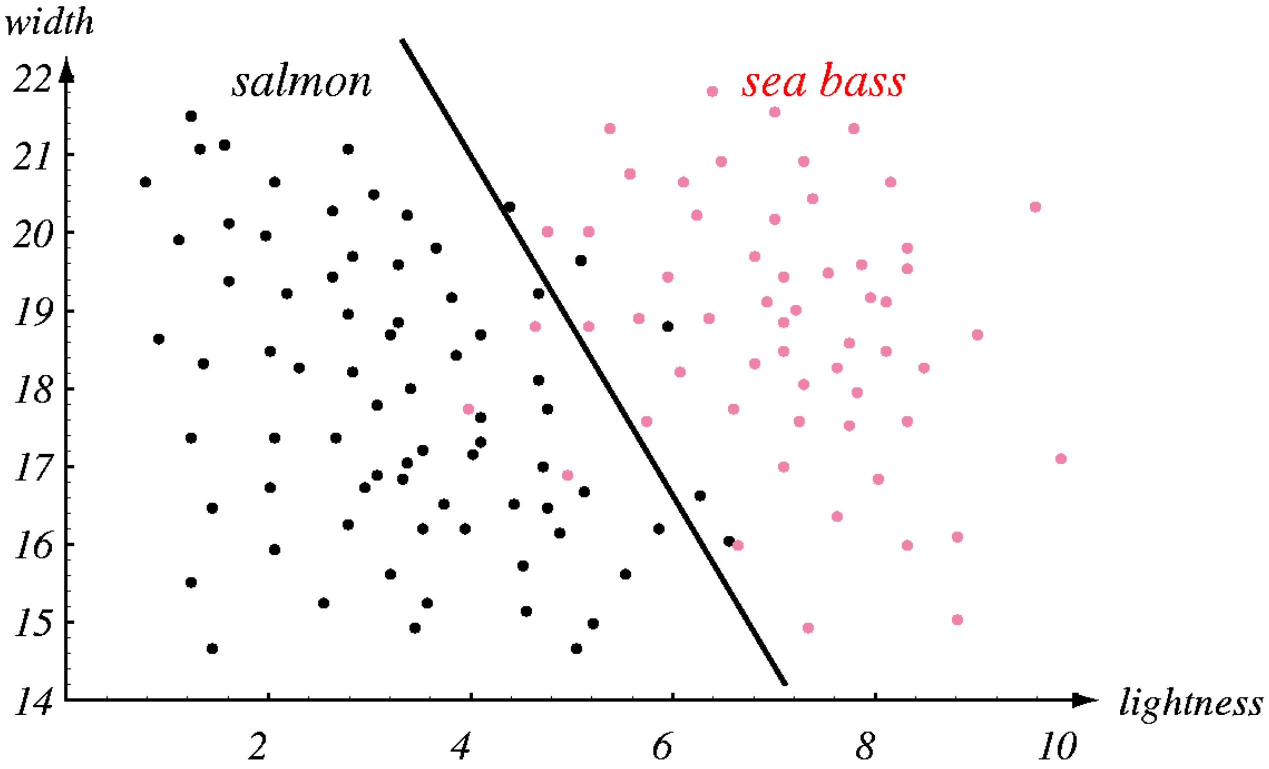
Bayes rule: prior*likelihood/evidence

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)\ P(\omega_j)}{p(x)}$$

Probabilistic approach to classification

For each point, estimate the probability for each class.

Choose the class with the highest probability.



(from Duda, Hart, Stork (2001) Pattern classification)

How to estimate probabilities

THE CENTRAL PROBLEM IN THE PROBABILISTIC APPROACH TO CLASSIFICATION

Priors

Classes

ω_1 - sea bass

ω_2 - salmon

a two-class problem

A priory probabilities (or prior probabilities)

$P(\omega_1)$ - probability of finding sea bass

$P(\omega_2)$ - probability of finding salmon

A simple decision rule

$$\begin{cases} \omega_1, if\ P(\omega_1) > P(\omega_2) \\ \omega_2, otherwise \end{cases}$$

Class conditional probability density function and likelihood

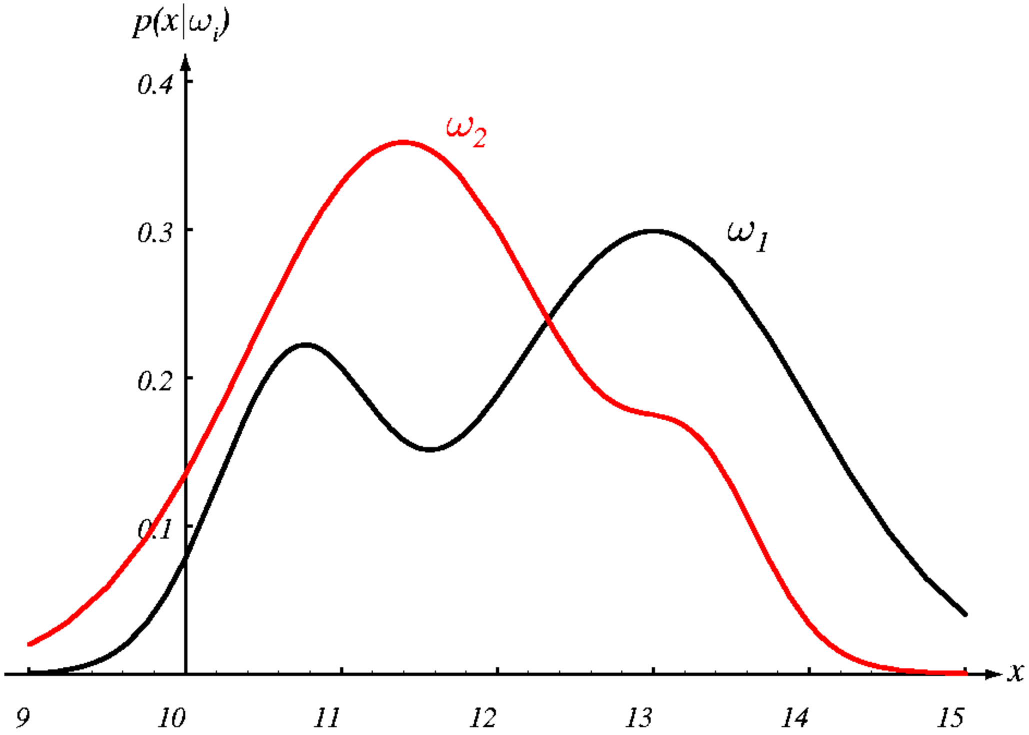
Likelihood

pdf as a function of

the first argument

(feature value x) with the second argument (class) fixed

$$p(x \mid \omega_1)$$
$$p(x \mid \omega_2)$$



x: lightness

(from Duda, Hart, Stork (2001) Pattern classification)

Bayes formula/rule

$$p(x,\omega_j) = p(x\,|\,\omega_j)\,P(\omega_j)$$ Joint probability

$$p(x,\omega_j) = P(\omega_j\,|\,x)\,p(x)$$

$$P(\omega_j\,|\,x)\,p(x) = p(x\,|\,\omega_j)\,P(\omega_j)$$

$$P(\omega_j\,|\,x) = \frac{p(x\,|\,\omega_j)\;\;P(\omega_j)}{p(x)}$$ Bayes rule

$$p(x) = p(x\,|\,\omega_1)\,P(\omega_1) + p(x\,|\,\omega_2)\,P(\omega_2)$$

$$posterior = \frac{likelihood\;\;\mathrm{x}\;\;prior}{evidence}$$

Bayes decision rule

Probability of making an error:

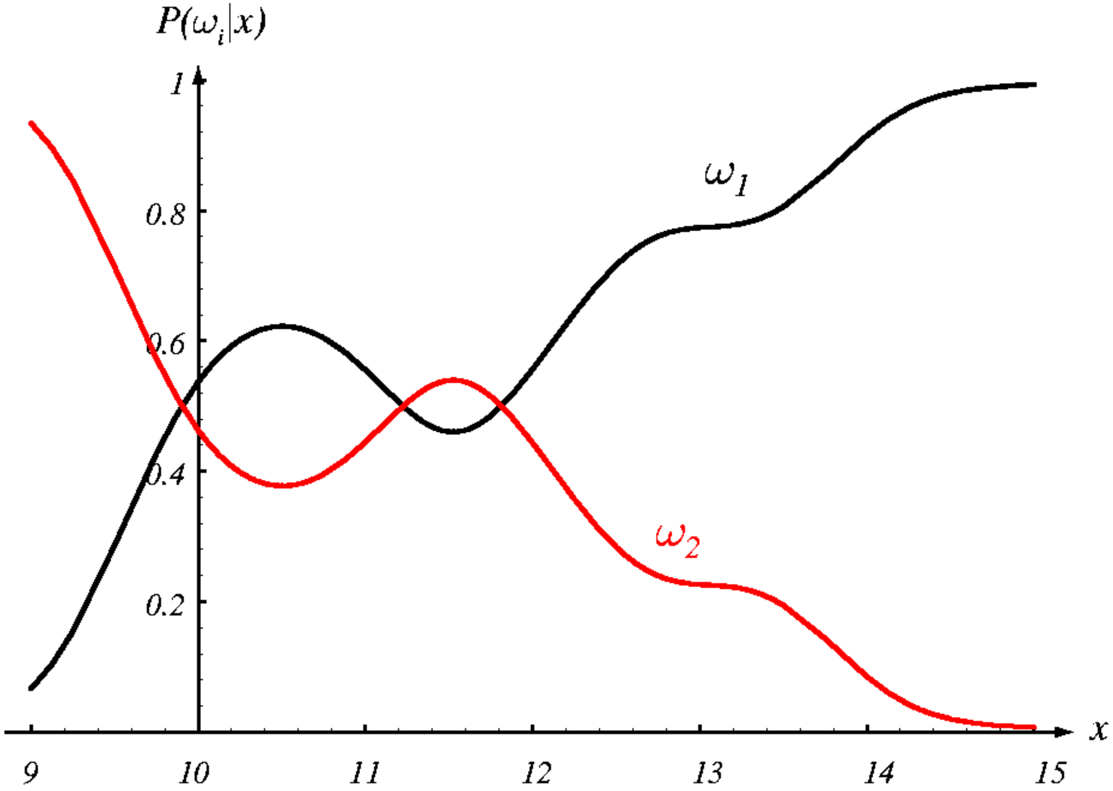
$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x), if\ we\, decide\ \ \omega_2 \\ P(\omega_2 \mid x), if\ we\, decide\ \ \omega_1 \end{cases}$$

Bayes decision rule:

$$\begin{cases} \omega_1, if\ P(\omega_1 \mid x) > P(\omega_2 \mid x) \\ \omega_2, otherwise \end{cases}$$

Posterior probability plots

Use priors as coefficients of likelihoods and normalize so that their sum is 1 for any x



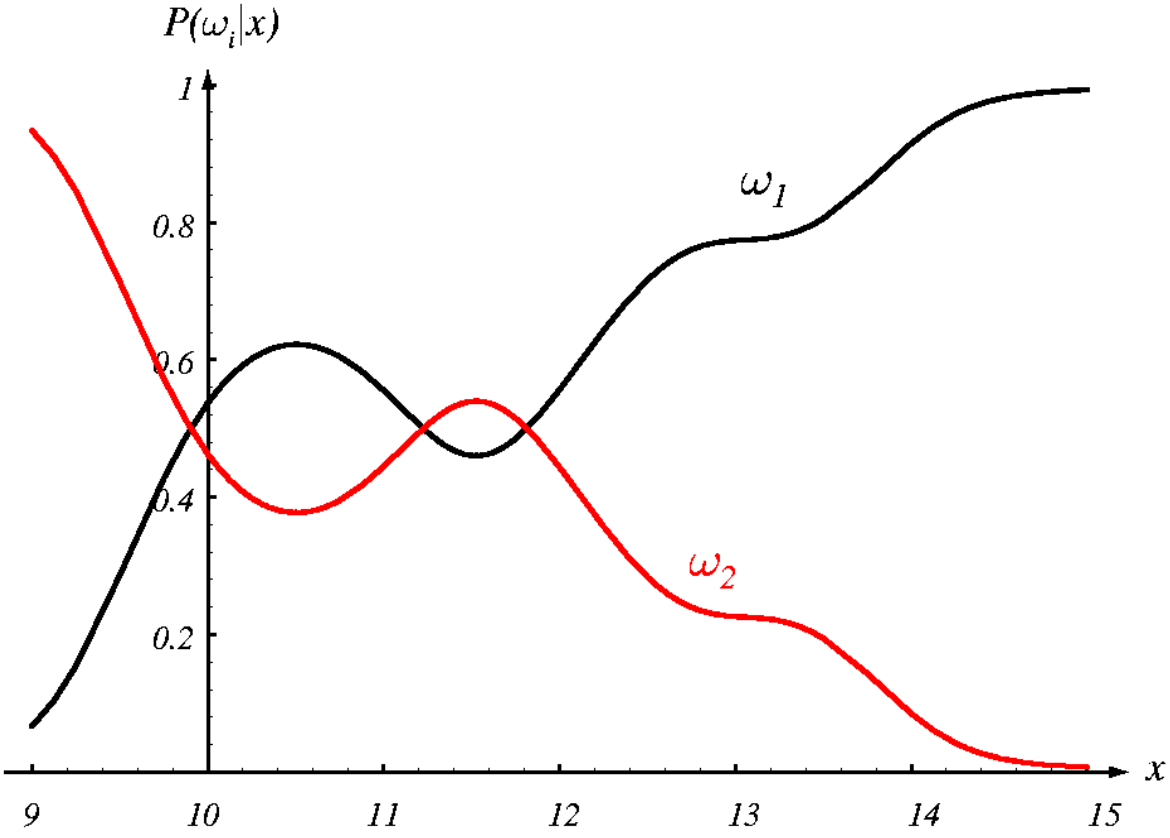
$$P(\omega_1) = 2/3$$

$$P(\omega_2) = 1/3$$

(from Duda, Hart, Stork (2001) Pattern classification)

Error probability of Bayes decision rule

$$P(error \mid x) = \min[P(\omega_1 \mid x), P(\omega_2 \mid x)]$$



(from Duda, Hart, Stork (2001) Pattern classification)

Generalizations of Bayesian Decision Theory

We replace the scalar x with the feature vector $\underline{\mathrm{x}} \in \mathrm{R}^d$

We introduce a cost or a loss function λ which states how costly each classification decisions is.

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ - categories (classes)

$\{\alpha_1, \alpha_2, \ldots, \alpha_c\}$ - possible actions

The loss function $\lambda(\alpha_i \,|\, \omega_j)$ describes the loss incurred for taking action α_i when the category is ω_j

Bayes formula

$$P(\omega_j \mid \underline{\mathrm{x}}) = \frac{p(\underline{\mathrm{x}} \mid \omega_j)\ P(\omega_j)}{p(\underline{\mathrm{x}})}$$

Evidence

$$p(\underline{\mathrm{x}}) = \sum_{j=1}^{c} p(\underline{\mathrm{x}} \mid \omega_j)\ P(\omega_j)$$

Bayesian decision theory

Taking action α_i , the loss, also called *conditional risk*, is:

$$R(\alpha_i \mid \underline{\mathrm{x}}) = \sum_{\mathrm{j=1}}^{\mathrm{c}} \lambda(\alpha_{\mathrm{i}} \mid \omega_j)\, P(\omega_j \mid \underline{\mathrm{x}})$$

Rule to minimize the expected loss:

- *Select that action which minimizes the conditional risk.*

Generalized Bayesian decision theory

Let P(melanoma|x) = 0.1 and P(benign nevus|x) = 0.9

Bayesian classification: **benign nevus** (since it has higher probability)

Let now consider the actions: α_1 – remove, α_2 – do not remove, with costs

$\lambda(\alpha_1|mel) = 50$ $\quad\quad$ $\lambda(\alpha_1|nev) = 50$

$\lambda(\alpha_2|mel) = 100000$ $\quad\quad$ $\lambda(\alpha_2|nev) = 0$

Expected cost R_i as weighted average over many cases with same x:

$R_1 = \lambda(\alpha_1|mel)P(mel \mid x) + \lambda(\alpha_1|nev)P(nev|x) = 50*0.1 + 50*0.9 = 50$

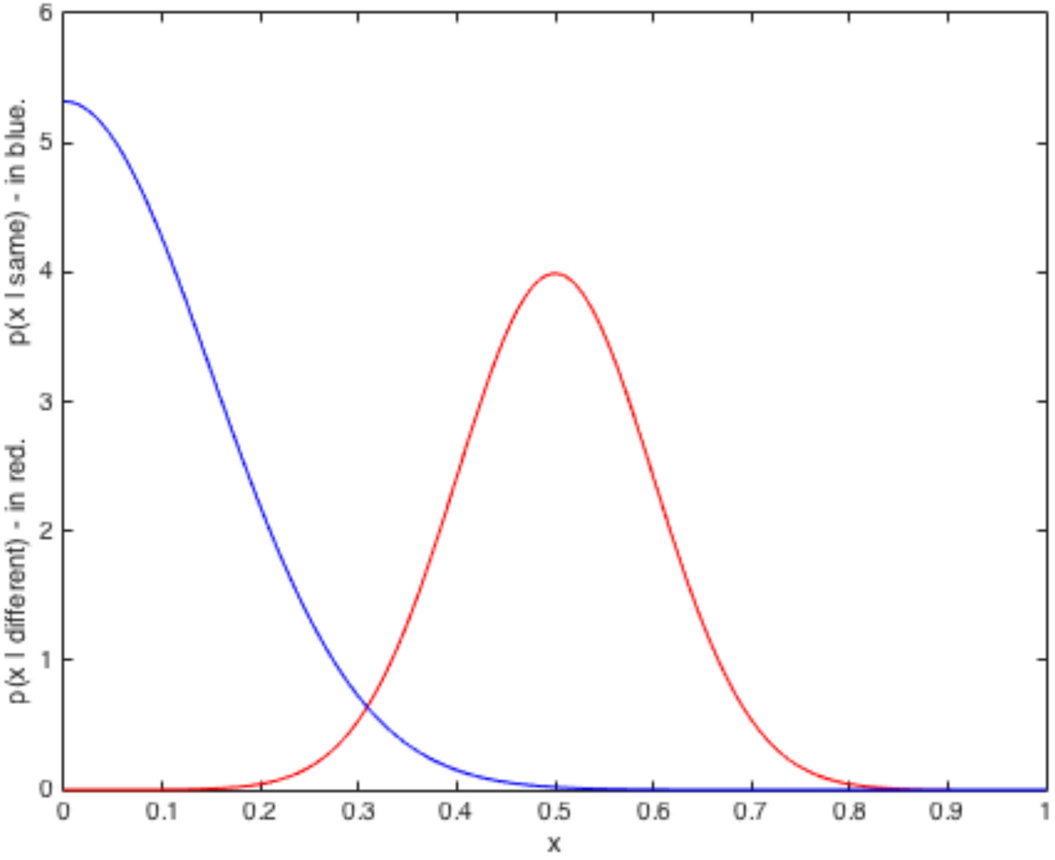
$R_2 = \lambda(\alpha_2|mel)P(mel \mid x) + \lambda(\alpha_2|nev)P(nev|x) = 100000*0.1 + 0*0.9 = 10000$

-> we choose for the action with lower cost: α_1 - 'remove'

Rule to minimize the expected loss:

Select that action which minimizes the conditional risk.

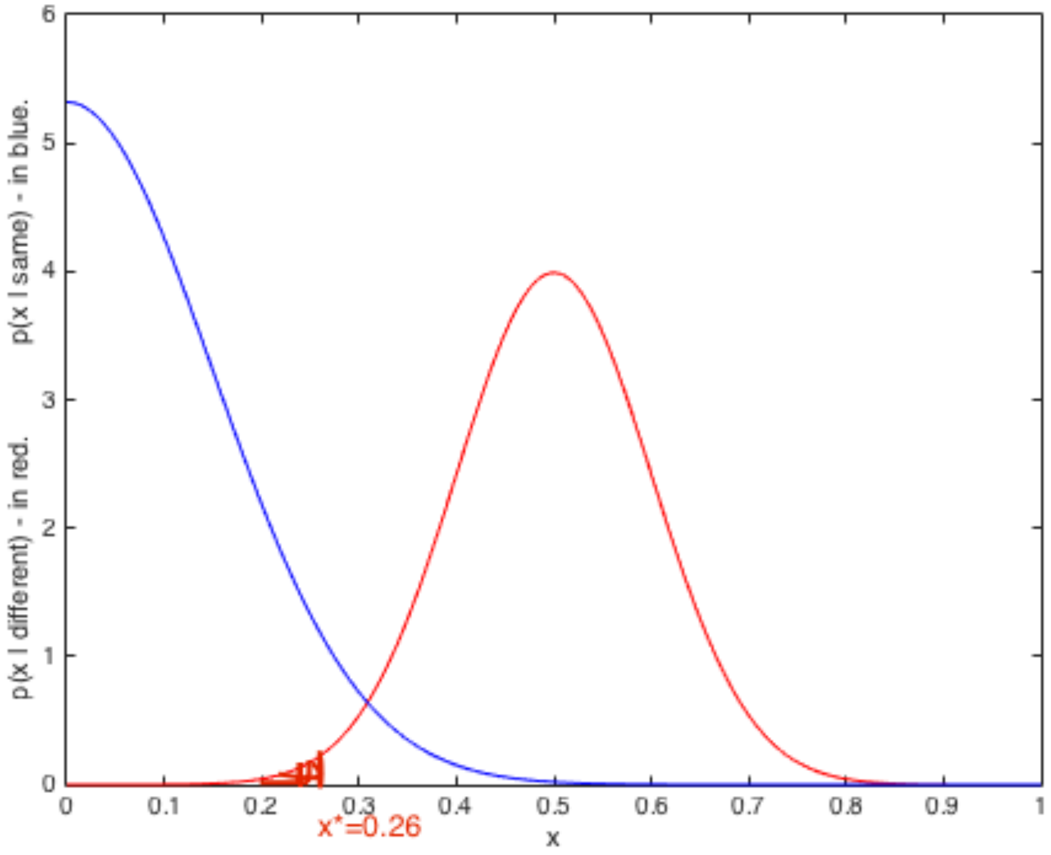
Differences between decision theory based on (A) hypothesis testing and (B) Bayesian rule



For decisions based on **hypothesis testing**, we use **one distribution** only – the one that corresponds to the null hypothesis. Here: ‘The two irises belong to different persons’ – hence, use the red distribution.

For decisions based on **Bayes rule**, for any value of x we use the values of the **two distributions**, here - the red and the blue.

(A) Hypothesis testing

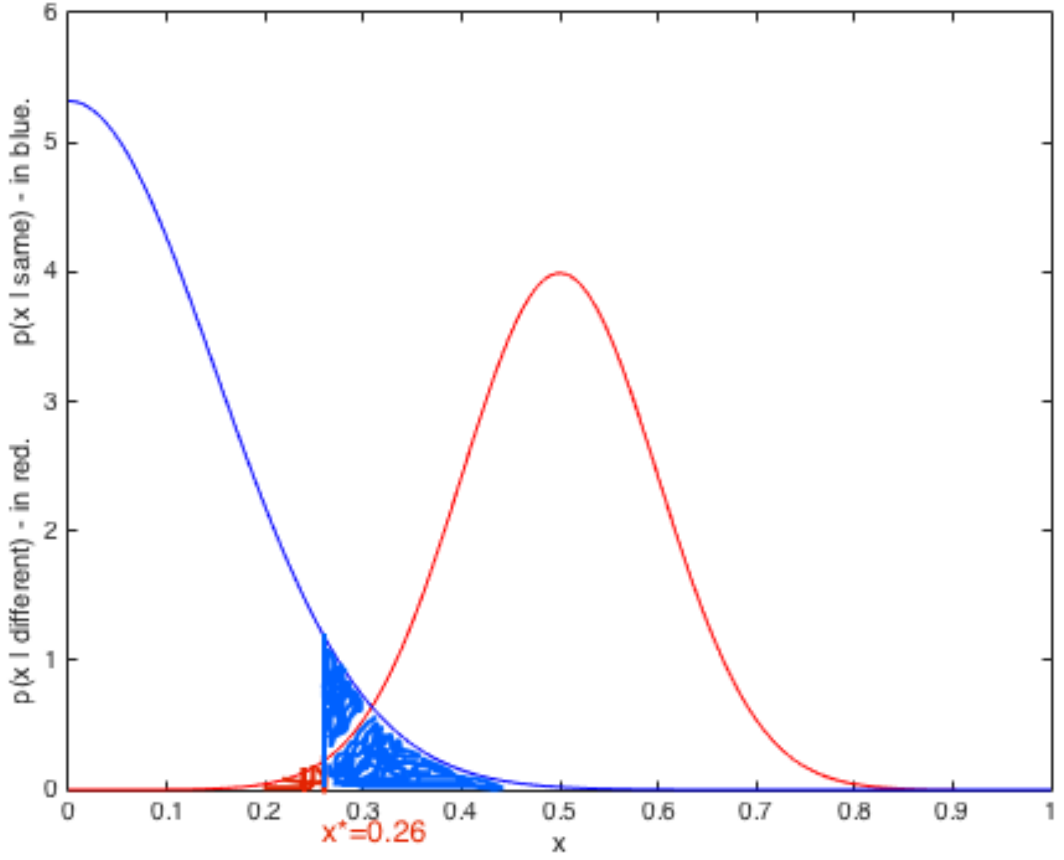


We use the red distribution to set the value of our decision criterion x*. E.g. if we accept an error of of 2% of admitting an imposter (deciding that two irises belong to the same person while they belong to two different persons), we need to set x* = 0.26.

If the comparison of two irises yields distance x > x*, we decide that they belong to different persons.

The tail of the red distribution for x < x* contains the percentage of erroneously admitted imposters.

(A) Hypothesis testing



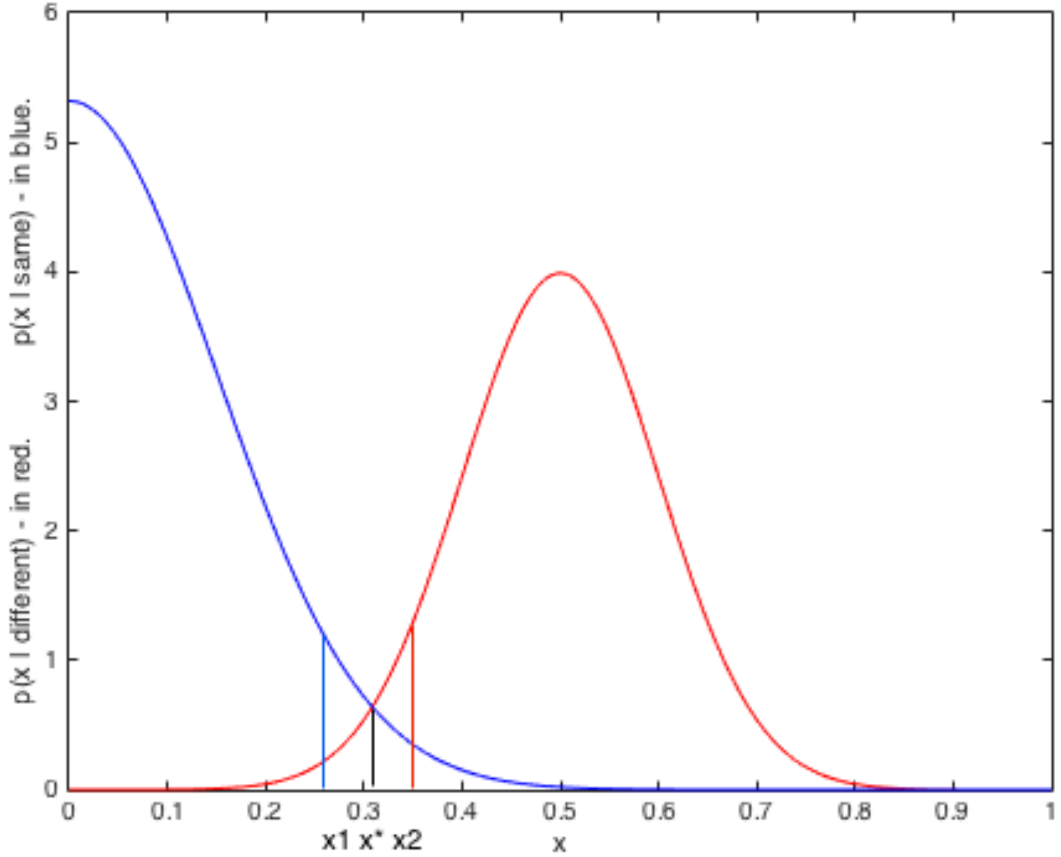
Is the blue distribution used at all?

Yes, it is used to determine the percentage of erroneously rejected authentic persons. This percentage is in the tail of the blue distribution for x > x*.

Two important aspects:

1) We **use only one** distribution (the red one) to choose the value of the decision criterion.
2) For any measured value of x, we do not use the values of the two distributions (as in Bayes rule) to take a decision. The decision solely depends on the relation of x to x*.

(B) Bayes rule



For any value of x, we use the two distributions by comparing the two corresponding probabilities.

x_1 is classified as 'same person', x_2 as 'different'.

x* is set where the two probabilities are equal.

Note: here we use the likelihoods as probabilities, i.e. we assume the priors to be equal. More generally, the two curves need to be scaled by the priors and risks.

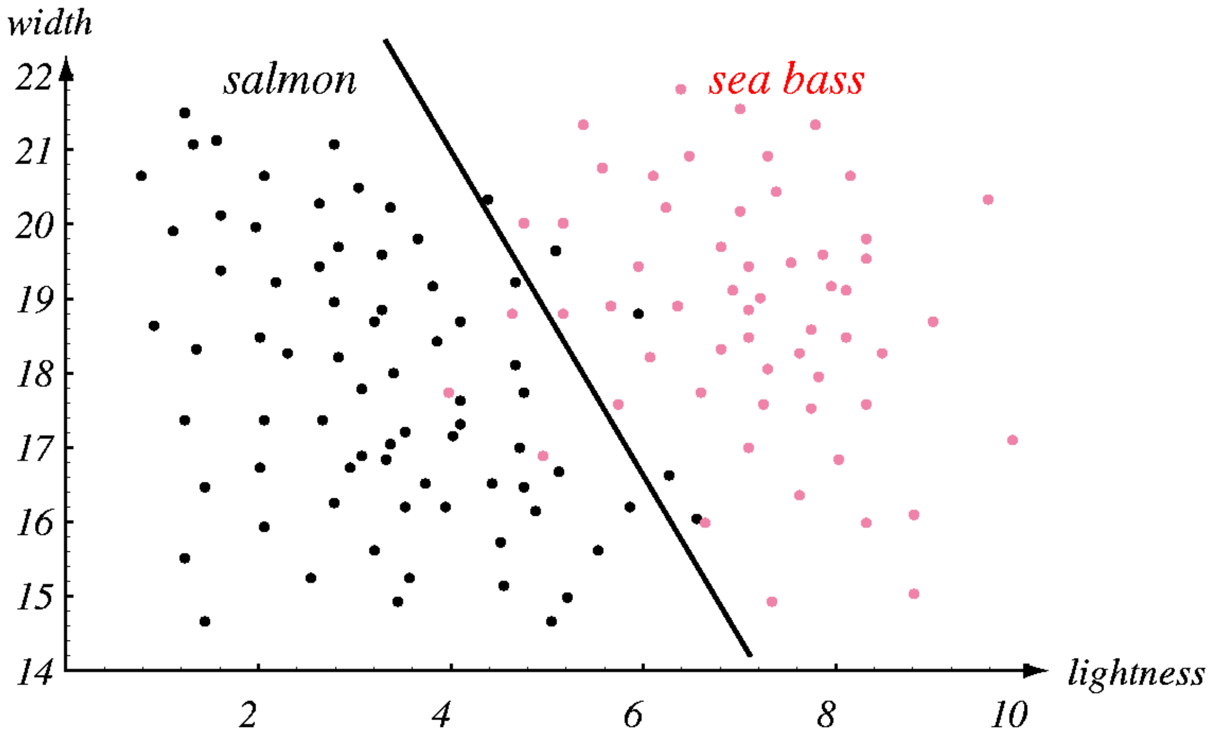
Naïve Bayes pdf estimation

How to estimate class-conditional probabilities

THE CENTRAL PROBLEM IN THE PROBABILISTIC APPROACH TO CLASSIFICATION

Estimation of pdf is a problem for high dimensional data

The more dimensions we have, the more data point we need for reliable estimation of the pdf’s



(from Duda, Hart, Stork (2001) Pattern classification)

Naive Bayes rule

Bayes rule:

$$P(\omega_j \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n \mid \omega_j)\ P(\omega_j)}{p(x_1, x_2, \ldots, x_n)}$$

Simplifying assumption: the features are statistically independent

$$p(x_1, x_2, \ldots, x_n \mid \omega_j) = p(x_1 \mid \omega_j) p(x_2 \mid \omega_j) \ldots p(x_n \mid \omega_j) \qquad j = 1 \ldots c$$

Naïve Bayes rule:

$$P(\omega_j \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1 \mid \omega_j) p(x_2 \mid \omega_j) \ldots p(x_n \mid \omega_j)\ P(\omega_j)}{p(x)}$$

Naive Bayes rule - Advantages

Each distribution can be independently estimated as a 1D distribution

No need for large data sets that scale exponentially with the number of features (curse of dimensionality)

Empirical observation: In many cases it works.

Naïve explanation: Correct classification as long as the correct class is more probable than any other class (hence class probabilities do not have to be estimated very well)

Naive Bayes rule – Example: Spam filter

$$P(spam|w_1, w_2, \ldots, w_n) = \frac{p(w_1|spam)p(w_2|spam)\ldots p(w_n|spam)P(spam)}{p(w_1, w_2, \ldots, w_n)}$$

$$P(\neg spam|w_1, w_2, \ldots, w_n) = \frac{p(w_1|\neg spam)p(w_2|\neg spam)\ldots p(w_n|\neg spam)P(\neg spam)}{p(w_1, w_2, \ldots, w_n)}$$

$$\frac{P(spam|w_1, w_2, \ldots, w_n)}{P(\neg spam|w_1, w_2, \ldots, w_n)} = \frac{p(w_1|spam)p(w_2|spam)\ldots p(w_n|spam)P(spam)}{p(w_1|\neg spam)p(w_2|\neg spam)\ldots p(w_n|\neg spam)P(\neg spam)}$$

Naive Bayes classifier – Example: Spam filter

An email contains the words *viagra, purchase, love, romantic, happy*

$$\frac{P(spam|viagra, purchase, love, romantic, happy)}{P(\neg spam|viagra, purchase, love, romantic, happy)} =$$

$$= \frac{p(viagra|spam)}{p(viagra|\neg spam)} \frac{p(purchase|spam)}{p(purchase|\neg spam)} \frac{p(love|spam)}{p(love|\neg spam)} \frac{p(romantic|spam)}{p(romantic|\neg spam)} \frac{p(happy|spam)}{p(happy|\neg spam)} \frac{P(spam)}{P(\neg spam)} =$$

$$= \frac{0.1}{10^{-4}} \frac{0.3}{3 \cdot 10^{-3}} \frac{0.2}{4 \cdot 10^{-2}} \frac{10^{-2}}{0.05} \frac{4 \cdot 10^{-4}}{0.4} \frac{0.75}{0.25} = 10^{3} \cdot 10^{2} \cdot 5 \cdot 0.2 \cdot 10^{-3} * 3 = 300$$

Summary of concepts and facts

Prior probability

Class conditional probability density function, likelihood

Posterior probability

Bayes formula/rule for posteriors

Bayes decision rule

Minimum cost/loss/risk classification

Naïve Bayes and examples