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≈0.02

Solution in a formula

$P(sick|positive) = 2/(2+0.01*9998) =$

$= p(positive|sick)*P(sick)*10000/(p(positive|sick)*P(sick)*10000+p(positive|not\ sick)*P(not\ sick)*10000)=$

$= p(positive|sick)*P(sick)/(p(positive|sick)*P(sick)+p(positive|not\ sick)*P(not\ sick))$

Solution in a formula

P(sick), P(not sick) – prior probabilities

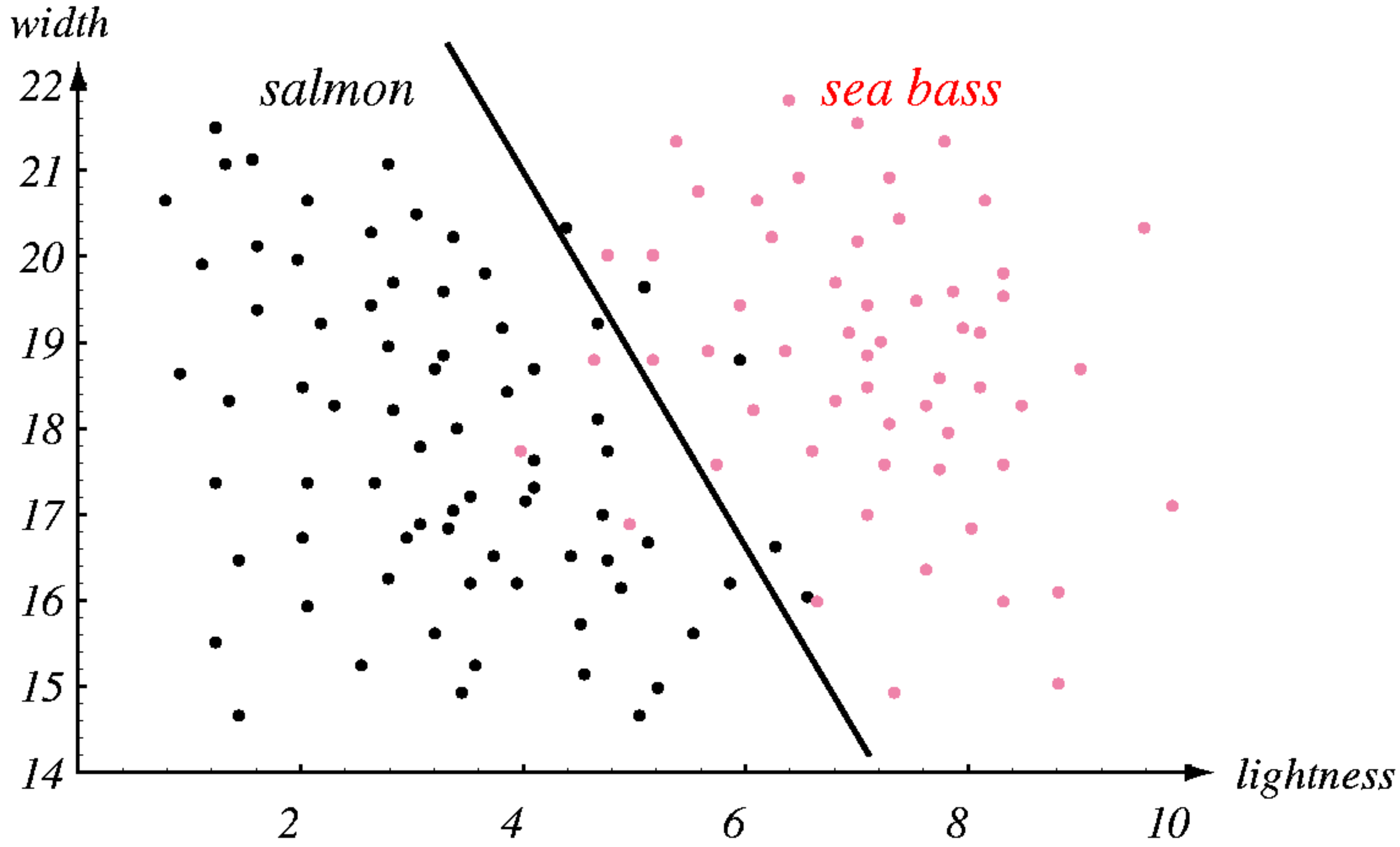
p(positive|sick), p(positive|not sick) – class conditional probabilities (likelihoods)

p(positive|sick)*P(sick)+p(positive|not sick)*P(not 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.

The central problem in the probabilistic approach to classification: How to estimate probabilities

Priors

Classes

- sea bass ω_1
- salmon ω_2

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

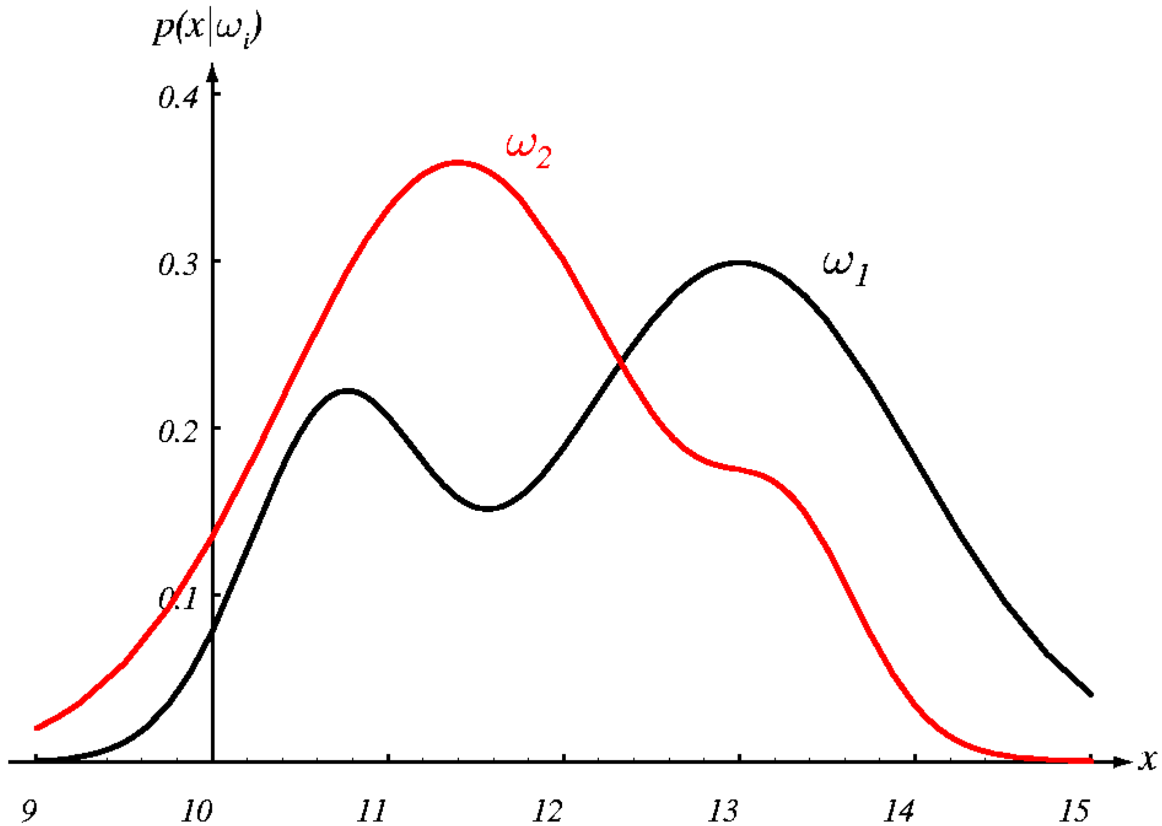
A simple decision rule

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

Class conditional probability density function and likelihood

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

Likelihood –
pdf as a function of the second argument (class) with the first argument (feature value x) fixed



feature x (e.g. lightness)

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

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

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

Bayes formula/rule

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

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

$$posterior = \frac{likelihood \text{ x } prior}{evidence}$$

Bayes decision rule

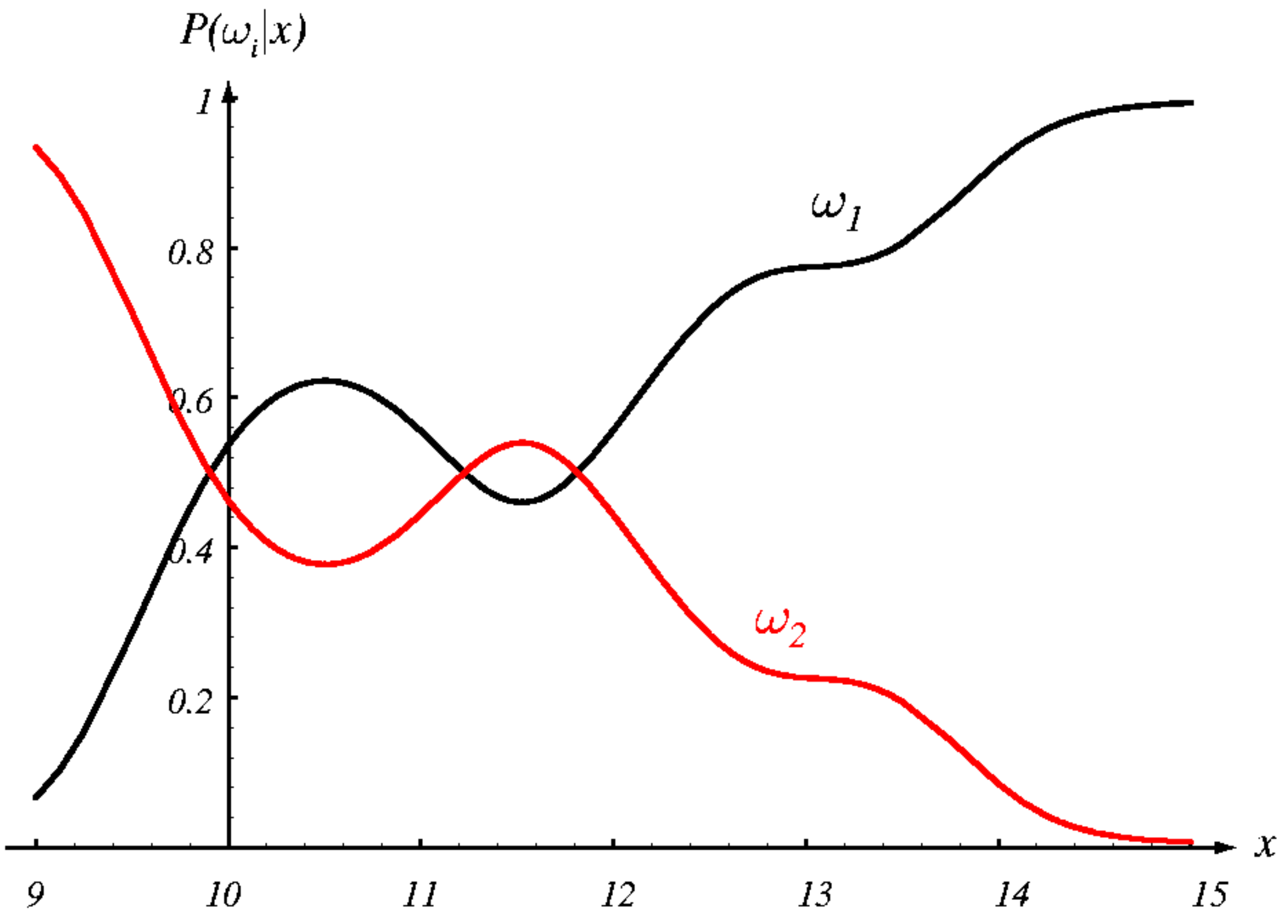
Probability of making an error:

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

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



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

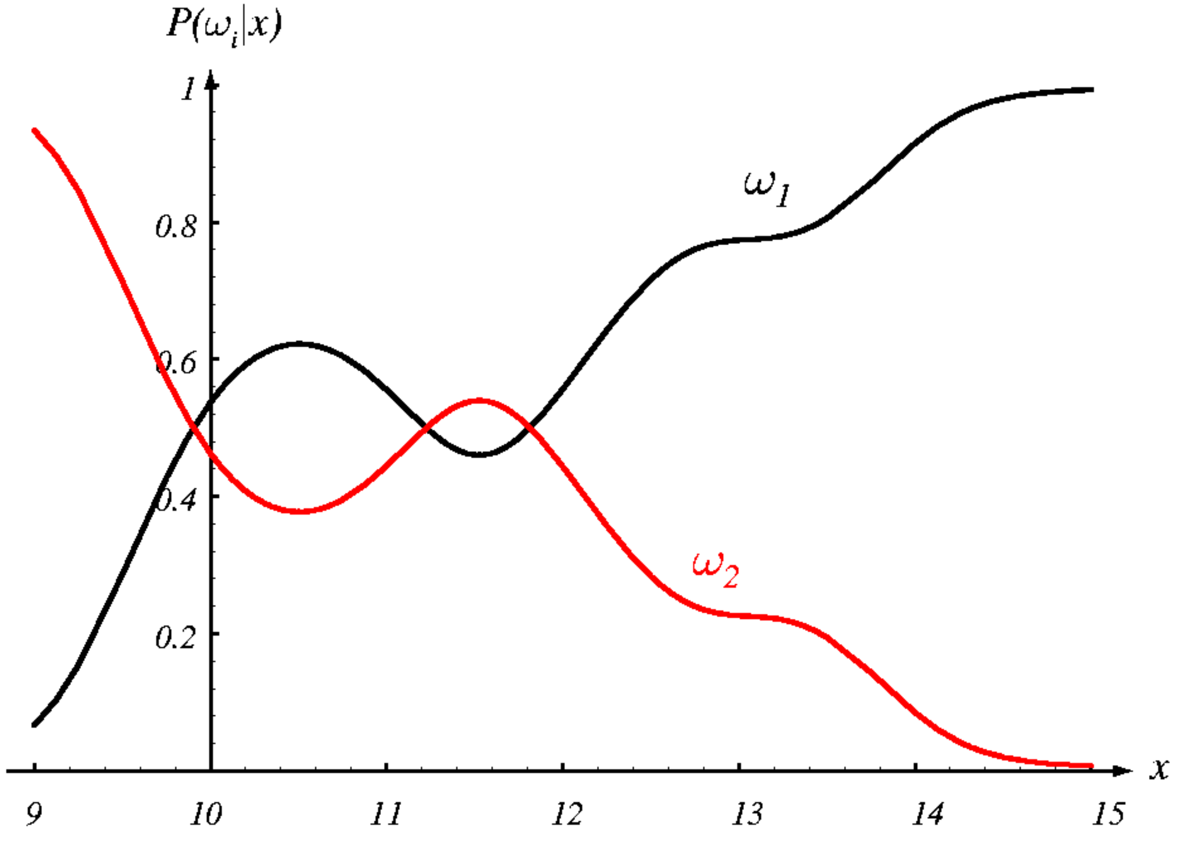
$$P(\omega_2) = 1/3$$

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

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

Error probability of Bayes decision rule

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



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

Generalizations of Bayesian Decision Theory

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

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 $\lambda(\alpha_i \mid \omega_j)$ describes the loss incurred for taking action α_i when the category is ω_j.

Bayes formula

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

Evidence

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

Bayesian decision theory

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

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

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 in Euro $\lambda(\alpha_1|\text{mel}) = 50$ $\lambda(\alpha_1|\text{nev}) = 50$

$\lambda(\alpha_2|\text{mel}) = 100000$ $\lambda(\alpha_2|\text{nev}) = 0$

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

$R_1 = \lambda(\alpha_1|\text{mel})\, P(\text{mel} \mid \mathbf{x}) + \lambda(\alpha_1|\text{nev})\, P(\text{nev} \mid \mathbf{x}) =$

$= 50*0.1 + 50*0.9 = 50$

$R_2 = \lambda(\alpha_2|\text{mel})\, P(\text{mel} \mid \mathbf{x}) + \lambda(\alpha_2|\text{nev})\, P(\text{nev} \mid \mathbf{x}) =$

$= 100000*0.1 + 0*0.9 = 10000$

-> we choose for the action with lower cost: α_2 - ‘remove’

Dealing with missing features in Bayesian decision theory

How can we classify a feature vector $(*, x_2)$ in which the value of the first feature x_1 is missing?

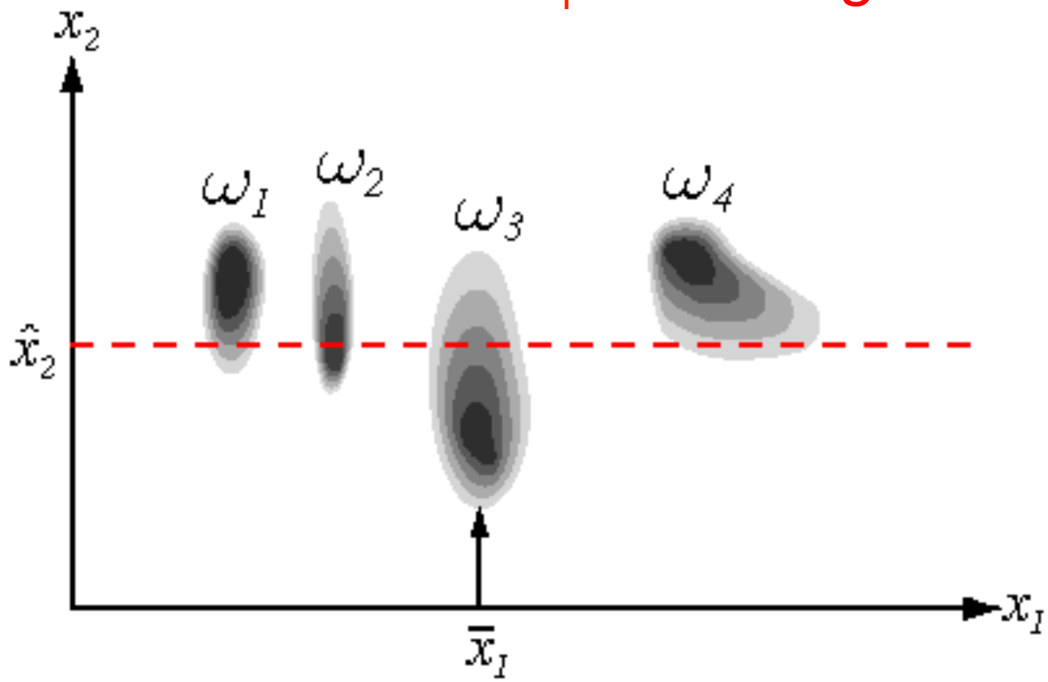
$$P(\omega_i / x_2) = \frac{p(\omega_i, x_2)}{p(x_2)} =$$

$$= \frac{\int p(\omega_i, x_1, x_2)\,dx_1}{p(x_2)} =$$

$$= \frac{\int P(\omega_i) p(x_1, x_2 | \omega_i)\,dx_1}{p(x_2)}$$

$$= \frac{P(\omega_i) \int p(x_1, x_2 | \omega_i)\,dx_1}{p(x_2)}$$

$$= \frac{P(\omega_i)\, p(x_2 | \omega_i)}{p(x_2)}$$



Intuitively one may wish to take some average x_1 – this will result in choosing ω_3.
Correct is however to select ω_2.

from Duda, Hart, Stork (2001) Pattern classification

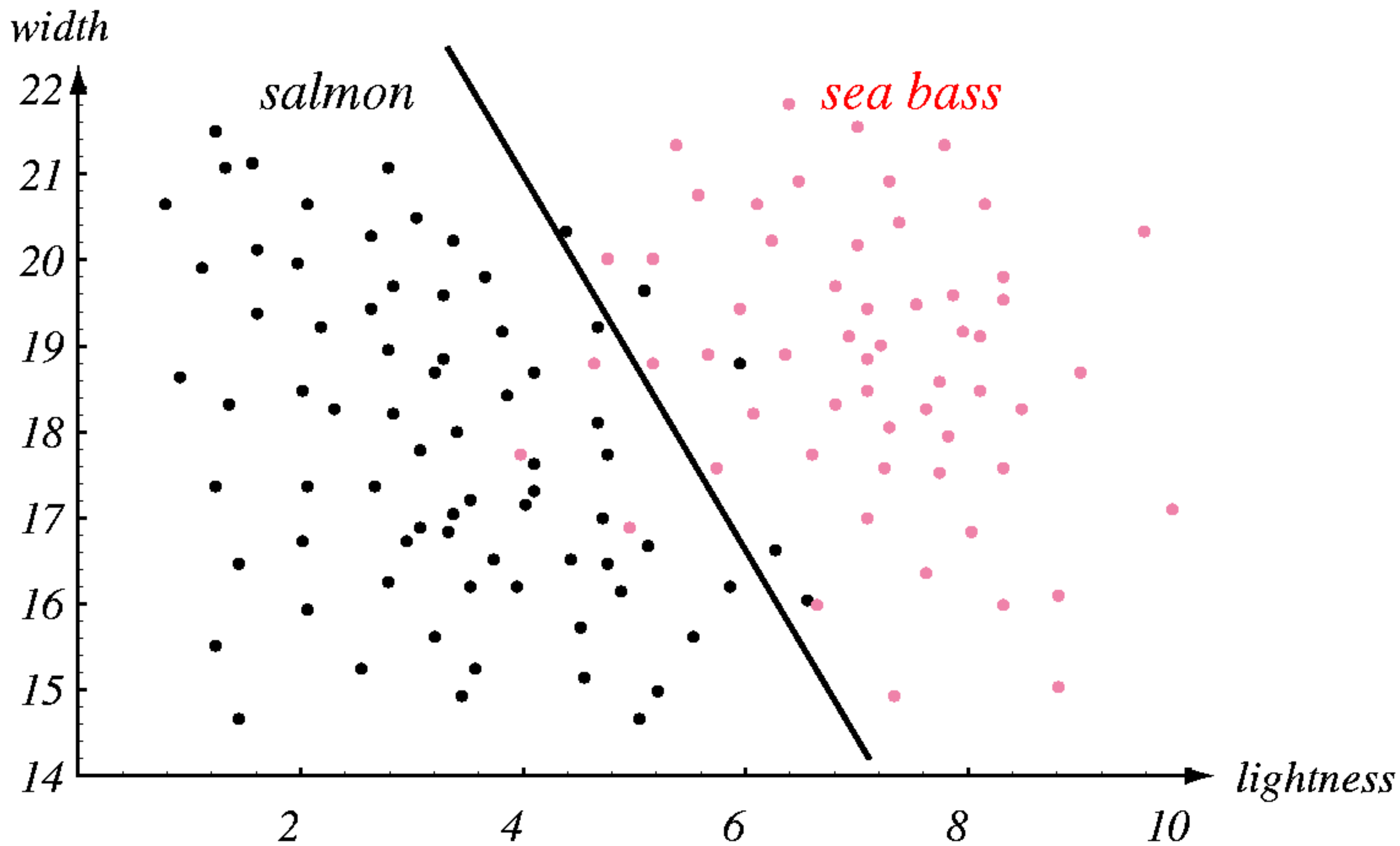
Summary of concepts and facts

- Prior probability
- Class conditional probability density function, likelihood
- Posterior probability
- Bayes formula/rule for posteriors
- Bayes decision rule
- Minimum cost/loss/risk classification
- Dealing with missing features

Naïve Bayes pdf estimation

The central problem in the Bayesian approach to classification: How to estimate class conditional probabilities

Estimation of pdf's 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.

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 | word1, word2 … word n) =
p(word1 | spam) p(word2 | spam) … p(word n | spam) P(spam) /
p(word1, word2 … word n)

P(non-spam | word1, word2 … word n) =
p(word1 | non-spam) p(word2 | non-spam) … p(word n | non-spam) P(non-spam) /
p(word1, word2 … word n)

P(spam | word1, word2 … word n) / P(non-spam | word1, word2 … word n) =

(p(word1 | spam)/p(word1 | non-spam)) … (p(word n | spam)/p(word n | non-spam))
(P(spam) / P(non-spam))

Naive Bayes classifier – Example: Spam filter

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

P(spam | viagra, purchase, love, romantic, happy) / P(non-spam | viagra, purchase, love, romantic, happy) =
(p(viagra | spam)/p(viagra | non-spam)) (p(purchase | spam)/p(purchase | non-spam))
(p(love | spam)/p(love | non-spam)) (p(romantic | spam)/p(romantic | non-spam))
(p(happy | spam)/p(happy | non-spam)) (P(spam) /P(non-spam)) =
1000*100*10*0.01*0.1*5 = 5000