Pattern Recognition

Hertentamen 2, August 27, 2004, 14:00–17:00

The problems are to be solved within 3 hrs. **The use of supporting material (books, notes, calculators) is not allowed.** In each of the five problems you can achieve up to 2 points, with a total maximum of 10 points. The exam is "passed" with 5.5 or more points.

---

1. Basics

a) Define and explain in words the terms *likelihood, prior, evidence,* and *posterior*. What is the relation between these quantities?

b) Consider the following sets of feature vectors, representing
class 1: $S_1 = \{(2,6),(3,4),(3,8),(4,6)\}$ and
class 2: $S_2 = \{(3,0),(3,-4),(1,-2),(5,-2)\}$, respectively.
They originate from two two-dimensional normal distributions. Compute the covariance matrices for each class from the sample data and write down the corresponding bivariate normal densities. (You can use the naive ML-estimates of the covariances.)

c) Assuming equal prior probabilities, evaluate the optimal decision boundary between the classes based on the densities obtained in part b).

2. Minimum error rate classification

Consider a simple, binary classification problem which is based on a single feature x. Assume that the corresponding class conditional probabilities are

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-6)^2\right] \quad \text{and} \quad p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-10)^2\right].$$

The classifier decides for ω_1 if $x < x^*$ and else decides for ω_2.

a) Which value of the decision boundary x^* gives the lowest expected classification error if the prior probabilities are $P(\omega_1) = P(\omega_2) = 1/2$? Visualize the situation, i.e. sketch the class conditionals and mark x^*.

b) Assume the value x^* from part a) is used, but the the true priors are $P(\omega_1) > 1/2$ and $P(\omega_2) = 1 - P(\omega_1)$. Does the expected classification error increase, decrease, or remain the same as in the case $P(\omega_1) = 1/2$?

c) Is the optimal boundary for $P(\omega_1) > 1/2$ greater or smaller then x^* for $P(\omega_1) = 1/2$?

Remarks:
Explicit calculations are not necessary here. You can exploit symmetry and use plausibility arguments, instead. However, it is not sufficient to "guess" the correct results, explain your answers!

3. Density estimation

a) Define and explain Maximum Likelihood (ML) estimation in the context of density estimation.

b) What are the ML estimates of the mean and standard deviation in case of a unidimensional normal distribution as obtained from sample data $\{x_1, x_2, \ldots, x_n\}$? (Just write down the estimates, you don't have to show that they maximize the likelihood.)

c) The ML estimate of the standard deviation is a so–called *biased estimate.* Explain precisely what this means (you don't have to prove that the estimate is biased). Write down an alternative, unbiased estimate of the standard deviation.

4. Kullback–Leibler divergence

An important measure of the difference between two distributions in the same space is the so–called *Kullback–Leibler (KL) divergence.* For two densities $p_1(x)$ and $p_2(x)$ (real random number x) it is defined as

$$D_{KL}\left[p_1(x), p_2(x)\right] = \int_{-\infty}^{\infty} p_1(x) \ln\left(\frac{p_1(x)}{p_2(x)}\right) dx$$

a) Suppose we want to approximate an arbitrary distribution $p_1(x)$ by a normal $p_2 = N(\mu, \sigma^2)$ with adjustable mean value μ and variance σ^2. Show explicitly (by calculation, not hand-waving arguments) that the "obvious" choice

$$\mu = \epsilon_1[x] \quad \text{and} \quad \sigma^2 = \epsilon_1[(x-\mu)^2]$$

satisfies the necessary conditions for minimizing the KL divergence. Here, ϵ_1 denotes the expectation over p_1.

b) One can show that the KL divergence is non-negative (you don't have to show it). Hence, it is sometimes called the *KL distance.* Explain why this "distance" is not a metric in the space of distributions $p(x)$. It is sufficient to argue that one of the properties of metrics is violated.

5. Competitive Learning

a) What is the purpose of *On–line competitive learning*? Present the *Winner Takes All* algorithm in terms of a "pseudocode computer program" and sketch an example scenario for a two–dimensional feature space.

b) Suggest a variation of *On–line competitive learning* in which the number of prototype vectors can increase and decrease in the course of learning.