Tentamen: Introduction to Pattern Recognition for AI (NB054B) (English translation)

6 July 2007, 09:00-12:00

Write your ***name and student number at the top of each sheet****. On each page, indicate page number and total number of pages.*
Please, write clearly! *Make sure to properly motivate all answers, and do not forget to include intermediate steps in your calculations: even if your final answer is wrong, you may still gain some points in that way. You may refer to the Bishop book for relevant equations, etc. One personal "cheat sheet" (a single A4 paper sheet) is allowed.*

Assignment 1

A factory produces intermediate products X. 40% of the intermediate products has quality $x = 1$ and the rest has quality $x = 2$. There is a test Z to assess the quality of these intermediate products. The result of Z is a number between 0 and 1. Let's assume that the test result, dependent on the quality is distributed as follows:

$$p(z|x=1) = \frac{1}{C_1}(1-z^2)$$
$$p(z|x=2) = \frac{1}{C_2}(1+z)$$

for $0 \leq z \leq 1$. C_1 and C_2 are constants. $p(z|x) = 0$ if z is outside of the given interval.

Question 1.1 *Show that $C_1 = \frac{2}{3}$ and $C_2 = \frac{3}{2}$*

Question 1.2 *Use Bayes' rule to compute the probability of quality $x = 1$ and $x = 2$ for the case that the test result is $z = 0$. Do the same for the case that the test result is $z = 1$*

Assignment 2

Given the following probability distribution

$$p(x, k|\mu, \sigma^2, \alpha) = p(x|k, \mu, \sigma^2)p(k|\alpha) \tag{1}$$

with $x \in \mathbb{R}$ and 'classes' $k \in \{1, ..., K\}$. The probability distribution has parameters $\mu = (\mu_1, ..., \mu_K), \sigma^2$ and $\alpha = \alpha_1, ..., \alpha_K$, where $\alpha_k \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$

The classes are distributed following $p(k|\alpha) = \alpha_k$. The class-conditional densities are Gaussian distributions with the same variance σ^2, which means:

$$p(x|k, \mu\sigma^2) = \mathcal{N}(x|\mu_k, \sigma^2) \tag{2}$$

Given the trainingset $\{x_n, k_n\}$ with $n = 1, ..., N$. The data points were sampled independent out of the distribution.
Question 2.1 *Show that the negative log-likelihood for this data is given by*

$$E = \frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{n=1}^{N}\delta_{kk_n}(x_n-\mu_k)^2 + N\log\sigma + \frac{N}{2}\log(2\pi) - \sum_{k=1}^{K}N_k\log\alpha_k \tag{3}$$

with $N_k = \sum_{n=1}^{N}\delta_{kkN}$ which means: the amount of data point with $k_n = k$

Question 2.2 *Show by minimalization of E that the maximum likelihood estimates of μ_k, σ^2 and α_k are given by*

$$\begin{aligned}\mu_k &= \frac{1}{N_k}\sum_{n=1}^{N}\delta_{kk_n}x_n \\ \sigma^2 &= \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}\delta_{kk_n}(x_n-\mu_k)^2 \\ \alpha_k &= \frac{N_k}{N}\end{aligned}$$

Assignment 3

Question 3.1 *Discuss the advantages and disadvantages of the maximum likelihood method regarding Bayesian inference.*

Assignment 4

Given the distribution $p(t|\lambda) = \lambda\exp(-t\lambda)$ for $t > 0$. The distribution is parameterized by $\lambda > 0$.
Question 4.1 *Show that*

$$\int_0^\infty p(t|\lambda)dt = 1 \tag{4}$$

Assume we have data $D = \{t_n\}_{n=1}^N$ with mean $\frac{1}{N}\sum_{n=1}^N t_n = \langle t\rangle$

Question 4.2 *Show that the maximum likelihood solution is given by $\lambda = \frac{1}{\langle t\rangle}$*

Now we are going to perform Bayesian inference. We choose $p(\lambda) \propto \lambda^\nu\exp(-\nu\tau\lambda)$ as prior with hyperparameters ν and τ.

Question 4.3 *Show that the posterior can also be written as*

$$p(\lambda|D) \propto \lambda^{N+\nu}\exp\big(-(N\langle t\rangle + \nu\tau)\lambda\big) \tag{5}$$

The prior and posterior are Gamma distributions (see Bishop page 688). The Gamma distribution has two parameters $a > 0$ and $b > 0$ and has the following form

$$Gamma(\lambda|a,b) \propto \lambda^{a-1}\exp(-b\lambda) \tag{6}$$

By comparing (5) and (6) you can see that the posterior is actually a Gamma distribution with $a = N+\nu+1$ and $b = N\langle t\rangle + \nu\tau$

In this particular case we can calculate the posterior exactly, or to phrase it a bit better: seeking for it. In most cases this is not possible and we have to approximate the posterior. For example, this can be done with the Laplace approximation as we can see in the following questions. But in this case where we have knowledge about the exact posterior, we can see how good Laplace approximation actually is, i.e., for comparing statistical properties of the distributions.
The Laplace approximation of the posterior is a Gaussian approximation around the MAP solution (as well the mode λ_o of the distribution),

$$q(\lambda|D) = \mathcal{N}(\lambda|\lambda_0, s^2) \tag{7}$$

Question 4.4 *The mode of $Gamma(\lambda|a,b)$ is*

$$\lambda_0 = \frac{a-1}{b} \tag{8}$$

Verify this by maximizing $\lambda^{a-1}\exp(-b\lambda)$ *with regards to* λ. *Why is the normalization in this case not important?*
Furthermore verify that the mode of the posterior $p(\lambda|D)$ *is given by*

$$\lambda_0 = \frac{N+\nu}{N\langle t\rangle + \nu\tau} \tag{9}$$

Question 4.5 *Describe how the variance in the Laplace approximation is calculated and show that this is given by*

$$s^2 = \frac{N+\nu}{(N\langle t\rangle + \nu\tau)^2} \tag{10}$$

Now we take a look at the quality of the approximation. The expected value and variance of $Gamma(\lambda|a, b)$ are

$$\mathbb{E}[\lambda] = \frac{a}{b}$$
$$var[\lambda] = \frac{1}{b^2}$$

Question 4.6 *Out of convenience we assume that* $\nu \approx 0$, *so we can neglect* ν *with regards to* N. *Compare the approximation of* $\mathbb{E}[\lambda], var[\lambda]$ *according to the Laplace approximation of the posterior (5) with their exact values. Do this by expressing the approximation in terms of the exact values, which means*

$$\mathbb{E}_{Laplace}[\lambda] = (1 + \epsilon(N, \langle t\rangle))\mathbb{E}_{Exact}[\lambda]$$
$$var_{Laplace}[\lambda] = (1 + \eta(N, \langle t\rangle))var_{Exact}[\lambda]$$

What can you say about the quality of the approximation as a function of N *and* $\langle t\rangle$*? Finally, how good is the approximation of the mode* λ_0*?*