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* $C_2 = \frac{3}{2}$

Answer:

C_1 and C_2 are normalizing constants.

$$\int 1 - z^2 dz = (z - \frac{1}{3}z^3)\Big|_0^1 = \frac{2}{3}$$
$$\int 1 + z dz = (z + \frac{1}{z}z^2)\Big|_0^1 = \frac{3}{2}$$

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* $z = 1$

Answer:

Bayes' rule:

$$(x_1|z) = \frac{p(z|x_1)p(x_1)}{p(z|x_1)p(x_1) + p(z|x_2)p(x_2)}$$

Filling in $z = 0$

$$p(x_1|z = 0) = \frac{\frac{3}{2}\frac{2}{5}}{\frac{3}{2}\frac{2}{5} + \frac{2}{3}\frac{3}{5}} = \frac{\frac{3}{5}}{\frac{3}{5} + \frac{2}{5}} = 0.6$$

Therefore $p(x_2|z = 0) = 1 - p(x_1|z = 0) = 0.4$.
Filling in $z = 1$. Because $p(z = 1|x = 1) = 0$ and $p(z = 1|x = 2) \neq 0$ the measured value can only generated by $x = 2$. Therefore $p(x_1|z = 1) = 0$ and $p(x_2|z = 1) = 1$

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 $k_n = k$

Answer:

Likelihood is

$$p(\{x_n, k_n\}_{n=1}^{N}) = \prod_{n=1}^{N} p(x_n, k_n|\mu, \sigma^2, \alpha)$$

The negative loglikelihood is

$$\begin{aligned}
E &= -\sum_{n=1}^{N} \log p(x_n, k_n|\mu, \sigma^2, \alpha) \\
&= -\sum_{n=1}^{N} \log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \Big(-\frac{(x_n - \mu k_n)^2}{2\sigma^2} \Big) \alpha k_n \Big) \\
&= -\sum_{n=1}^{N} \log(\frac{1}{\sqrt{2\pi}}) - \sum_{n=1}^{N} \log \frac{1}{\sqrt{\sigma^2}} - \sum_{n=1}^{N} \log \exp \Big(-\frac{(x_n - \mu k_n)^2}{2\sigma^2} \Big) - \sum_{n=1}^{N} \log \alpha k_n \\
&= N \log(\sqrt{2\pi}) + N \log \sigma - \sum_{n=1}^{N} \Big(-\frac{(x_n - \mu k_n)^2}{2\sigma^2} \Big) - \sum_{n=1}^{N} \log \alpha k_n \\
&= \frac{N}{2} \log(2\pi) + N \log \sigma + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu k_n)^2 - \sum_{n=1}^{N} \log \alpha k_n
\end{aligned}$$

Use the general fact that $F(k_n) = \sum_{k=1}^{K} \delta_{kk_n} F(k)$. Therefore:

$$(x_n - \mu k_n) = \sum_{k=1}^{K} \delta_{kk_n} (x_n - \mu k)^2$$

And therefore:

$$
\begin{aligned}
E &= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu k_n)^2 - \sum_{n=1}^{N}\log\alpha k_n \\
&= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu k_n)^2 - \sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}\log\alpha k_n \\
&= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu k_n)^2 - \sum_{k=1}^{K}\sum_{n=1}^{N}\delta_{kk_n}\log\alpha k_n \\
&= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu k_n)^2 - \sum_{k=1}^{K}N_k\log\alpha k
\end{aligned}
$$

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

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

Answer:

You can find μ_k, σ^2 and α_k by solving

$$
\begin{aligned}
\frac{\partial E}{\partial \mu_k} &= 0 \\
\frac{\partial E}{\partial \sigma^2} &= 0 \\
\frac{\partial E}{\partial \alpha_k} &= 0
\end{aligned}
$$

Keep in mind that α_k has the constraint that $\sum_k \alpha_k = 0$. Furthermore you should remember this general rule

$$
\frac{\partial}{\partial x_k}\sum_{k'=1}^{K}F(x'_k) = \frac{\partial}{\partial x_k}F(x_k)
$$

Partial derivative with regards to μ_k:

$$
\frac{\partial E}{\partial \mu_k} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\delta_{kk_n}(x_n - \mu_k)
$$

Zero crossings (Nullstellen):

$$
\sum_{n=1}^{N}\delta_{kk^n}(x^n - \mu_k) = 0
$$

Therefore

$$
\sum_{n=1}^{N}\delta_{kk^n}\mu_k = \sum_{n=1}^{N}\delta_{kk^n}x^n
$$

With the definition of N_k that is

$$N_k \mu_k = \sum_{n=1}^{N} \delta_{kk^n} x^n$$

Therefore

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \delta_{kk^n} x^n$$

The same for σ

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

Zero crossings/Nullstellen, multiplying both sides with σ^3 and dividing by N yields the asked result.

Finally α_k: use the constraint for the lagrangemultiplier λ, construct Lagrangian $E+\lambda(\sum_k \alpha_k - 1)$ Minimalize with regards to $P(k)$ and solve λ out of the constraints

$$\frac{\partial E + \lambda(\sum_k \alpha_k - 1)}{\partial \alpha_k} = \frac{N_k}{\alpha_k} + \lambda$$

This leads to:

$$\alpha_k = \lambda N_k$$

From $\sum_k \alpha_k = 1$ follows $\lambda = \sum_k N_k = N$, and from that follows the asked result.

Assignment 3

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

Answer:

Maximum likelihood gives a single parameter (vector), namely that one that describes/explains/fits the dataset the best. Bayesian inference on the other hand gives a complete distribution of parameters given the data, and therefore takes the uncertainty in the parameters into acount.
Advantages of ML:

- Maximum likelihood is simpler
  - Working with a parameter vector is much simpler than struggling with a whole distribution
  - The computation of the ML-solution is the minimalization of an error function. This is in general numerically easy to do. Bayesian inference on the other hand requires the computation of a posterior distribution. Normally, this is still analytically feasible in the most basic cases. However, numerical integration in a high dimensional parameter space is more difficult and costs alot more computing time than the numerical minimalization of an error function.
- Maximum likelihood does not need a prior. This can be an advantage because you do not have to think about a suiting prior.
- If the amount of data available is relatively large, then the result of ML and Bayes are identical and ML is the most simple alternative.

Disadvantages of ML:

- Maximum likelihood does not take parameter uncertainty into account and therefore has probably too much confidence in the parameters that it computes. If there is only a small amount of data available compared to the complexity of the model then this could lead to overfitting. On the other hand, Bayes allows much more complex models because Bayesian inference automatically regularizes.
- Because Bayesian inference carries a whole distribution with it, you can also inspect other statistics, i.e., the uncertainty of the parameters. This could be important to answer questions such as if the model needs more data for training.
- In Bayesian inference it is possibe to take different model classes of different complexities. This is more tedious in ML. ML would then choose just one parameter from the most complex model (the one that fits the data best).
- Bayesian inference requires a prior and forces you to think about the model assumptions. In the case of ML you do not have to think about that.

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*

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

Answer:

$$\int_0^\infty \lambda \exp(-t\lambda)dt = -\exp(-t\lambda)\Big|_{t=0}^{\infty} = 0 - (-1) = 1$$

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* $\lambda = \frac{1}{\langle t \rangle}$

Answer:

The negative likelihood is

$$E(\lambda) = -\sum_n \log(\lambda \exp(-t_n\lambda)) = -N\log(\lambda) + \sum_n t_n\lambda = -N\log(\lambda) + N\langle t \rangle\lambda$$

Taking the derivative

$$\frac{dE}{d\lambda} = -\frac{N}{\lambda} + N\langle t \rangle$$

Zero crossings (Nullstellen) are giving the answer.

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*

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

Answer:

Application of Bayes' rule

$$p(\lambda|D) \propto p(D|\lambda)p(\lambda)$$

Likelihood is

$$p(D|\lambda) = \prod_n \lambda \exp(-t_n\lambda) = \lambda^N \exp(-\sum_n t_n\lambda) = \lambda^N \exp(-N\langle t\rangle\lambda)$$

Prior is

$$p(\lambda) \propto \lambda^\nu \exp(-\nu\tau\lambda)$$

Multiplying with each other gives

$$p(D|\lambda)p(\lambda) \propto \lambda^N\lambda^\nu \exp(-N\langle t\rangle)\exp(-\nu\tau\lambda)$$

Cancelling the exponents gives the final result.

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*

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

Answer:

Take the derivative with regards to λ and calculate zero crossings (equal to zero/ Nullstellen berechnen).

$$\frac{d}{d\lambda}Gamma(\lambda|a,b) = (a-1)lambda^{a-2}\exp(-b\lambda) - b\lambda^{a-1}\exp(-b\lambda) = 0$$

$\lambda^{a-2}\exp(-b\lambda)$ is positive, therefore

$$(a-1) - b\lambda = 0$$

which is the same as

$$\lambda = (a-1)/b$$

Normalization is constant in λ. If $\lambda*$ is the maximum of $f(\lambda)$ then it is also the maximum of $\alpha f(\lambda)$. Filling in $a = N + \nu + 1$ and $N\langle t\rangle + \nu\tau$ in $\lambda = (a-1)/b$ gives the answer.

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

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

Answer:

If $p(\lambda) \propto f(\lambda)$ then the Laplace approximation is calculated by determining the maximum λ_0 of $f(\lambda)$. Therefore, determining the variance is done by the second derivative of $\log f$ at λ_0. Then the variance is

$$s^2 = \frac{1}{d^2/d\lambda^2 \log f(\lambda_0)}$$

Here: $g = \log f = (a-1)\log\lambda - b\lambda$. First derivative is $g' = (a-1)/\lambda - b$. Second derivative is $g'' = -(a-1)/\lambda^2$, therefore $s^2 = \lambda_0^2/(a-1)$ Filling in $\lambda_0 = (a-1)/b$ gives us $s^2 = (a-1)/b^2$. Filling in $a = N + \nu + 1$ and $b = N\langle t\rangle + \nu\tau$ gives us the result.

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?
Answer:

We neglect terms with ν. Filling in $a = N + 1$ and $b = N\langle t\rangle$ gives

$$\mathbb{E}_{Exact}[\lambda] = \frac{N+1}{N\langle t\rangle}$$
$$var_{Exact}[\lambda] = \frac{N+1}{N^2\langle t\rangle^2}$$

The Laplace values are λ_0 and s^2, therfore

$$\mathbb{E}_{Laplace}[\lambda] = \frac{N}{N\langle t\rangle}$$
$$var_{Laplace}[\lambda] = \frac{N}{N^2\langle t\rangle^2}$$

Or

$$\mathbb{E}_{Laplace}[\lambda] = (1 - \frac{1}{N+1})\mathbb{E}_{Exact}[\lambda]$$
$$var_{Laplace}[\lambda] = (1 - \frac{1}{N+1})var_{Exact}[\lambda]$$

Therefore the Laplace approximation gives a little underestimation of the mean and variance. The relative error decreases proportionally if N decreases. The mode is correct.