Tentamen: Statistical Machine Learning (NB054E)

21 January 2015, 08:30-11:30 in HG00.068

*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.*

Total possible points: 80

Assignment 1 (18pt)

A factory produces *fliggrs* X. Seventy-five percent (75%) of the fliggrs has quality $x = 1$ and the rest has quality $x = 0$. Unfortunately, the quality of fliggrs cannot be determined directly. (To asses these probabilities, destructive methods have been applied).
There is a possibility to test fliggr quality with a new type of test Y . A test result can be positive ($y = 1$) or negative ($y = 0$). Studies have shown that 40% of fliggrs with quality $x = 1$ have a positive test result. However, 30% of fliggrs with $x = 0$ also test positive.

Question 1.1 (4pt) *What are the following probabilities according to the above stated description?*

i $P(x = 1)$

ii $P(y = 1|x = 1)$

iii $P(y = 1|x = 0)$

iv $P(y = 1)$

Question 1.2 (6pt) *Compute, using Bayes' rule, the probability of quality $x = 1$ if the test result is positive. Do the same for the case that the test result is negative.*

The test Y is cheap, but still there are costs (per fliggr) involved. The question is whether the test is economically beneficial, i.e. whether its use increases the expected profit.

1. If a fliggr of quality $x = 1$ is correctly classified, it yields a profit of €40.-
2. If a fliggr is classified as $x = 0$, it yields a profit of €18.-, regardless of its true quality.
3. If a fliggr with true quality $x = 0$ is wrongly classified as $x = 1$, it causes a loss of €60.-

Without the test we could either adopt a policy of classifying every fliggr as $x = 1$ or as $x = 0$.

Question 1.3 (4pt) *What is the expected profit per fliggr under optimal classification, without implementation of the test?*

Using test Y we could adopt a more refined policy (dependent on the outcome of the test) to maximize the expected profit per fliggr. There are now four different classification policies: if the test result is positive ($y = 1$), we can either classify the fliggr as $x = 1$ or as $x = 0$, and if the test result is negative ($y = 0$), we also can choose between classifying the fliggr as $x = 1$ or as $x = 0$.

Question 1.4 (4pt) *Compute the expected profit per fliggr, assuming optimal classification dependent on the outcome of the test. What is the maximum price of the test per fliggr, if the test is to be economically beneficial?*

Assignment 2 (24pt)

The distribution of the number of occurences in a fixed period of time in systems with a large number of possible events, each of which is relatively rare, is modelled by the *Poisson* distribution. Examples are the number of accidents in a week for a certain stretch of road, or the number of photons per second received by a detector from a distant star. The probability of k such events in a fixed period of time is given as

$$Pois(k|\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!} \tag{1}$$

with parameter $\lambda > 0, k \in \{0, 1, 2, ...\}$ and $k!$ the factorial of k.

Question 2.1 (4pt) *Verify that the Poisson distribution (1) represents a proper probability distribution. Hint: use the fact that* $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$. *Hint 2: do not confuse the variable with the parameter!*

For a 1 km stretch of road the number of accidents per week has been recorded over a period of several months, resulting in a data set $\mathbf{X} = \{x_1, ..., x_N\}$. We assume the data can be modelled as independent samples from a Poisson distribution and want to obtain an estimate for λ

Question 2.2 (6pt) *Show that the log-likelihood of* λ *for this data set is given by*

$$\ln p(\mathbf{X}|\lambda) = -N\lambda + K \ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) \tag{2}$$

with $K = \sum_{i=1}^{N} x_i$.

Question 2.3 (4pt) *From (2) Show that the maximum likelihood estimate* λ_{ML} *is given by*

$$\lambda_{ML} = \frac{K}{N} \tag{3}$$

Extra Question *Is* λ_{ML} *a biased or unbiased estimator? Why?*

The recorded number of accidents per week over a 2 month period was as follows:

$$\mathbf{X} = \{4, 1, 0, 5, 2, 3, 0, 1\}$$

Question 2.4 (2pt) *Calculate* λ_{ML} *for this data set.*

Background information based on observations for similar types of roads has resulted in the following *Gamma* distribution as prior over λ per kilometer of road:

$$Gam(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{4}$$

with hyperparameters $a = 3$ and $b = 2$ (see 2.146 in Bishop). The Gamma distribution is the conjugate prior to the likelihood function for the parameter λ in the Poisson distribution, meaning that the posterior has the same functional form as the prior.

Question 2.5 (4pt) *Show that with the Gamma prior (4) and a dataset* $\boldsymbol{X}$*, the posterior distribution of* λ *takes the form*

$$p(\lambda|\mathbf{X}) = Gam(\lambda|a + K, b + N) \tag{5}$$

Hint: In the derivation, you can ignore factors not involving λ

Question 2.6 (4pt) *Use the posterior distribution (5), together with (B.26-29) to obtain*

- *the Bayesian maximum a posteriori estimate* λ_{ML}*, i.e., the mode of the posterior distribution of* λ
- *the posterior expected value* $\mathbb{E}[\lambda|\boldsymbol{X}]$

for the given stretch of road. Compare with the maximum likelihood estimate λ_{ML}*; why are these estimates lower than the maximum likelihood estimate?*

Assignment 3 (18pt)

Consider a probability distribution $p(u, v, w) = p(u)p(v)p(w|u, v)$

Question 3.1 (4pt) *Show this implies that variable u is independent of v. (If you think that this makes it easier, you may assume that all three variables are discrete).*

We now consider a multivariate Gaussian probability distribution $p(u, v, w) = p(u)p(v)p(w|u, v)$ defined in terms of conditional distributions as

$$p(u) = \mathcal{N}(u|\alpha, \rho^2) \tag{6}$$
$$p(v) = \mathcal{N}(v|\beta, \sigma^2) \tag{7}$$
$$p(w|u, v) = \mathcal{N}(w|\gamma(u + 2v), \tau^2) \tag{8}$$

with $\alpha, \beta, \gamma, \rho, \sigma$ and τ constant model parameters.

We are looking for the marginal distribution $p(w)$. Unfortunately, the equations (2.113-2.117) only consider the relation between two variables. To deal with this, we view u and v as two partitioned Gaussian components of a single multivariate variable $\mathbf{x} = (u, v)^T$, and write the distribution over this new variable as $p(x) = \mathcal{N}(x|x_0, \Sigma)$

Question 3.2 (6pt) *Give an expression for the mean x_0 and covariance Σ in $p(x)$ in terms of the model parameters.*

The joint distribution can now be written in the form $p(x, w) = p(x)p(w|x)$. For the conditional distribution we write $p(w|x) = \mathcal{N}(w|Ax, L^{-1})$

Question 3.3 (8pt) *(a) Give an expression for $\boldsymbol{A}$ and $\boldsymbol{L}^{-1}$ in $p(w|x)$ in terms of the model paramters.*
(b) Use this to obtain an expression for the mean and variance of the marginal distribution $p(w) = \mathcal{N}(w|w_0, \sigma_w^2)$ in terms of the model parameters.

Assignment 4 (20pt)

In a recent survey under master students at the RU, we collected a data set of 400 records of 3 variables $\mathbf{x} = \{gender, IQ, haircolour\}$, in which each value is represented as an integer number. We would like to use the 'kernel trick' to analyse this data.

Question 4.1 (4pt) *For this data set, show that*

$$k(x, y) = (x^T y)^2 + 4x^T y + 1 \tag{9}$$

is a valid kernel function. What are the dimensions of the corresponding kernel matrix and the number of implied features?

As this data set is too ambituous to tackle by hand in an exam, we als have another, more modest data set of observations: $(x_1, t_1) = (-1, 0)$, and $(x_2, t_2) = (1, 1)$. We assume that there is some underlying function $y_i = f(x_i)$, for which we have noisy observations $t_i = y_i + \epsilon_i$ governed by independent Gaussian noise, with precision parameter $\beta = 2$:

$$p(t_i|y_i) = \mathcal{N}(t_i|y_i, \beta^{-1}) \tag{10}$$

We want to know the value t we can expect to observe at $x = 0$. We decide to use a Gaussian process (GP), with a standard Gaussian kernel defined as

$$k(x, x') = \theta_0 \exp\left\{-\frac{\theta_1}{2}||x - x'||^2\right\} \tag{11}$$

with $\theta_0 = 1, \theta_1 = \ln(2) \approx 0.6931$

Question 4.2 (4pt) *In this GP approach, the marginal distribution $p(t)$ over $t = (t_1, t_2)^T$ (before the actual observation $t_1 = 0, t_2 = 1$), conditioned on input values $x_1 = -1, x_2 = 1$, takes the form of a multivariate Gaussian. Compute the mean μ and covariance matrix Σ of this distribution.*

Question 4.3 (6pt) *On observing $t_1 = 0, t_2 = 1$, the resulting probability distribution for the observation at $x = 0$ is again Gaussian. Compute mean and covariance for this distribution.*

Hint: remember that

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Question 4.4 (4pt) *Does the maximum of the expected mean pass through the data points $(-1, 0)$ and $(1, 1)$? If so, explain why this is logical for the given data set; if not, explain what should be changed to make it so.*

Question 4.5 (2pt) *What happens to the predictive mean of our GP regression result with Gaussian kernel as $x \to \pm\infty$? How does this differ from the solution for the Bayesian linear regression model $y(x, w) = w_0 + w_1 x$ (Bishop, §3.3) with a zero mean Gaussian weight prior?*