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.*

Assignment 1

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)* $P(y = 1)$

ANSWER:

$$\begin{aligned} P(x = 1) &= 0.75 \quad (1)\\ P(y = 1|x = 1) &= 0.4 \quad (2)\\ P(y = 1|x = 0) &= 0.3 \quad (3)\\ P(y = 1) &= 0.375 \quad (4) \end{aligned}$$

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.*

ANSWER:

$$\begin{aligned} P(x = 1|y = 1) &= \frac{P(y = 1|x = 1)P(x = 1)}{P(y = 1|x = 1)P(x = 1) + P(y = 1|x = 0)P(x = 0)}\\ &= \frac{0.4 \cdot 0.75}{0.4 \cdot 0.75 + 0.3 \cdot 0.25} = \frac{4}{5} = 0.8 \end{aligned}$$

and so $P(x = 0|y = 1) = 1 - P(x = 1|y = 1) = 0.2$

$$\begin{aligned} P(x = 1|y = 0) &= \frac{P(y = 0|x = 1)P(x = 1)}{P(y = 0|x = 1)P(x = 1) + P(y = 0|x = 0)P(x = 0)}\\ &= \frac{0.6 \cdot 0.75}{0.6 \cdot 0.75 + 0.7 \cdot 0.25} = \frac{18}{25} = 0.72 \end{aligned}$$

and so $P(x = 0|y = 0) = 1 - P(x = 1|y = 0) = 0.28$

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?*

ANSWER: (use c for the classification)

(i) $\text{Profit}(c = 1) = P(x = 1)€40 - P(x = 0)€60 = 0.75 * 40 - 0.25 * 60 = €15.-$

(ii) $\text{Profit}(c = 0) = €18$

Classifying as $x = 0$ is optimal. Expected profit is €18.-

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?*

ANSWER:

- *What is the expected profit of a fliggr if:*

  1. *(the test result is $y = 1$ and the quality is classified as $x = 1$,)*
     answ:

     $$\text{Profit}(c = 1|y = 1) = P(x = 1|y = 1)*€40 - P(x = 0|y = 1)*€60 = 0.8*40 - 0.2*60 = €20$$

  2. *(test result is $y = 1$ and the quality is classified as $x = 0$,)*
     answ:

     $$\text{Profit}(c = 0|y = 1) = €18$$

  (What is the optimal classification for a fliggr with test result $y = 1$.)

- *(What is the optimal classification for a fliggr with with test result $y = 0$ (and why)?)*
  answ: Without the test, the optimal classification is $c = 0$. With test result $y = 0$, the probability of $x = 1$ is even less than before doing the test. So $c = 0$ remains optimal.

- *(Compute the expected profit per fliggr, assuming optimal classification after doing the test. What is the maximum price of the test per fliggr, for the test being economically beneficial?)*
  answ: First compute $P(y = 1)$ and $P(y = 0)$

$$\begin{aligned} P(y = 1) &= P(y = 1|x = 1)P(x = 1) + P(y = 1|x = 0)P(x = 0) && (5) \\ &= 0.4 * 0.75 + 0.3 * 0.25 = 15/40 = 3/8 = 0.375 && (6) \end{aligned}$$

So $P(y=0) = 5/8 = 0.625$ In the case of $y = 1$, the profit is €20.- In the case of $y = 1$, the profit is €18,- So the expected profit is

$$\begin{aligned} ExpProf &= P(y=1) * €20 + P(y=0) * €18 \qquad (7)\\ &= 0.375 * €20 + 0.625 * €18 = €18.75 \qquad (8)\end{aligned}$$

With test: expected profit is €18.75, without test, expected profit is €18. Difference is €0.75, this is the maximum profitable price of the test.

With alternative expected profits:

$$\begin{aligned} ExpProf &= P(y=1) * €30 + P(y=0) * €20 \qquad (9)\\ &= 0.375 * €30 + 0.625 * €20 = €23.75 \qquad (10)\end{aligned}$$

Difference is €5.75

Assignment 2

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

$$\mathrm{Pois}(k|\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!} \qquad (11)$$

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

Question 2.1 *(4pt) Verify that the Poisson distribution* (11) *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!

ANSWER:
For discrete variables a proper probability mass function satisfies $p(x) \geq 0$ and $\sum_x p(x) = 1$. Since $k \geq 0$ (by definition: less than zero events makes no sense) and $\lambda > 0$ none of the elements in (11) *can become negative and so $p(x) \geq 0$. Normalizing over all possible events gives*

$$\begin{aligned} \sum_k Pois(k;\lambda) &= \sum_{k=0}^{\infty} \frac{\lambda^k \exp(-\lambda)}{k!} \\ &= \exp(-\lambda) \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= \exp(-\lambda)\exp(\lambda) \\ &= 1 \end{aligned}$$

and so (11) *is indeed a proper discrete probability distribution.*

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 $\mathbf{X} = \{x_1, \ldots, x_N\}$. 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*

$$\ln p(\mathbf{X}|\lambda) = -N\lambda + K\ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) \qquad (12)$$

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

ANSWER: For the likelihood we have

$$p(\mathbf{X}|\lambda) = \prod_{i=1}^{N} Pois(x_i; \lambda) \tag{13}$$

Taking the log, converting products to sums and collecting terms this gives

$$\begin{aligned}
\ln p(\mathbf{X}|\lambda) &= \ln \prod_{i=1}^{N} \left(\frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} \right) && (14) \\
&= \sum_{i=1}^{N} \Big(-\lambda + x_i \ln(\lambda) - \ln(x_i!) \Big) && (15) \\
&= -N\lambda + \left(\sum_{i=1}^{N} x_i \right) \ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) && (16) \\
&= -N\lambda + K \ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) && (17)
\end{aligned}$$

Question 2.3 *(4pt) From* (12) *Show that the maximum likelihood estimate* λ_{ML} *is given by*

$$\lambda_{ML} = \frac{K}{N} \tag{18}$$

ANSWER: Maximizing by taking the derivative of (12) *w.r.t.* λ *and setting equal to zero results in*

$$\begin{aligned}
0 &= \frac{d}{d\lambda} \ln p(\mathbf{X}|\lambda) \\
&= \frac{d}{d\lambda} \Big(-N\lambda + K \ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) \Big) \\
&= -N + \frac{K}{\lambda} - (0)
\end{aligned}$$

and so the maximum likelihood estimate for λ *is given by* $\lambda_{ML} = \frac{K}{N}$*, corresponding to the sample mean.*
ExtraQ: Is λ_{ML} *a biased or unbiased estimator? Why? An estimator is unbiased if its expectation is equal to the true value. For the expectation of the Poisson distribution we have*

$$\begin{aligned}
E[k] &= \sum_{k=0}^{\infty} k Pois(k; \lambda) \\
&= \sum_{k=0}^{\infty} k \frac{\lambda^k \exp(-\lambda)}{k!} \\
&= \lambda \exp(-\lambda) \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} \\
&= \lambda \exp(-\lambda) \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\
&= \lambda \exp(-\lambda) \exp(\lambda) \\
&= \lambda
\end{aligned}$$

where we again used the hint in Q2.1: $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$.
So, in this case each measurement has expectation λ*, and therefore the sample mean* $\frac{K}{N}$ *as well, which shows that* λ_{ML} *is an unbiased estimator.*

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

$$\mathbf{X} = \{4, 1, 0, 5, 2, 3, 0, 1\} \tag{19}$$

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

ANSWER: Freebie2: $\lambda_{ML} = \frac{K}{N}$*. For the given data set* $K = (4 + 1 + 0 + 5 + 2 + 3 + 0 + 1) = 16$ *and* $N = 8$*, resulting in* $\lambda_{ML} = 2$.

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:

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{20}$$

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* (20) *and a dataset* $\mathbf{X}$*, the posterior distribution of* λ *takes the form*

$$p(\lambda|\mathbf{X}) = Gam(\lambda|a + K, b + N) \tag{21}$$

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

ANSWER: A prior is conjugate to a distribution if the posterior has the same functional form as the prior. The posterior is equal to the likelihood times the prior times some normalization constant dependent on the data. The latter, as the hint states, can be ignored in the derivation as it is taken care of automatically by the functional form of the posterior.
From (13) *we have for the likelihood*

$$p(\mathbf{X}|\lambda) = \prod_{i=1}^{N} \left(\frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} \right)$$

multiplying by the Gamma distribution (20) *then gives*

$$\begin{aligned}
p(\lambda|\mathbf{X}) &\propto p(\mathbf{X}|\lambda)p(\lambda) \\
&= \prod_{i=1}^{N} \left(\frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} \right) \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \\
&= \frac{\lambda^{\sum_{i=1}^{N} x_i} \exp(-N\lambda}{\prod_{i=1}^{N} x_i!} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \\
&= \left(\frac{b^a}{\Gamma(a) \prod_{i=1}^{N} x_i!} \right) \lambda^{(a+(\sum_{i=1}^{N} x_i)-1)} \exp(-(b + N)\lambda) \\
&\propto \lambda^{(a+K-1)} \exp(-(b + N)\lambda) \\
&\propto Gam(\lambda|a + K, b + N)
\end{aligned}$$

The normalization constant follows from the standard form of the Gamma distribution.

Question 2.6 *(4pt) Use the posterior distribution* (21)*, together with (B.26-29) to obtain*

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

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

ANSWER: According to (21) *the posterior distribution takes the form*

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

For the given data set and prior information we have $a = 3$, $b = 2$, $K = 16$, $N = 8$. Filling in gives

$$p(\lambda|\mathbf{X}) = Gam(\lambda|19, 10) \tag{22}$$

From appendix B we see that the mode (=MAP) and expectation value of the Gamma distribution are given by $(a-1)/b$ and a/b respectively. In this case this corresponds to $\lambda_{MAP} = 1.8$ and $E[\lambda] = 1.9$. These are both slightly below the maximum likelihood estimate $\lambda_{ML} = 2$, which should not come as a surprise, as the background information by itself has an expectation value of $\lambda_{prior} = 3/2 = 1.5$, indicating that this roads seems more prone to accidents than usual. The fact that max. and expectation value for the Gamma distribution are not identical has no significance (Gamma is clearly a non-symmetric distribution).

Assignment 3

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).*

ANSWER: Comparing the standard factorization $p(u, v, w) = p(u)p(v|u)p(w|u, v)$ to the given $p(u, v, w) = p(u)p(v)p(w|u, v)$ it follows that $p(v|u) = p(v)$ which implies that v is independent of u and vice versa.

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:

$$\begin{aligned} p(u) &= \mathcal{N}(u|\alpha, \rho^2) && (23)\\ p(v) &= \mathcal{N}(v|\beta, \sigma^2) && (24)\\ p(w|u, v) &= \mathcal{N}(w|\gamma(u + 2v), \tau^2) && (25) \end{aligned}$$

with α, β, γ, ρ, σ 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 $\mathbf{x} = \big(u, v\big)^{\mathrm{T}}$, and write the distribution over this new variable as $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{x}_0, \mathbf{\Sigma})$.

Question 3.2 *(6pt) Give an expression for the mean $\mathbf{x}_0$ and covariance $\mathbf{\Sigma}$ in $p(\mathbf{x})$ in terms of the model parameters.*

ANSWER: The distribution of $\mathbf{x}$ can be written as $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{x}_0, \mathbf{\Sigma})$, with

$$\begin{aligned} \mathbf{x} &= \big(u, v\big)^T && (26)\\ \mathbf{x}_0 &= \big(\alpha, \beta\big)^T && (27)\\ \mathbf{\Sigma} &= \big(\rho^2, 0; 0, \sigma^2\big) && (28) \end{aligned}$$

(remember that u and v are independent components of $\mathbf{x}$).

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

Question 3.3 *(8pt) (a) Give an expression for $\mathbf{A}$ and $\boldsymbol{L}^{-1}$ in $p(w|\mathbf{x})$ in terms of the model parameters.*
(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.

ANSWER: (a) The variable w is a scalar that only depends on u and v through its mean $\gamma(u+2v)$, so conditional on $\mathbf{x}$ the distribution can be written as $p(w|\mathbf{x}) = \mathcal{N}(w|\mathbf{A}\mathbf{x}, \boldsymbol{L}^{-1})$, with

$$\begin{aligned} \mathbf{A} &= \left(\gamma, 2\gamma\right) & (29) \\ \boldsymbol{L}^{-1} &= \tau^2 & (30) \end{aligned}$$

(b) Using eq.2.115, with $\mathbf{y} = w$ (match other parameter as well), we find $p(w) = \mathcal{N}(w|w_0, \sigma_w^2)$ with

$$\begin{aligned} w_0 &= \gamma(\alpha + 2\beta) & (31) \\ \sigma_w^2 &= \boldsymbol{L}^{-1} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T & (32) \\ &= \tau^2 + \left(\gamma, 2\gamma\right)\left(\rho^2, 0; 0, \sigma^2\right)\left(\gamma, 2\gamma\right)^T & (33) \\ &= \tau^2 + \gamma^2(\rho^2 + 4\sigma^2) & (34) \end{aligned}$$

Assignment 4

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*

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^2 + 4\mathbf{x}^T\mathbf{y} + 1 \qquad (35)$$

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

ANSWER:
In the course we defined a valid' kernel function as a function that can be written as the inner product of some implied feature vector, i.e.: $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T\phi(\mathbf{y})$. In this case

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}^T\mathbf{y})^2 + 4\mathbf{x}^T\mathbf{y} + 1 \\ &= (x_1y_1 + x_2y_2 + x_3y_3)^2 + 4(x_1y_1 + x_2y_2 + x_3y_3) + 1 \\ &= x_1^2y_1^2 + x_2^2y_2^2 + x_3^2y_3^2 + 2x_1x_2y_1y_2 + 2x_1x_3y_1y_3 + 2x_2x_3y_2y_3 + 4x_1y_1 + 4x_2y_2 + 4x_3y_3 + 1 \\ &= (x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3, 2x_1, 2x_2, 2x_3, 1)^T \bullet \ldots \\ &\quad \ldots (y_1^2, y_2^2, y_3^2, \sqrt{2}y_1y_2, \sqrt{2}y_1y_3, \sqrt{2}y_2y_3, 2y_1, 2y_2, 2y_3, 1) \\ &= \phi(\mathbf{x})^T\phi(\mathbf{y}) \end{aligned}$$

This corresponds to 9 features + 1 bias constant in the implied feature vector $\phi(\mathbf{x})$ (or simply 10 features). The size of the corresponding kernel matrix equals the number of data points, i.e. $[400 \times 400]$.

Alternatively, one can argue that (6.12) in combination with (6.17) implies it is a valid kernel, and reason analogous to (6.12) that for 3 input variables this corresponds to 10 implied features: 3 squares, 3 cross-terms, 3 from the inner product and 1 constant.

As this data set is too ambitious 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 $\beta = 2$:

$$p(t_i|y_i) = \mathcal{N}(t_i|y_i, \beta^{-1}) \tag{36}$$

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

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left\{-\frac{\theta_1}{2}||\mathbf{x} - \mathbf{x}'||^2\right\} \tag{37}$$

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

Question 4.2 *(4pt) In this GP approach, the marginal distribution $p(\mathbf{t})$ over $\mathbf{t} = (t_1, t_2)^T$ (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 $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of this distribution.*

ANSWER:
A Gaussian process assumes an a priori zero-mean multivariate Gaussian over the output values (Bishop,6.60), so also on the noisy observations of those output values (Bishop, eq.6.61)

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = \mathbf{C}_2$ defined as in (6.62).

Filling in for (x_1, x_2) in $\mathbf{C}_2(i, j) = k(x_i, x_j) + \beta^{-1}\delta_{ij}$ with precision $\beta = 2$ then gives:

$$\begin{aligned}
\mathbf{C}_2(1,1) &= \exp -1/2\ln(2) * ||(-1+1)||^2 + \beta^-1\delta_{1,1} = 1 + 2^{-1} = 3/2 \\
\mathbf{C}_2(1,2) &= \exp -1/2\ln(2) * ||(-1-1)||^2 + \beta^-1\delta_{1,2} = 1/4 + 0 = 1/4 \\
\mathbf{C}_2(2,1) &= \mathbf{C}_2(1,2) \\
\mathbf{C}_2(2,2) &= \exp -1/2\ln(2) * ||(1-1)||^2 + \beta^-1\delta_{1,1} = 1 + 2^{-1} = 3/2
\end{aligned}$$

So that

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.5 & 0.25 \\ 0.25 & 1.5 \end{pmatrix} \tag{38}$$

Question 4.3 *(6pt) On observing $t_1 = 0, t_2 = 1$, the resulting probability distribution for the observation at $x = 0$ is again a Gaussian. Compute mean and covariance for this distribution.*
Hint: remember that $\mathbf{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{det(\mathbf{A})}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

ANSWER:
For that we can extend previous covariance matrix $\mathbf{C}_2$ with the additional point of interest $x_3 = 0$, and use this $\mathbf{C}_3$ to compute the implied distribution over t_3. Filling in in accordance with (6.65):

$$\mathbf{C}_3 = \begin{pmatrix} \mathbf{C}_2 & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix} \tag{39}$$

with

$$\begin{aligned}
\mathbf{k}(1) &= \exp -1/2\ln(2) * ||1||^2 + \beta^-1\delta_{1,3} = 1/\sqrt{2} + 0 \approx 0.7071 \\
\mathbf{k}(2) &= \mathbf{k}(1) \\
c &= \exp -1/2\ln(2) * ||0||^2 + \beta^-1\delta_{1,1} = 1 + 2^{-1} = 3/2
\end{aligned}$$

Then by (6.66+6.67) the resulting distribution over t at $x = 0$ is a Gaussian $p(t_3|(t_1, t_2) = \mathcal{N}(t_3|m(x_3), \sigma(x_3)$ given by

$$\begin{aligned} m(x_3) &= \mathbf{k}^T \mathbf{C}_2^{-1} \mathbf{t} \\ \sigma(x_3 &= c - \mathbf{k}^T \mathbf{C}_2^{-1} \mathbf{k} \end{aligned}$$

Computing the inverse: $\mathbf{C}_2^{-1} = \begin{pmatrix} 1.5 & 0.25 \\ 0.25 & 1.5 \end{pmatrix}^{-1} = \frac{4}{35} \begin{pmatrix} 6 & -1 \\ -1 & 6 \end{pmatrix} \approx \begin{pmatrix} 0.6857 & -0.1143 \\ -0.1143 & 0.6857 \end{pmatrix}$, *and filling in then gives:*

$$\begin{aligned} m(x_3) &= (1/\sqrt{2}, 1/\sqrt{2}) \frac{4}{35} \begin{pmatrix} 6 & -1 \\ -1 & 6 \end{pmatrix} (0, 1)^T & (40) \\ &\approx 0.4041 & (41) \\ \sigma(x_3 &= 1.5 - (1/\sqrt{2}, 1/\sqrt{2}) \frac{4}{35} \begin{pmatrix} 6 & -1 \\ -1 & 6 \end{pmatrix} (1/\sqrt{2}, 1/\sqrt{2})^T & (42) \\ &\approx 0.9286 & (43) \end{aligned}$$

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.*

ANSWER:
No: it is not a 'perfect interpolator' (see below). This is due to the zero mean 'prior' on $y(x)$ (Bishop, eq.6.60), in combination with the Gaussian observation noise ϵ_i with finite precision β, which tends to 'drag' the most likely outcome towards the zero (x-axis).

Eliminating observation noise by setting $\beta^{-1} = 0$ gets rid of this effect, and results in a solution that passes through each of the observed data points.

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 $y(x, \mathbf{w}) = w_0 + w_1 x$ (Bishop, §3.3) with a zero mean Gaussian weight prior?*

ANSWER:
With Bayesian linear regression the solution is a similar to a straight line $y = w_1 x + w_0$, where the prior has a regularizing impacts on the weights (tendency to drive a bit towards zero, resulting in a slightly flatter line a bit below average), but the solution remains a polynomial which means that $|y(x)|$ goes to infinity as x goes to $\pm\infty$.

With GPs on the standard Gaussian kernel the solution is also driven towards zero, but now over the entire x-axis, and is only 'drawn away' for points close to other measured points. It also means the answer is NOT a polynomial. In fact: based on the two data points, for arbitrary x the most likely solution t (corresponding to the mean of the conditional, Bishop eq.6.66) takes the form

$$\begin{aligned} E[t|x, \mathbf{t}] &= m(t|x, \mathbf{t}) = \mathbf{k}^T \mathbf{C}_2^{-1} \mathbf{t} \\ &= (k(x_1, x), k(x_2, x)) \frac{4}{35} (-1, 6)^T \\ &\approx -0.1143 * 2^{-\frac{1}{2}(x+1)^2} + 0.6857 * 2^{-\frac{1}{2}(x-1)^2}) \end{aligned}$$