Solutions to exercise 12, 20.4.2007

Problem 1.

In the lectures, the predictive distribution was given as

$$p(\tilde{y}|y) \propto \exp\left[-\frac{1}{2}\frac{1}{(c-k^TC^{-1}k)}(\tilde{y}-k^TC^{-1}y)^2\right].$$

We are asked to confirm this result. Here y is a vector of training data, $\tilde{y}$ is the scalar value we are trying to predict, and all other symbols will be defined shortly.

We can calculate the predictive distribution as

$$p(\tilde{y}|y) = p(y,\tilde{y})/p(y).$$

Here $p(y) = N(y|0,C)$ and $p(y,\tilde{y}) = N((y\ \tilde{y})|0,\tilde{C})$, where

$$\tilde{C} = \begin{bmatrix} C & k \\ k^T & c \end{bmatrix}.$$

Above, C is a $n \times n$ matrix, c is a scalar, and k is a $n \times 1$ vector.

Now we are able to employ the formulas given in the problem. We get

$$\begin{aligned} E(\tilde{y}|y) &= E(\tilde{y}) + \mathrm{Cov}(y,\tilde{y})(\mathrm{Var}(y))^{-1}(y-E(y)) \\ &= 0 + k^TC^{-1}(y-0) \\ &= k^TC^{-1}y, \end{aligned}$$

as required. Similarly,

$$\begin{aligned} \mathrm{Var}(\tilde{y}|y) &= \mathrm{Var}(\tilde{y}) - \mathrm{Cov}(y,\tilde{y})(\mathrm{Var}(y))^{-1}\mathrm{Cov}(\tilde{y},y) \\ &= c - k^TC^{-1}k. \end{aligned}$$

If we have n training points (the length of the vector y is n), the matrix C is of size $n \times n$ and its inverse takes $\mathcal{O}(n^3)$ multiplications to compute. All the other computations are at most $\mathcal{O}(n^2)$ so the total cost is $\mathcal{O}(n^3)$.

When the matrix C^{-1} has been computed once, it does not change when predicting new points. Only the vector k containing the covariances between the new point and all the training points changes. To compute the predictive mean, we only need an inner product $k^TC^{-1}y$ where $C^{-1}y$ is a fixed vector. This takes $\mathcal{O}(n)$ multiplications.

The predictive variance has a quadratic form $k^TC^{-1}k$ which can be written as $\sum_i\sum_j k_ik_j[C^{-1}]_{ij}$ and therefore takes $\mathcal{O}(n^2)$ multiplications.

To summarize: solving the regression first takes $\mathcal{O}(n^3)$ steps. Predicting the mean of new points takes $\mathcal{O}(n)$ steps, and predicting the variance of new points takes $\mathcal{O}(n^2)$ steps.

Problem 2.

i) At each time t_i, the expected value of $B(t_i) = 0$, since $B(t_i) - B(0) = B(t_i)$ is Normally distributed with zero mean. The covariance function $C(t_i,t_j)$ is then $\mathrm{E}(B(t_i)B(t_j))$. Assume $t_i > t_j$ and write

$$\begin{aligned} C(t_i,t_j) &= \mathrm{E}\left[B(t_i)B(t_j)\right] \\ &= \mathrm{E}\left[\{B(t_i)-B(t_j)\}B(t_j) + B^2(t_j)\right] \\ &= \mathrm{E}\left[\{B(t_i)-B(t_j)\}B(t_j)\right] + \mathrm{E}\left[B^2(t_j)\right] \\ &= \mathrm{E}\left[B^2(t_j)\right] \\ &= t_j. \end{aligned}$$

So the covariance is $C(t_i,t_j) = \min(t_i,t_j)$. This process actually exists and is continuous but nowhere differentiable, despite the innocent-looking covariance.

ii) The expected value is $\mathrm{E}(y) = \mathrm{E}(w^Tx + e) = 0$ given the noise assumption. The covariance function is then by definition

$$\begin{aligned} C(x_i,x_j) &= \mathrm{E}(y_iy_j) \\ &= \mathrm{E}((w^Tx_i + e_i)(w^Tx_j + e_j)) \\ &= \mathrm{E}(x_i^Tww^Tx_j) + \sigma^2\delta_{ij} \\ &= x_i^Tx_j + \sigma^2\delta_{ij}, \end{aligned}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

iii) The expected value is zero, since $\mathrm{E}(b) = \mathrm{E}(v_i) = 0$. The covariance function is then

$$C(x_i,x_j) = \mathrm{E}(f(x_i)f(x_j)) = \mathrm{E}\left[(b + \sum_k v_kh_{ik})(b + \sum_k v_kh_{jk})\right],$$

where $h_{ik} = \exp(-\frac{1}{2\sigma^2}\|x_i - u_k\|^2)$. Computing further gives

$$\begin{aligned} C(x_i,x_j) &= \sigma_b^2 + \sum_k \mathrm{E}(v_k^2h_{ik}h_{jk}) \\ &= \sigma_b^2 + \sum_k \sigma_v^2\mathrm{E}(h_{ik}h_{jk}) \\ &= \sigma_b^2 + K\sigma_v^2\mathrm{E}(h_{ik}h_{jk}). \end{aligned}$$

These steps follow from the independent zero-mean priors on the weights, and the i.i.d. prior for v_k's. It remains to compute the expectation. This is

$$\mathrm{E}(h_{ik}h_{jk}) = \int \exp(-\frac{1}{2\sigma^2}[(x_i-u)^T(x_i-u) + (x_j-u)^T(x_j-u)])p(u)du.$$

Now we assume that σ_u^2 is very large compared to σ^2 and omit the distribution $p(u) \approx$ constant.

The exponent can be written as a sum of an u-dependent and an x-dependent term:

$$-\frac{1}{2}[2u^Tu - 2(x_i + x_j)^Tu + x_i^Tx_i + x_j^Tx_j]\sigma^{-2} = -[(u-m)^T(u-m) + g(x_i, x_j)]\sigma^{-2}$$
$$= -[u^Tu - 2m^Tu + m^Tm + g(x_i, x_j)]\sigma^{-2}.$$

First find m: Comparing the terms in the left and right sides of the above equation, m must be $m = \frac{1}{2}[x_i + x_j]$. Then

$$\begin{aligned} g(x_i, x_j) &= \frac{1}{2}(x_i^Tx_i + x_j^Tx_j) - m^Tm \\ &= \frac{1}{4}(x_i^Tx_i + x_j^Tx_j) - \frac{1}{2}(x_i^Tx_j) \\ &= \frac{1}{4}(x_i - x_j)^T(x_i - x_j). \end{aligned}$$

This finishes the solution, since the integral over u simply integrates the term $\exp(-(u-m)^T(u-m))$ which results in a constant. What remains is $\exp(-\frac{1}{4}(x_i - x_j)^T(x_i - x_j))$.

The final covariance is approximately

$$C(x_i, x_j) \approx \sigma_b^2 + \sigma_v^2 K' \exp(-\frac{1}{4}(x_i - x_j)^T(x_i - x_j)),$$

where K' is a constant.

Problem 3.

i) To find the mode of $p(u|\tilde{x}D)$ we maximise $\log p(u|\tilde{x}D)$ over the latent variables u_i ($u = \{u_1, \ldots, u_n\}$). We use the Bayes Theorem to obtain

$$p(u|\tilde{x}, D) = p(u|\tilde{x}, x, y) \propto [\prod_i p(y_i|u_i, \tilde{x}, x)]p(u|\tilde{x}, x) = [\prod_i p(y_i|u_i)]p(u|\tilde{x}, x)$$

To find the conditional prior $p(u|\tilde{x}, x)$ we assume another set of latent variables w linearly related to u: $u_i = x_i^Tw \Rightarrow u = X^Tw$. Now we can reasonably assume all the dependence on the data x to be in the linear transformation matrix X^T and use a prior for w that is independent of x: $p(w|\tilde{x}, x) = p(w)$. As instructed, we take $p(w) = N(w|0, I)$. Since u is a linear combination of zero mean normally distributed variables w, its distribution also is a zero mean Gaussian distribution: $p(u|\tilde{x}, x) = N(u|0, C)$. The covariance matrix C is given by

$$C = E_{u|\tilde{x},x}[uu^T] = E_{w|\tilde{x},x}[X^Tw(X^Tw)^T] = X^T \underbrace{E_{w|\tilde{x},x}[ww^T]}_{=I} X = X^TX.$$

Inserting the prior into the function to be maximised, we have

$$\log p(u|\tilde{x}, D) = [\sum_i \log p(y_i|u_i)] - \frac{1}{2}u^TC^{-1}u + \text{constant}.$$

As hinted, we insert the assumption $w = Xa$ in $u = X^Tw$ and obtain $u = X^TXa = Ca$. This gives

$$u^TC^{-1}u = a^TCa.$$

But since $w = Xa$ we have that also $w^Tw = a^TX^TXa = a^TCa$. Therefore $u^TC^{-1}u = \|w\|^2$.

We can thus maximise

$$\log p(u|\tilde{x}, D) = [\sum_i \log p(y_i|u_i)] - \frac{1}{2}\|w\|^2 + \text{constant}$$

We may as well minimise

$$\|w\|^2 - 2\sum_i \log p(y_i|u_i)$$

Substitute the given distribution $p(y_i|u_i)$ to obtain

$$\|w\|^2 + 2\sum_i \log(1 + \exp(-2y_iw^Tx_i))$$

where we have used $u_i = w^Tx_i$.

ii) In the above cost function there are two parts. The $\|w\|^2$ part is independent of the training samples, whereas the sum evaluates the efficiency of the linear classiffier in classifying the training samples. Consider the effect of single training point i on the sum. From the expression for $p(y_i|u_i)$ we see that y_i is likely have the same sign as $u_i = w^Tx_i$. With large $|u_i|$ dependency is very sharp. $y_iw^Tx_i < 0$ is the indicator for sample i being probably misclassified.

In the case of almost certain misclassification $y_iw^Tx_i << 0$ the corresponding term in the sum is approximately $-2y_iw^Tx_i$, a large positive number. For a probable correct classification $y_iw^Tx_i >> 1$ the term in the sum is aproximately zero.

Similar considerations apply also to the soft-margin SVM cost function. The cost has also in this case a training sample independent term $\|w\|^2$. In the sum, samples classified succesfully with large enough margin $(y_i(w^Tx_i) \geq 1)$ are not penalised at all. Misclassifications $y_i(w^Tx_i) << 0$ result in a large positive cost.

Generally, the GP classifier is more or less close to the soft-margin SVM, depending on the distribution $p(y_i|u_i)$.