T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Exercises 12, 20.4.2007

Problem 1.

The solution for Gaussian Process regression was given in the lectures in the form of a Normal predictive distribution for a point $\tilde{y}$. Check that the predictive mean and variance are correct, using the following formulas for a joint normal distribution $p(u, v)$:

$$\mathrm{E}(u|v) = \mathrm{E}(u) + \mathrm{Cov}(v, u)(\mathrm{Var}(v))^{-1}(v - \mathrm{E}(v))$$
$$\mathrm{Var}(u|v) = \mathrm{Var}(u) - \mathrm{Cov}(v, u)(\mathrm{Var}(v))^{-1}\mathrm{Cov}(u, v).$$

What is the computational complexity of finding the solution, as a function of n, the number of training points? What is the extra cost of predicting another point? Assume that it is cheap to evaluate the covariance function.

Problem 2.

The covariance function $C(x_i, x_j)$ is the key element in a Gaussian Process prior. In this problem we see examples of processes of varying properties, which are not always apparent from the covariance function. Find the covariance function of the following processes:

i) Brownian motion $B(t)$, where $B(0) = 0$, increments $B(s) - B(t)$ with $s > t$ are Normally distributed random variables with distribution $N(0, s - t)$. Disjoint increments $B(s) - B(t)$ and $B(v) - B(w)$ are independent if $s > t \geq v > w$. Find the covariance function $C(t_i, t_j)$.

ii) Linear model $y_i = w^T x_i + e_i$ where the i.i.d. noise e_i is distributed as $N(0, \sigma^2)$. Inputs x_i are d-dimensional vectors and the prior for w is $p(w) = N(w|0, I)$. Find $C(x_i, x_j)$.

iii) A neural network $f(x) = b + \sum_k v_k \exp(-\frac{1}{2\sigma^2}\|x - u_k\|^2)$, where the network weights have Normal priors $p(u) = N(u|0, \sigma_u^2 I)$, $p(v) = N(v|0, \sigma_v^2 I)$, $p(b) = N(b|0, \sigma_b^2)$. The weights are independent of each other. For simplicity, assume that σ_u^2 is very large and $\sigma^2 = 1$.

Hint: use the assumption that the weights are independent to write the covariance as a function of $E\left[\exp(-\frac{1}{2\sigma^2}\|x_i - u\|^2)\exp(-\frac{1}{2\sigma^2}\|x_j - u\|^2)\right]$. The expectation is computed over $p(u)$, which you can assume to be constant since its variance is very large. Now you should be able integrate u out by rearranging the exponent term. Note that the integral can be computed only up to a proportionality constant.

Problem 3.

Using the notation of the lectures, the distribution $p(u|\tilde{x}, D)$ in GP classification is difficult to formulate since the training data D contains only the inputs x_i and corresponding class labels $y_i \in \{-1, +1\}$. Various approximations can be used, and some of them require that the mode is computed. Choose a linear classifier, so $u = X^T w$ with $X = [x_1, \ldots, x_n]$. Use the prior $p(w|\tilde{x}, x) = p(w) = N(w|0, I)$ for the classifier w. Assume that the linear classifier w is obtained as a linear combination of inputs, i.e. $w = \sum_i x_i a_i = Xa$. For the class label distribution, use $p(y_i|u_i) = (1 + \exp(-2y_i u_i))^{-1}$

i) Write the problem of finding the mode of $p(u|\tilde{x}, D)$ as a minimization problem wrt w.
Hints:
- find the prior $p(u|\tilde{x}, x) = p(u|x)$ induced by $p(w)$ by writing $u = X^T w$ and computing $\mathrm{E}[uu^T]$.
- to replace u by w you need to use the representation $w = Xa$.

ii) Compare the result with Soft Margin SVM, where the optimization is to minimize $\|w\|^2 + K \sum_i (1 - y_i(w^T x_i))_+$ where $z_+ = \max(z, 0)$.
Hint:
- to solve this problem, examine what happens approximately in the cost functions when $w^T x_i$ has a very large absolute value and has either the same or different sign as the correct class label y_i.