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

**Exercises 9, 23.3.2007**

## Problem 1.

In variational learning the following expression is maximised:

$$C = \int q(\theta) \log \left( \frac{p(y|\theta)p(\theta)}{q(\theta)} \right) d\theta$$

i) Justify variational learning by showing that it minimizes the Kullback-Leibler divergence $D(q\|p)$ between $q(\theta)$ and $p(\theta|y)$.
ii) The evidence $p(y)$, or the prior predictive distribution, is the probability of the data according to the model $p(y, \theta)$. The evidence can not be computed if the model can not be integrated. Show that maximising $C$ maximises a lower bound for $p(y)$. Hint: Jensen's inequality for a random variable $X$ and a convex function $f$: $E[f(X)] \geq f(E[X])$.

## Problem 2.

Consider what happens in variational learning with a Normal approximate posterior $q(\theta)$, when the true posterior $p(\theta|y)$ is a mixture of two Normal distributions $N(\theta|\mu_i, \sigma_i^2)$, $i = 1, 2$ with prior probabilities $a_1$ and $a_2 = 1 - a_1$. Assume that the mixture distributions $N(\theta|\mu_i, \sigma_i^2)$ are separated well enough to warrant fitting $q$ separately to the mixture components.

## Problem 3.

The Poisson distribution with intensity $\lambda$ is

$$p(k|\lambda) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

i) Compute the Laplace approximation for $p(\lambda|k)$ using the prior $p(\lambda) \propto \lambda^{-1}$.

ii) Also compute the Laplace approximation for $p(\log \lambda|k)$. Note that the prior is now $p(\log \lambda) \propto 1$ (This follows from the formula for the density of a transformation: Let $l = g(\lambda) = \log \lambda$. Now $p_l(l) = (\frac{\partial g(\lambda)}{\partial \lambda})^{-1} \cdot p_\lambda(\lambda)$.)

**Problem 4.**

Approximate the posterior $p(\theta|y)$ by another distribution $q(\theta)$. Use $q(\theta) = N(\theta|\theta_0, \sigma^2)$ where *the variance* $\sigma^2$ *is known*. What is $\theta_0$ when

i) you minimize the KL divergence $D(p(\theta|y)\|q(\theta))$?

ii) you minimize the KL divergence $D(q(\theta)\|p(\theta|y))$?

Note that in part ii), the solution can be found only approximately.