GI01/4C55 – Homework #5 (Due 12am, December 16, 2005)

Aim: To get familiarity with neural networks and Gaussian processes. Presentation, clarity, and synthesis of exposition will be taken into account in the assessment of these exercises.

This document is available at `http://www.cs.ucl.ac.uk/staff/M.Pontil/courses/SL-homework5.pdf`

1. [**40 pts**] `basic` (*weight decay*)

   Consider a one-hidden-layer neural network with a single output,

   $$f(\mathbf{x}) = g\left(\sum_{n=1}^{N} u_n h(\mathbf{w}_n^\top \mathbf{x}_i)\right) \tag{1}$$

   where the activation function h and the output function g are prescribed. Derive the gradient descent equations for this neural network trained by weight decay, that is the weights are computed by minimizing the following penalized empirical error

   $$\sum_{i=1}^{m} \Big(y_i - f(\mathbf{x}_i)\Big)^2 + \lambda \left(\sum_{n=1}^{N} \|\mathbf{w}_n\|^2 + \|\mathbf{u}\|^2\right).$$

   Comment on the back-propagation algorithm in this case.

2. [**30 pts**] `basic` (*neural networks*)

   Consider again the neural network in equation (1) using the square-error loss, identity output function ($g(t) = t$) and sigmoid activation function h, that is

   $$h(z) = \frac{1}{1 + e^{-z}}.$$

   Suppose that the weights $\mathbf{w}_n$ from the input to the hidden layer are nearly zero. Show that the resulting model is nearly linear in the inputs.

3. [**30 pts**] `medium/advanced` (*Gaussian processes*)

   Let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^N$ be a prescribed feature map and $K(\mathbf{x}, \mathbf{t}) = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{t})$ the associated kernel function. Consider on the space of functions $\mathcal{H} := \{f(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) : \mathbf{w} \in \mathbb{R}^N\}$ the Gaussian prior probability distribution

   $$P(\mathbf{w}) \propto e^{-\mathbf{w}^\top \mathbf{w}}.$$

   Let $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \ldots, m$ be distinct training points, $y_i \in \mathbb{R}, i = 1, \ldots, m$ target outputs and assume that the kernel matrix $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, \ldots, m)$ is invertible.

   Show that the conditional distribution of the random variable $f(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$ given the information that $f(\mathbf{x}_i) = y_i$, $i = 1, \ldots, m$, is Gaussian with mean

   $$\mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{y}$$

   and variance

   $$K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})$$

   where we used the notation $\mathbf{y} = (y_1, \ldots, y_m)^\top$ and $\mathbf{k}(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_m))^\top$.