GI01/M055 – Homework #1 (Due 12am, October 23, 2009)

Aim: To get familiarity with the theory of linear regression and least squares. 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/J.Shawe-Taylor/courses/SL-homework1.pdf`

1. **[10 pts]** `basic` (*least squares*)

   (a) Consider the constant function model $f(\mathbf{x}; b) = b$, where $b \in \mathbb{R}$. Derive the least squares solution for this model, that is, compute a minimizer of the empirical error

   $$\mathcal{E}(b) := \sum_{i=1}^{m} \big(y_i - f(\mathbf{x}_i; b)\big)^2.$$

   Does the solution depend on the dimension d of $\mathbf{x}$? Explain your observation.

   (b) Now, consider a linear model in one dimension: $f(x) = ax + b$, where $a, b \in \mathbb{R}$. Derive the least square solution for this model.

2. **[20 pts]** `medium` (*non-linear least squares*)

   Suppose that the regression function $f^*(\mathbf{x})$ is quadratic in $\mathbf{x}$ and goes through the origin, that is, $f^*(\mathbf{0}) = 0$ (here, $\mathbf{0}$ is the d–dimensional vector all of whose components are equal to zero). Which hypothesis space of models would you choose in this case? Explain why this would be a good choice. Show how the learning problem can be reduced to solving a multi-dimensional linear regression problem (*Hint:* $\mathbf{x}^\top \mathbf{A} \mathbf{x} = \mathrm{vec}(\mathbf{A})^\top \mathrm{vec}(\mathbf{x}\mathbf{x}^\top)$, where $\mathrm{vec}(\mathbf{A})$ converts the matrix $\mathbf{A}$ into a vector by concatenating its columns).

3. **[20 pts]** `medium` (*decomposition formula for the expected error*)

   Let $d = 1$ and define the expected error of a function $f : \mathbb{R} \to \mathbb{R}$ as

   $$\mathcal{E}(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \big(y - f(x)\big)^2 dP(x, y).$$

   Derive the following formula:

   $$\mathcal{E}(f) = \mathcal{E}(f^*) + \int_{-\infty}^{\infty} \big(f^*(x) - f(x)\big)^2 dP(x),$$

   where f^* is the optimal regression function. Explain the meaning of this formula and argue that it also holds true in the general case that $\mathbf{x} \in \mathbb{R}^d$ and $dP(\mathbf{x}, y)$ is an arbitrary probability measure on $\mathbb{R}^d \times \mathbb{R}$.

4. **[50 pts]** `old exam question` (*nearest neighbour*) Given a training set

   $$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$$

   (a) describe the k-nearest neighbour algorithm for classifying a test point $\mathbf{x}$ given a distance function $d(\cdot, \cdot)$ defined between inputs.

   (b) What assumption about the conditional probability distribution $P(y|\mathbf{x})$ is used to motivate the nearest neighbour algorithm?

   (c) Show how the assumption is used to justify the accuracy of the algorithm and discuss convergence to the Bayes optimal classifier.

   (d) For a fixed training set size m, how does varying k affect the complexity of the class of functions?

   (e) For a noisy training set describe with a graph how we might expect training and test error to vary with k, indicating the underfitting and overfitting regimes.

Remarks: In all questions, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \mathbb{R}$ denotes a training set and $\mathbb{R} = (-\infty, \infty)$ is the set of real numbers. It is always assumed that S is generated *i.i.d.* from a probability measure $P(\mathbf{x}, y)$. The function f^* denotes the regression function, namely, $f^*(\mathbf{x}) = \int_{-\infty}^{\infty} y dP(y|\mathbf{x}) \equiv \mathbf{E}_{y|\mathbf{x}}[y]$. If you find it easier you may also assume that the probability measure $dP(\mathbf{x}, y)$ has a density, in which case you can write $dP(\mathbf{x}, y) = P(\mathbf{x}, y) d\mathbf{x} dy$. For example, in the additive Gaussian noise model $P(\mathbf{x}, y) \propto \exp(-\beta(y - f^*(\mathbf{x}))^2) P(\mathbf{x})$.