CS 101.2 - Active Learning
Problem Set 2

Handed out: 3 Feb 2009
Due: 17 Feb 2009

---

1 Gaussian Processes and Bayesian Linear Regression

The regression problem involves estimating the functional dependence between an input variable $\mathbf{x}$ in $\mathbb{R}^d$ and an output variable y in $\mathbb{R}$. We assume the relationship

$$y(\mathbf{x}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) + \epsilon = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + \epsilon, \tag{1}$$

where $\mathbf{w}^T \mathbf{\Phi}(\mathbf{x})$ is a linear combination of M predefined nonlinear basis functions $\phi_j(\mathbf{x})$ with input in $\mathbb{R}^d$ and output in $\mathbb{R}$. The observations are additively corrupted by i.i.d. noise with normal distribution

$$\epsilon \sim N(0, \sigma_n^2) \tag{2}$$

which has zero mean and variance σ_n^2.

Our goal is to estimate the weights w_j given a training set consisting of pairs $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$. We use a multivariate normal distribution as prior on the weights

$$\mathbf{w} \sim N(0, \mathbf{\Sigma}_w), \tag{3}$$

with zero mean and M-by-M sized covariance matrix $\mathbf{\Sigma}_w$.

The goal of this problem is to show that the Bayesian linear regression defined above is an example of a Gaussian process. Recall that a Gaussian Process is a probability measure over $y(\mathbf{x})$ defined by a mean function $\mu(\mathbf{x}) = E[y(\mathbf{x})]$ and a covariance kernel $K(\mathbf{x}, \mathbf{x}') = E[(y(\mathbf{x}) - \mu(\mathbf{x}))(y(\mathbf{x}') - \mu(\mathbf{x}'))]$. We can write this as

$$y(\mathbf{x}) \sim \mathrm{GP}(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')). \tag{4}$$

1. Show that the Bayesian linear regression functions defined above have mean function

   $$\mu(\mathbf{x}) = 0 \tag{5}$$

   and covariance function

   $$K(\mathbf{x}, \mathbf{x}') = \mathbf{\Phi}(\mathbf{x})^T \mathbf{\Sigma}_w \mathbf{\Phi}(\mathbf{x}') + \sigma_n^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{6}$$

   where $\delta(\mathbf{x}, \mathbf{x}') = 1$ if $\mathbf{x} = \mathbf{x}'$ and zero otherwise.

2. Prove that the covariance function (eq. 6) is a valid kernel function. That is, prove that it is symmetric and positive semi-definite.

3. Derive an expression for

$$P(y'|\mathbf{x}', (\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)) \tag{7}$$

which is the predictive distribution of the output variable y' associated with test point $\mathbf{x}'$ given that we have observed a training data set $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$.

4. Derive an expression for the 95-th percentile of the predictive distribution of y'.

5. Use $[1\ x\ x^2\ x^3]^T$ as your basis functions, $\mathbf{\Sigma}_w$ equal to the identity matrix, and $\sigma_n^2 = 0.1$. Implement Bayesian linear regression and run it on the following data set: $(x_1 = 1, y_1 = 0.5), (x_2 = 2, y_2 = 1.6), (x_3 = 3, y_3 = 1.1), (x_4 = 4, y_4 = 3), (x_5 = 5, y_5 = 4.2)$. What are the predictive distributions associated with the following test points: $x' = \{1.5, 2.5, 3.5, 4.5, 5.5\}$? What is the value associated with the 95-th percentile for each test point?

2 VC Dimension

1. **Unions of intervals.** Consider data x that lies in the interval $[0, 1]$. Suppose the hypothesis space consists of indicator functions of unions of two intervals. More specifically, let the hypothesis space be parameterized by a, b, c, d satisfying $a < b$ and $c < d$ so that data points that fall in either interval $[a, b]$ or $[c, d]$ are classified as positive, and data that falls outside both intervals is classified as negative. What is the VC dimension of this hypothesis class?

2. **Decision Stumps in $\mathbb{R}^D$.** Decision stumps are a particularly simple family of binary classifiers for data $\mathbf{x}$ that lies in $\mathbb{R}^D$. Their classification rule has parameters q, i, α and takes the form $f(\mathbf{x}; i, q, \alpha) = q * \text{sign}(\mathbf{x}_i - \alpha)$. Decision stumps classify example $\mathbf{x}$ based only on the value of its i-th coordinate. α is a threshold value in $\mathbb{R}$ and q is either $+1$ or -1.

   (a) Consider n non-overlapping data points lying in $\mathbb{R}^D$. What is the maximum number of ways they can be classified using the decision stump family? That is, how many different binary labelings of the n points are there in the decision stump hypothesis space? Your result should be a function of n and D.

   (b) Show that the above result implies the following about the VC dimension of decision stumps:

$$VC_{ds} < 2(\log_2 D + 1) \tag{8}$$

3 Active Learning

The purpose of this question is to design an active learning strategy for the hypothesis space from Problem 2.1. To simplify things, assume that $0 < a < b < c < d < 1$. Also assume that the parameters a, b, c, d are all separated by at least $\eta > 0$ and both intervals are at least length η such that, i.e., $a > \eta$, $b - a > \eta$, $c - b > \eta$, $d - c > \eta$, and $d < 1 - \eta$. Hereby, η is a constant, known to the algorithm. Suppose that the distribution over the inputs $P(x)$ is the uniform distribution over $[0, 1]$.

(a) Develop an active learning scheme that only requires $O(\log 1/\varepsilon \log 1/\delta)$ labels to find a hypothesis with error at most ε with probability $1 - \delta$. Bound the number of labels that your algorithm requires as a function of ϵ, δ and η.

(b) Generalize this scheme to the hypothesis class containing of hypotheses that are indicator functions of unions of k intervals, i.e., for each hypothesis h there exists $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, $b_i - a_i > \eta$, $a_{i+1} - b_i > \eta$ and $a_1 > \eta$, $b_k < 1 - \eta$, such that h classifies inputs x positive if x is contained in one of the intervals $[a_i, b_i]$, negative otherwise.