1 Gaussian Processes and Bayesian Linear Regression

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

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

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

\[
\epsilon \sim \mathcal{N}(0, \sigma_n^2)
\]

which has zero mean and variance \( \sigma_n^2 \).

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

\[
w \sim \mathcal{N}(0, \Sigma_w),
\]

with zero mean and \( M \)-by-\( M \) sized covariance matrix \( \Sigma_w \).

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 \( y(x) \) defined by a mean function \( \mu(x) = E[y(x)] \) and a covariance kernel \( K(x, x') = E[(y(x) - \mu(x))(y(x') - \mu(x'))] \). We can write this as

\[
y(x) \sim \text{GP}(\mu(x), K(x, x')).
\]

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

\[
\mu(x) = 0
\]

and covariance function

\[
K(x, x') = \Phi(x)^T \Sigma_w \Phi(x') + \sigma_n^2 \delta(x, x'),
\]

where \( \delta(x, x') = 1 \) if \( x = x' \) 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

$$P(y'|x', (x_1, y_1), \ldots, (x_n, y_n))$$

(7)

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

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, $\Sigma_w$ 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 $x$ that lies in $\mathbb{R}^D$. Their classification rule has parameters $q, i, \alpha$ and takes the form $f(x; i, q, \alpha) = q \cdot \text{sign}(x_i - \alpha)$. Decision stumps classify example $x$ based only on the value of its $i$-th coordinate. $\alpha$ 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:

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

(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 $\eta$ such that, i.e., $a > \eta$, $b - a > \eta$, $c - b > \eta$, $d - c > \eta$, and $d < 1 - \eta$. Hereby, $\eta$ 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 $O(\log 1/\epsilon \log 1/\delta)$ labels to find a hypothesis with error at most $\epsilon$ with probability $1 - \delta$. Bound the number of labels that your algorithm requires as a function of $\epsilon$, $\delta$ and $\eta$.

(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.