Foundations of Machine Learning
Department of Computer Science, NYU
Homework assignment 2 – Solution

(1) [15 points] Shattering coefficients of intervals in \( \mathbb{R} \).

Let \( \{x_1, \ldots, x_m\} \) be a set of \( m \) distinct ordered real numbers and let \( I \) be an interval. If \( I \cap \{x_1, \ldots, x_m\} \) contains \( k \) numbers, their indices are of the form \( j, j + 1, \ldots, j + k - 1 \). How many different set of indices are possible? The answer is \( m - k + 1 \) since \( j \) can take values in \( \{1, \ldots, m - k + 1\} \). Thus, the total number of dichotomies for a set of size \( m \) is:

\[
1 + \sum_{k=1}^{m} (m - k + 1) = 1 + \frac{m(m + 1)}{2} = \binom{m}{0} + \binom{m}{1} + \binom{m}{2},
\]

which matches the general bound on shattering coefficients.

(2) [15 points] Bound on shattering coefficients.

Let \( X \) be an arbitrary set. Consider the set of all subsets of \( X \) of size less than or equal to \( d \). The indicator functions of these sets form a concept class whose shattering coefficient is equal to the general upper bound.

(3) [20 points] VC dimension of a vector space of real functions.

Show that no set of size \( m = r + 1 \) can be shattered by \( H \). Let \( x_1, \ldots, x_m \) be \( m \) arbitrary points. Define the linear mapping \( l : F \rightarrow \mathbb{R}^m \) defined by:

\[
l(f) = (f(x_1), \ldots, f(x_m))
\]

Since the dimension of \( \text{dim}(F) = m - 1 \), the rank of \( l \) is at most \( m - 1 \) and there exists \( \alpha \in \mathbb{R}^m \) orthogonal to \( l(F) \):

\[
\forall f \in F, \sum_{i=1}^{m} \alpha_i f(x_i) = 0
\]

We can assume that at least one \( \alpha_i \) is negative. Then,

\[
\forall f \in F, \sum_{i: \alpha_i \geq 0} \alpha_i f(x_i) = -\sum_{i: \alpha_i < 0} \alpha_i f(x_i)
\]
Now, assume that there exists a set \( \{ x : f(x) \geq 0 \} \) selecting exactly the \( x_i \)'s on the left-hand side. Then all the terms on the left-hand side are non-negative, while those on the right-hand side are negative, which cannot be. Thus, \( \{ x_1, \ldots, x_m \} \) cannot be shattered.

(4) [20 points] Soft margin hyperplanes

The corresponding dual problem is:

\[
\max_{\alpha, \beta} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \left( \alpha_i + \beta_i \right)^{k/(k-1)} \\
\frac{(kC)^{1/(k-1)}}{(k-1)} \left( \frac{1}{k} - 1 \right)
\]

subject to:

\[
\sum_{i=1}^{m} \alpha_i y_i = 0 \quad \alpha \geq 0 \quad \beta \geq 0.
\]

(5) [30 points] SVM classification

- [(i)] The data was already generated.
- [(ii)] Downloading and installing the library should not have been a problem.
- [(iii)] [7.5 points] Plotting the test error as a function of the parameters should not have been a problem.
- [(iv)] [7.5 points] Plotting the training and test errors as a function of the size of the training data is also straightforward.
- [(v)] [15 points] The density functions of the two classes have the same covariance matrix, thus the best classifier \( h^* \) is the one based on the separating hyperplane halfway between the centers of the Gaussians passing through \( \frac{\mu_2 - \mu_1}{2} = 0 \). The corresponding error is the area where the two density functions overlap, which can be proved to be:

\[
\Pr[error(h^*]) = \int_{\rho/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2} dx
\]

with \( \rho = (\mu_2 - \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1) \) and \( \Sigma = 50I \). An approximate value of the integral is: \( \Pr[error(h^*]) \approx 0.16 \). This seems to be close to the empirical test error, but it must be a lower bound.