A. VC Dimension

1. What is the VC dimension of the family of finite unions of closed intervals over the real line?

2. What is the VC dimension of the intervals of the real line \([x, x + 1], x \in \mathbb{R}\)?

3. What is the VC dimension of triangles in the plane? For this question, try to give a direct proof without using the result given in class for polygons. (hint: consider two cases for eight points: (1) one of the points is in the convex hull of all the others; (2) none of the points is in the convex hull of all the others. In the latter case, you can show that an alternating labeling is not possible.)

4. What is the VC dimension of the family of indicator functions 
\[\{x \mapsto 1_{\sin(x + \phi) > 0}: \phi \in \mathbb{R}\}\]?

5. Let \(C\) be a concept class of finite VC dimension \(d \geq 1\). For \(k \geq 1\), let 
\[C_k = \bigcup_{i=1}^{k} c_i: c_i \in C\].

   (a) Let \(S\) be a sample of size \(m\) and let \(\Pi_K(S)\) denote the set of all labelings of \(S\) using a concept class \(K\). Show that \(|\Pi_{C_k}(S)| \leq |\Pi_{C}(S)|^k\).

   (b) Prove that \(\text{VCdim}(C_k) = O(dk \log k)\) (hint: you can use the fact that \(\log(3k) < \frac{9k}{2e}\) for \(k \geq 2\)).

A. Support Vector Machines

1. Download and install the \texttt{libsvm} software library from:

   \url{http://www.csie.ntu.edu.tw/~cjlin/libsvm/}

2. Download the \texttt{abalone} data set:
Use the libsvm scaling tool to scale the features of all the data. Use the first 3133 examples for training, the last 1044 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the binary classification that consists of distinguishing classes 1 through 9 from the rest. Use SVMs combined with polynomial kernels to tackle this binary classification problem.

To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, \( d = 1, 2, 3, 4 \), plot the average cross-validation error plus or minus one standard deviation as a function of \( C \) (let other parameters of polynomial kernels in libsvm be equal to their default values), varying \( C \) in powers of 2, starting from a small value \( C = 2^{-k} \) to \( C = 2^k \), for some value of \( k \). \( k \) should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with libsvm as the value of \( C \) increases.

4. Let \((C^*, d^*)\) be the best pair found previously. Fix \( C \) to be \( C^* \). Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of \( d \). Plot the average number of support vectors obtained as a function of \( d \). How many of the support vectors lie on the margin hyper-planes?

5. In class, we gave two types of argument in favor of the SVMs algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose that instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the norm \( p \) of the vector \( \alpha \) that defines the weight vector \( w \), for some \( p \geq 1 \). For simplicity, fix \( p = 2 \). This gives the following optimization problem for a kernel function \( K \):

\[
\min_{\alpha,b} \frac{1}{2} \sum_{i=1}^{m} \alpha_i^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to } y_i \left( \sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) + b \right) \geq 1 - \xi_i, i \in [1, m] \\
\xi_i, \alpha_i \geq 0, i \in [1, m].
\]
(a) Show that modulo the non-negativity constraint on $\alpha$, the problem coincides with an instance of the primal optimization problem of SVMs (indicate exactly how to view it as such).

(b) Is the positive-definiteness of the kernel function $K$ needed to ensure that this is a convex optimization problem? Justify your response.

(c) Derive the dual optimization of problem of (1).

(d) Suppose we omit the non-negativity constraint on $\alpha$. Use libsvm to solve the problem. Plot the ten-fold cross-validation training and test errors for the hypotheses obtained based on the solution $\alpha$ as a function of $d$, for the best value of $C$ measured on the validation set.