Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
Due: March 19, 2010

A. VC Dimension

1. Let H and H' be two families of functions mapping from X to $\{0, 1\}$ with finite VC dimensions. Show that

$$\mathrm{VCdim}(H \cup H') \leq \mathrm{VCdim}(H) + \mathrm{VCdim}(H') + 1. \qquad (1)$$

   Use that to determine the VC dimension of the hypothesis set formed by the union of axis-aligned rectangles and triangles in dimension 2.

2. For two sets A and B, let $A\Delta B$ denote the symmetric difference of A and B, that is $A\Delta B = (A \cup B) - (A \cap B)$. Let H be a non-empty family of subsets of X with finite VC dimension. Let A be an element of H and define $H\Delta A = \{X\Delta A \colon X \in H\}$. Show that

$$\mathrm{VCdim}(H\Delta A) = \mathrm{VCdim}(H). \qquad (2)$$

B. SVMs

1. Download and install the `libsvm` software library from:
   `http://www.csie.ntu.edu.tw/˜cjlin/libsvm/`.

2. Download the `ISOLET` data set from
   `http://archive.ics.uci.edu/ml/datasets/ISOLET`. There are two files: `isolet1+2+3+4`, and `isolet5`. `isolet1+2+3+4` should be used for training and validation, `isolet5` for testing only.

   The dataset corresponds to a number of people pronouncing each of the 26 letters in the alphabet. Each person pronounces the whole alphabet twice. We will consider the binary classification that consists of distinguishing the first 13 letters from the last 13 letters. Thus, assign label '-1' to the first 13, '+1' to the rest.

3. Normalize all input vectors (`isolet1+2+3+4` and `isolet5`): compute the scaling and offset on the `isolet1+2+3+4` so that each feature has zero mean and standard deviation 1, and apply the same scaling on the `isolet5` data.

4. Split the data `isolet1+2+3+4`, containing the alphabet spoken 240 times, into 10 folds, ensuring that the same speaker is only in one of the folds by splitting after each 24 pronunciations (modulo a few that are missing; check that all your files start with the label '1').

5. Let $x_1, \ldots, x_m$ denote the sample formed by the ten folds. For each feature f, let $f_1, \ldots, f_m$ denote its values for $x_1, \ldots, x_m$, and let $y_1, \ldots, y_m$ denote the labels. Compute the empirical correlation of each non-constant feature f with the labels

$$\widehat{\rho}(f, y) = \frac{\widehat{\sigma}_{fy}}{\sqrt{\widehat{\sigma}_{ff}\widehat{\sigma}_{yy}}},$$

   where $\widehat{\sigma}_{fy} = \frac{1}{m}\sum_{i=1}^{m}(f_i - \overline{f})(y_i - \overline{y})$, with $\overline{f}$ the average value of f, and $\overline{y}$ the average value of y. $\widehat{\sigma}_{ff}$ and $\widehat{\sigma}_{yy}$ are defined in a similar way. Sort all features in decreasing order of the absolute value of the correlation and save the correlation values.

6. We first consider a *Naive* algorithm. Given a kernel K, this algorithm assigns a label to a new point simply based on its similarity with respect to the set of positively labeled versus its similarity with the negatively labeled points of the training set. Thus, for any point x, if we denote by $\mathbf{y}$ the vector of the labels in the training set and use the notation $\mathbf{K}_x = \begin{bmatrix} K(x,x_1) \\ \vdots \\ K(x,x_m) \end{bmatrix}$, the label it assigns to x is

$$h(x) = \operatorname{sgn}(\mathbf{K}_x \cdot \mathbf{y}).$$

   Determine and report the performance of the Naive algorithm on the test set when using a polynomial kernel of degree d with $d = 1, 2, 3, 4$, when using a percentage p of the most correlated features, with $p = 100\%, 80\%, 40\%, 20\%$.

7. Determine and report the performance of SVMs for the same set of kernels and the same sets of features. To determine the trade-off parameter C, use 10-fold cross validation with the ten folds previously defined (let the other parameters of polynomial kernels in `libsvm`, γ and c, be equal to their default values 1). Give a plot comparing the performance of SVMs with that of the Naive algorithm for each value of p.

8. Now, first multiply each feature f by its empirical correlation $\widehat{\rho}(f, y)$ and retrain SVMs with the full set of features. Report the test results obtained for the four values of d.

9. For each d and p, fix C to its best value obtained in question 7. For this value of C, for each d and p, train ten models, each time by excluding one

of the ten folds. Compute the average number of support vectors and the average test error for each d and p. Plot the average error as a function of the average fraction of support vectors. Discuss your results and compare with the leave-one-out theorem presented in class.