Machine Learning: Foundations **Fall Semester, 2010/11**

Homework 3: Dec 26, 2010

Lecturer: Yishay Mansour

Final Project.

Prepare a one page proposal for the final project. The proposal should give enough details to understand what you are planning to do in the project.

Submit it on a **different** page with the names and e-mails of the students (single or pairs), and send it also by e-mail (with a subject `ML Final Project`) to the lecturer.

Homework number 3.

Theory question I:

1. Compute the VC dimension of the class of convex polygons in the plane with d edges.(Show that it is $\Theta(d)$. Try to get the tightest bound.)

2. Compute the Rademacher Complexity of the class of linear functions $h_w(x) = \sum_{i=1}^{d} x_i w_i$, where $\|x\| = R$ and $\|w\| = \Lambda$.

Theory question II:

1. For each d, show that if two concept classes C_1 and C_2 have both VC dimension d, their union has VC dimension at most $2d+1$. (If you are unable to prove $2d+1$, prove the best upper bound you can.)

2. For each d, show two concept classes C_1 and C_2 whose VC dimension is d, and whose union has VC dimension $2d+1$. (If you are unable to prove $2d+1$, prove the best lower bound you can.)

3. *Bonus:* Show that if the concept classes C_i , $1 \leq i \leq d^d$ have VC dimension at most d, then their union $C = \cup_{i \in [1,d^d]} C_i$ has VC dimension at most $O(d \log d)$.

Theory question III: Consider training a soft-margin SVM with the cost C set to some positive constant. Suppose the training data is linearly separable. A student claimed:

> Since increasing the slack variables ψ_i can only increase the objective of the primal problem (which we are trying to minimize), at the optimal solution to the primal problem, all the training examples will have functional margin at least 1 and all the ψ_i will be equal to zero.

Is the claim True or False? Explain!
(Namely, Given a linearly separable dataset, is it necessarily better to use a a hard margin SVM over a soft-margin SVM?)

Programming assignment:

1. First download and install `libsvm` from `www.csie.ntu.edu.tw/∼cjlin/libsvm/`. (There are interfaces for Matlab, R and simple line commands, any one of those would be sufficient.)

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`. The file `isolet1+2+3+4` should be used for training, and the file `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, and $+1$ to the rest.

3. Train using a polynomial kernel with costs $0.01, 0.1, 0.5, 1.0, 2.0, 10, 100$ (use the default `degree=3` as well as the other defaults parameters for `gamma` and `coef0`). Report the accuracy (on the test set `isolet5`) and the number of support vectors. Discuss your observations.

4. Train using a polynomial kernel with degrees $1, 2, 3, 4, 5$ (use the default `cost=1` as well as the other defaults parameters for `gamma` and `coef0`). Report the accuracy (on the test set `isolet5`) and the number of support vectors. Discuss your observations.

5. A naive kernel classifier, given a training set $D = \{(x_i, y_i) : 1 \le i \le m\}$, generates the hypothesis $h_D(x) = sign(\sum_{(x_i,y_i)\in D} y_i K_d(x_i, x))$, where K_d is a polynomial kernel of degree d, i.e., $K_d(x, z) = (\frac{1}{F} < x, z >)^d$ where F is the number of attributes (for our case $F = 617$). Build h_D using the training set (`isolet1+2+3+4`), and report the accuracy of h_D on the test set (`isolet5`) when using polynomial kernels of degrees $1, 2, 3, 4, 5$. Would the results improve significantly if you would add a threshold b, i.e., $h'_D(x) = sign(-b + \sum_{(x_i,y_i)\in D} y_i K_d(x_i, x))$. Discuss your observations.

The homework is due in two weeks