Assignment 1: Decision Trees, Agnostic PAC Learning and VC dimension

CS489/698 – Winter 2010

Out: January 21, 2010
Due: February 9, 2010

Be sure to include your name and student number with your assignment.

1. Let X be some domain set and H a collection of functions from X to $\{0,1\}$. For a probability distribution P over $X \times \{0,1\}$ let $Err^P(H)$ denote $inf\{Err^P(h) : h \in H\}$ and, for sample $S \subset X \times \{0,1\}$, let $Err^S(H)$ denote $min\{Err^S(h) : h \in H\}$. We say that H is *ERM-learnable* (where ERM stands for Empirical Risk Minimization), if for every positive ϵ and δ there exist a number $m(\epsilon, \delta)$ so that, for every probability distribution P over $X \times \{0,1\}$, if S is an i.i.d. sample from P of size $m \geq m(\epsilon, \delta)$, then

$$P_{S \sim P^m}[sup\{|Err^P(h) - Err^P(H)| : h \in H \text{ and } Err^S(h) = Err^S(H)\} > \epsilon] < \delta$$

(That is, with high probability over the choice of S, an h that minimizes the S-empirical error has true error that is close to that of the best hypothesis in H).

   (a) **[10 pts]** Let X be the 12-dimensional Euclidean space $\Re^{12}$ and define a function h by

$$h(\bar{x} = (x_1, \ldots, x_{12})) = 1 \text{ if } \sum_{i=1}^{12} \frac{x_i}{2^i} > 0.5$$

   and $h(\bar{x}) = 0$ otherwise. Find a sample size, m_0, such that for any probability distribution P over $\Re^{12} \times \{0,1\}$ for which $E^P(h) = 0.3$, if S is an m-size sample drawn i.i.d by P, then with probability greater than 0.9, $0.2 \leq E^S(h) \leq 0.4$. The smaller the sample size m_0 you come up with, the higher your mark will be. However, you should prove your claim.

   (b) **[20 pts]** Let $X = [0,1]$, let

$$H_{mirrow} = \{h : X \rightarrow \{0,1\} : \forall x h(x) = 1 - h(1-x)\}$$

   Prove that H_{mirror} is *not* ERM-learnable. Hint: for P, consider the uniform distribution over $[0, 0.5] \times \{1\}$.

   (c) **[10 pts]** Given a class of functions, H, as above, define a new class $H^C = \{h^c : h \in H\}$ where, for $h : X \rightarrow \{0,1\}$, the function h^c is defined by setting, for every x, $h^c(x) = 1 - h(x)$. Prove that a class H is ERM learnable if and only if the class H^C is ERM learnable.

2. **[10 pts]** What is the VC-dimension of the class of all disks in the Euclidenan plane? (By a disk we mean a set of the form

$$d_{x_0,y_0,r} = \{(x,y) : (x - x_0)^2 + (y - y_0)^2 \leq r\}$$

for some $(x_0, y_0) \in \Re^2$ and $r \in \Re$). Prove your claim.

3. **[50 pts]** Decision Tree Learning

   (a) **[25 pts]** Implement the decision tree learning algorithm described in the lecture notes. More specifically, implement the ID3 algorithm and select attributes by minimizing the expected training error. A dataset will be posted on the course website. Report the tree found by your algorithm (either print it or draw it by hand) with the training error and testing error. Here training error is the percentage of misclassified instances in the training set and similarly for the testing error with respect to the testing set.

   (b) **[25 pts]** Implement the tree pruning procedure described in the lecture notes. Use Equation 13 from the lecture note as a bound on the true error to find the best pruned tree. Using the same dataset as in the previous question, report the tree found after pruning with the training error and testing error.