10-601 Machine Learning: Homework Assignment 5

Professor Tom Mitchell
Carnegie Mellon University
March 30, 2009

- The assignment is due at 1:30pm (beginning of class) on **Monday, April 6, 2009**.
- Submit writeups to the two problems *separately* with your name on each problem. Please do not staple the writeups together. Write your name at the top right-hand corner of each page submitted.
- Each student must hand in their own answers to the following questions. See the course webpage for the collaboration policies.
- Each question has the name of the TA who is the primary contact point for that question. Feel free to ask the other instructors about any question, but that TA is the authority on that question.

1 Computational Learning Theory [Andy, 25 points]

1.1 VC dimension

Consider the space of instances X corresponding to all points in the x, y plane. Give the VC dimension of the following hypothesis spaces:

1. H_r = the set of all rectangles in the x, y plane. That is, $H = \{((a < x < b) \wedge (c < y < d))|a, b, c, d \in \mathbb{R}\}$.
2. H_c = circles in the x, y plane. Points inside the circle are classified as positive examples.
3. H_t = triangles in the x, y plane. Points inside the triangle are classified as positive examples.

1.2 Probably approximately correct (PAC) learning

Consider a decision tree learning algorithm that considers only examples described by Boolean features $\langle X_1, \ldots X_n \rangle$, learns only Boolean-valued functions ($Y \in \{+, -\}$), and outputs only 'regular, depth-2 decision trees.' A 'regular, depth-2 decision tree' is a depth two decision tree (a tree with four leaves) in which the left and right child of the root are *required to test the same attribute*. For example, the tree in Figure 1 is a 'regular, depth-2 decision tree.'

1. Suppose you have noise-free training data for target concept c which you know can be described by a regular, depth-2 decision tree. How many training examples must you provide the learning algorithm in order to assure that with probability .99 the learner will output a tree whose true accuracy is at least .97? Assume you have data with 20 attributes in total (though of course you believe only two of these twenty will be needed to describe the correct tree).
2. Suppose you modify the algorithm to handle instances that have real-valued attributes instead of Boolean attributes, and you allow each decision tree node to perform a Boolean threshold test of the form $X_i > a$ where a is allowed to take on any real value. The tree is further

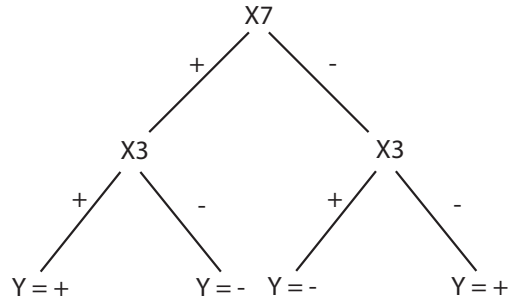

Figure 1: An example of a regular Boolean depth-2 decision tree.

constrained such that the two second level nodes, must both test the same attribute *and* use the same threshold. In this case, re-answer the above question: How many training examples must you provide the learning algorithm in order to assure that with probability .99 the learner will output a tree whose true accuracy is at least .97? In this case, assume that each example has only two attributes, so the tree will end up using both. You can still assume that the target concept c is in the new hypothesis space.

2 Support Vector Machines [Purna, 25 points]

Consider the one-dimensional dataset consisting of points $\{x, y\}$, where x is a real value, and y is the class variable s.t. $y \in \{-1, 1\}$. The points are $\{-1, -1\}, \{0, 1\}$ and $\{1, -1\}$.

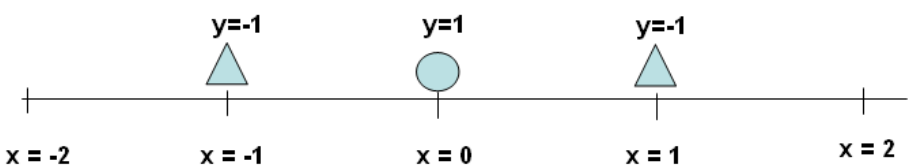

Figure 2: Data for SVM

2.1 SVM with Finite Number of Features

1. Can the dataset in figure 2 (in its current feature space) be perfectly separated using a linear separator? Why or why not?

2. Now we will define a simple feature map $\phi(u)$ to transform the the points from one to two dimensional feature space. We will define $\phi(u) = (u, u^2)$. Can a linear separator perfectly separate the points in this new feature space? Why or why not? *Hint: Plotting the points in the new two dimensional feature space might help.*

3. Construct a maximum margin separating hyperplane. This will be a line in this new two dimensional feature space. This line will have three parameters, i.e. $w_0 + w_1 y_1 + w_2 y_2 = 0$. Here $y_1 = x$, and $y_2 = x^2$. Give the values of w_0, w_1 and w_2. Compute the margin of this hyperplane. *Use your geometric intuition to solve this problem.*

4. Draw the transformed points (in two dimensions) and the maximum margin hyperplane. Also circle the support vectors. Now draw the decision boundary in the original one dimensional setting.

5. If we add a negative point (a triangle) to position $x = 2$, would the decision boundary change? Why or why not? If you were asked to add a negative point x such that $x > 0$ and the decision boundary is different on the modified dataset (one including the newly added negative point) where would you add it?

2.2 SVM with Infinite Number of Features [Extra Credit]

Now we will consider a more complicated feature mapping. Consider the following function $\phi_n(x)$ which maps a point from 1 to n dimentions.

$$\phi_n(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, ..., \frac{e^{-x^2/2}x^i}{\sqrt{i!}}, ..., \frac{e^{-x^2/2}x^n}{\sqrt{n!}} \right\}$$

In the above mapping if we let $n \to \infty$ we obtain the new feature transformation in (1). In the former problem we defined a feature map which transforms a point from one dimension to two. In this question we will work with the above feature map, which transforms a point from one dimension to an infinite number of dimensions.

$$\phi_\infty(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, ..., \frac{e^{-x^2/2}x^i}{\sqrt{i!}}, ... \right\} \tag{1}$$

The kernel trick in SVM is that we do not need to explicitly construct a feature. We only need to be able to enumerate the inner product of points in the transformed space. Its clearly impossible to explicitly construct the transformed feature vector. In this question we will see how we can classify points in the infinite dimensional space without having to actually computing the feature vectors explicitly. We will start by defining inner product between two infinite vectors $a = \{a_1, ..., a_i, ...\}$, and $b = \{b_1, ..., b_i, ...\}$ as $K(a, b) = a \cdot b = \sum_{i=1}^{\infty} a_i b_i$.

For two scalar numbers x and y, Prove the following,

$$K(x, y) = \phi_\infty(x) \cdot \phi_\infty(y) = \exp^{-\frac{(x-y)^2}{2}} \tag{2}$$

Hint: The Taylor series expansion of e^x is given by $\sum_{i=0}^{\infty} \frac{x^i}{i!}$.