Assignment 4: Sample Complexity

CS485/685 – Winter 2012

Out: March 15, 2012
Due: March 29, 2012, at the beginning of the lecture. Late assignments may be submitted in the pink drop off box on the third floor of MC within 24 hrs for 50% credit.

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

1. **[20 pts]** Naive Bayes Model

   Consider a Naive Bayes model for classification. The naive Bayes model is similar to the mixture of Gaussians model, but deals with discrete inputs (instead of continuous inputs) in the sense that it corresponds to a mixture of discrete distributions. Suppose that we have d binary attributes $A_1, ..., A_d$ and a binary class C. The Naive Bayes model is parametrized by a prior distribution $\Pr(C)$ and d conditional distributions $\Pr(A_i|C)$. The probability that instance $(a_1, a_2, ..., a_d)$ is in class c is

   $$\Pr(C = c|A_1 = a_1, A_2 = a_2, ..., A_d = a_d) = kPr(C = c) \prod_{i=1}^{d} \Pr(A_i = a_i|C = c)$$

   where k is a normalization constant. We select the class c^* with the highest probability by verifying that

   $$\Pr(C = c^*|A_1 = a_1, A_2 = a_2, ..., A_d = a_d) \geq \Pr(C = c|A_1 = a_1, A_2 = a_2, ..., A_d = a_d) \;\; \forall c$$

   Prove that the VC dimension of the space of naive Bayes models for d binary attributes and a binary class is at most $4d + 3$. Hint: consider the log of the above inequality.

2. **[60 pts]** Perceptron

   In this question, we analyze the sample complexity of the perceptron algorithm for binary classification. Recall that the perceptron finds a linear separator parametrized by a vector of weights $\mathbf{w}$. This wector of weights is obtained by adding or subtracting misclassified instances.

   (a) **[10 pts]** Suppose that there is a linear separator that can achieve 100% accuracy for any dataset. What is your confidence that the perceptron algorithm will return a linear separator that misclassifies at most 10% of the future instances for any underlying distribution, if you train the linear classifier with a dataset of 10,000 instances in 10 dimensions?

   (b) **[10 pts]** Suppose that there is no perfect linear separator. You decide to use the perceptron algorithm anyway and to stop it after 2 hours of training. A linear sperator is found with a misclassification rate of 5% on the 10,000 training instances in 10 dimensions. What is the smallest misclassification rate that can be guaranteed to hold 90% of the time for any underlying distribution?

   (c) **[40 pts]** Suppose that the classification task is spam filtering. You decide to encode each email as a vector of 10 bits corresponding to the absence or presence of 10 words. In the perceptron algorithm, suppose that you initialize all the weights to 0 and then adjust the weights by adding or subtracting misclassified instances.

      i. **[5 pts]** What is the hypothesis space?

ii. **[5 pts]** Suppose that you bound the weights to have an absolute value no greater than 10. In other words, misclassified emails are used to update the weights only when their absolute value does not get larger than 10. How large is the hypothesis space?

iii. **[10 pts]** Suppose that you have reasons to believe that the best linear separator should have small weights. Propose a distribution that gives higher probability to weights with smaller absolute values. Many answers are possible here.

iv. **[10 pts]** Suppose that the linear separator found by the perceptron algorithm has the following weights: $(10, 0, -1, -5, 4, 3, 7, 0, -4, -1)$. The misclassification rate on the training data is 5%. Using your prior distribution over hypotheses, can you guarantee a lower misclassification rate for this linear separator than the bound that you found in b) for any underlying distribution with a confidence of at least 90%? If yes, give the lower misclassification rate. Either way, explain your answer.

v. **[10 pts]** How would you modify the perceptron algorithm to take into account your prior distribution over hypotheses in order to find a linear separator with a better guarantee for the misclassification rate?

3. **[20 pts]** Nearest Neighbor learning

Consider the spam filtering problem described above. However instead of using the perceptron algorithm, you will use a 1-nearest neighbor classifier. In other words, future emails are classified based on the label of the closest email (according to Euclidean distance) in the training set (10,000 emails) assuming that all emails are represented by a vector of 10 bits corresponding to the absence or presence of 10 words. Suppose that the true underlying labeling function is c-Lipschitz with $c = 0.01$.

(a) **[10 pts]** Suppose that the Bayes optimal classifier has an expected error rate of 10%. Give a bound on the expected error rate of the 1-nearest neighbor classifier?

(b) **[10 pts]** There is no reason to restrict ourselves to 10 word features. In fact, suppose that the expected error rate of the Bayes optimal classifier is 1 / # of word features. How many word features should you consider to guarantee the lowest expected misclassification rate for the 1-nearest neighbor algorithm?