CSCI567 Machine Learning (Fall 2008) Assignment #1

Instructor: Dr. Sofus A. Macskassy
TA: Cheol Han

Due time: 5:00pm, Sep 18, 2008

Student Name: _____

Student ID: _____

Score: _____

1. (**Logistic regression**, 10 points) In our definition of logistic regression, we defined the following equations:

$$P_1(x;w) = \frac{\exp w.x}{1+\exp w.x}$$

$$P_0(x;w) = 1 - p_1(x;w)$$

Show that the logistic function is as follows:

$$P_1(x_i;w) = \frac{1}{(1+\exp[-w.x_i\ ])}$$

Using the above function, show that this equality holds:

$$\log \frac{P_1(x;w)}{P_0(x;w)} = w.x$$

2. (**Reject option**, 20 points) In many applications, the classifier is allowed to "reject" a test example rather than classifying it into one of the classes. Consider, for example, a case in which the cost of a false positive is \$5 but the cost of having a human manually make the decision is only \$2, we can formulate this as the following loss matrix:

decision	true label y	
	0	1
predict 0	0	5
predict 1	5	0
reject	2	2

Suppose P(y=1|x) is predicted to be 0.2. In other words, the learned model predicts that y=1 (the instance should belong to class 1) with probability 0.2, given the input x. Which decision minimizes the expected loss? Now suppose P(y=1|x)=0.4, Now which decision minimizes the expected loss? Show that in cases such as this there will be two thresholds θ_0 and θ_1 such that the optimal decision is to predict 0 if $p_1 < \theta_0$, reject if $\theta_0 \le p_1 \le \theta_1$, and predict 1 if $p_1 > \theta_1$.

What are the values of θ_0 and θ_1 for the following loss matrix?

decision	true label y	
	0	1
predict 0	0	20
predict 1	15	0
reject	8	8

3. (**Hypothesis Spaces**, 30 points) Consider the hypothesis space of *m-of-n* functions. The *m-of-n* function states that if *m* out of *n* attributes are true, then predict +1. For example, consider the input space {x1,x2,x3,x4,x5}; a specific *m-of-n* function would be: 1-of-3 of {x1,x3,x5}, which means that if either x1, x3 or x5 are true in a given instance, then predict +1. In other words:

$$f(x1, x2, x3, x4, x5)_{(1-of-3\{x1,x3,x5\})} = \begin{cases} +1 & (x1 \vee x3 \vee x5) = T \\ -1 & otherwise \end{cases}$$

A) What is the size of the *m-of-n* hypothesis space as a value of *N*, the number of boolean input values. In other words, how many possible hypotheses are there? The answer should be a function of *m*, *n* and *N*.

B) Which of these two is the more specific hypothesis and why? (if needed for your answer, assume that the input space consists of only 5 variables: x1, x2, x3, x4 and x5).

2-of-3 of {x1,x2,x4}
1-of-4 of {x1,x2,x4,x5}

C) When considering the parameter space of 3 boolean values {x1,x2,x3}, what is an example of the most general (non-trivial). What is the most specific hypothesis?

4. (**Hypothesis Spaces**, 10 points) Imagine a hypothesis space of not one rectangle but a union of *m* rectangles (*m > 1*).

A) What is the advantage of such a hypothesis class?

B) Show that *any* class can be represented by such a hypothesis, given a large enough *m.*

C) If f(**x**) = +1 if **x** lies within one of the rectangles, how large would *m* need to be in the worst case?

5. (**Linear Threshold Units**, 30 points) Consider the points
<1,2,+1>
<1,1,+1>
<2,4, -1>
<2,1,+1>
<3,4, -1>
<3,3, -1>

As we can see, training instances have 2 inputs and the class label is either -1 or +1.
Assume we are using hinge loss as the loss function.
Review slides from lecture 4 for the algorithms to use.
Note that we here provide a different initial weight vector than given in class!

A) Batch Perceptron Algorithm, learning rate $\eta = 1$, weight vector **w0 =** <0.1,0.5,0.3>.
Calculate, by hand, the weight vectors (**w1,w2**) after the first two iterations. Show both the gradient vector and the weight vector.

B) Online Perceptron Algorithm, learning rate $\eta = 1$, weight vector **w0 =** <0.4,0.1,0.3>.
Calculate and show the final weight vector after having run through all the examples in the order given.

C) Batch Logistic Regression, learning rate $\eta = 1$, weight vector **w0 =** <0.3,0.2,0.5>.
NOTE: Use y=0 for negative instances.
Calculate, by hand, the weight vectors (**w1,w2**) after the first two iterations. Show both the gradient vector and the weight vector.