CS534 — Homework Assignment 1 — Due Monday April 4, 2004

1. (Probability Decision Boundary). Consider a case where we have learned a conditional probability distribution $P(y|\mathbf{x})$. Suppose there are only two classes, and let $p_0 = P(y = 0|\mathbf{x})$ and $p_1 = P(y = 1|\mathbf{x})$. Consider the following loss matrix:

predicted label $\hat{y}$	true label y	
	0	1
0	0	10
1	5	0

   Show that the decision $\hat{y}$ that minimizes the expected loss is equivalent to setting a probability threshold θ and predicting $\hat{y} = 0$ if $p_1 < \theta$ and $\hat{y} = 1$ if $p_1 \geq \theta$. What is this threshold for this loss matrix? Show a loss matrix where the threshold is 0.1.

2. (Reject Option). 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 misclassification is \$10 but the cost of having a human manually make the decision is only \$3. We can formulate this as the following loss matrix:

decision	true label y	
	0	1
predict 0	0	10
predict 1	10	0
reject	3	3

   Suppose $P(y = 1|\mathbf{x})$ is predicted to be 0.2. Which decision minimizes the expected loss? Now suppose $P(y = 1|\mathbf{x}) = 0.4$. 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 $\theta_0 \leq p_1 \leq \theta_1$, and predict 1 if $p_1 > \theta_1$.

   What are the values of these thresholds for the following loss matrix?

decision	true label y	
	0	1
predict 0	0	10
predict 1	5	0
reject	3	3

3. (Weighted hinge loss). In our derivation of the Perceptron algorithm, we used the hinge loss to approximate the 0/1 loss. Suppose that we have a general loss matrix with the cost of a false positive being $L(1, -1) = c_0$ and the cost of a false negative $L(-1, 1) = c_1$. Suppose we used

$$\tilde{J}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} z_i \max(0, -y_i \mathbf{w} \cdot \mathbf{x}_i)$$

   for our approximate objective function, where $z_i = c_0$ if $y = -1$ and $z_i = c_1$ if $y = 1$. Compute the gradient using this approximation, and show how the batch Perceptron algorithm is modified to incorporate this change.

4. (Dual Perceptron Algorithm). Consider the following learning algorithm known as the dual perceptron algorithm:

   Let $\alpha_i = 0$ for $i = 1, ..., N$
   Repeat forever
   Accept training example $\langle \mathbf{x}_i, y_i \rangle$
   if $\sum_\ell \alpha_\ell \mathbf{x}_\ell \cdot \mathbf{x}_i y_i y_\ell < 0$
   $\alpha_i = \alpha_i + 1$

   In other words, α_i is a counter of the number of times training example $\mathbf{x}_i$ has been misclassified.

   Prove that this algorithm is equivalent to the online perceptron algorithm with learning rate 1 and weight vector $\mathbf{w} = \sum_\ell \alpha_\ell \mathbf{x}_\ell y_\ell$.

5. In our definition of logistic regression, we defined

$$p_1(\mathbf{x}; \mathbf{w}) = \frac{\exp \mathbf{w} \cdot \mathbf{x}}{1 + \exp \mathbf{w} \cdot \mathbf{x}}.$$

$$p_0(\mathbf{x}; \mathbf{w}) = 1 - p_1(\mathbf{x}; \mathbf{w}).$$

   Show that this is equivalent to

$$\log \frac{p_1(\mathbf{x}; \mathbf{w})}{p_0(\mathbf{x}; \mathbf{w})} = \mathbf{w} \cdot \mathbf{x}.$$

   Show also that

$$p_1(\mathbf{x}_i; \mathbf{w}) = \frac{1}{(1 + \exp[-\mathbf{w} \cdot \mathbf{x}_i])}.$$

   This is known as the *logistic function*.