CS 35000 – Introduction to AI
The University of Chicago, Winter 2005

Problem Set 1. Due: Tuesday 4/1

(Please attempt all problems by yourself without consultation with classmates or friends. If any question is unclear, contact me (niyogi@cs) or the TA (matveeva@cs).)

1. In class, we covered the Perceptron Learning Algorithm (PLA) for the case when the data was linearly separable by functions of the sort

$$y = 1 \textbf{ if } (\mathbf{w}.\mathbf{x} \geq 0) \textbf{ otherwise } y = -1$$

where $\mathbf{w}$ and $\mathbf{x}$ are k dimensional vectors. We showed how to update the "weight vector" $\mathbf{w}$ from mistake to mistake so that ultimately the data was separated. This is equivalent to assuming that the data was separable by hyperplanes passing through the origin. Suppose, this is not the case, i.e., the data is linearly separable but by hyperplanes of the sort

$$y = 1 \textbf{ if } (\mathbf{w}.\mathbf{x} + b \geq 0) \textbf{ otherwise } y = -1$$

where b is a scalar offset. Show how to modify the PLA to update both the weight vector $\mathbf{w}$ and the offset b from mistake to mistake so that one arrives at values for the weight vector and the offset that will separate the data.

2. Linear hyperplanes such as those provided by the Perceptron Learning Algorithm are useful only when the data is linearly separable. Consider the following four labeled data points in two-dimensions, i.e., each labeled datapoint is an $(\mathbf{x}, y)$ pair where $y \in \{-1, 1\}$ and $\mathbf{x} \in R^2$. The four data points are as follows: (we denote by x_1 and x_2 the two "coordinates" of the data point $\mathbf{x}$ and by y the label of the data)

x_1	x_2	y
2	2	+1
0	0	+1
2	0	-1
0	2	-1

By plotting the data on a plane, it is clear that the data are not linearly separable. Consider the possibility of mapping the data into a new space Z where the mapped data points are linearly separable. Can you come up with two functions (with two real valued variables as input and one real valued variable as output) f_1 and f_2 such that (i) each data point $((x_1, x_2), y)$ is mapped onto $((f_1(x_1, x_2), f_2(x_1, x_2)), y)$ (ii) the mapped data points can now be separated by the classical hyperplanes (with no offset term) that we discussed in class.

3. Linear hyperplanes (with no offset) are required to satisfy the following property to completely separate the labeled data. If the data point $\mathbf{x}$ has label $y = +1$ then $\mathbf{w}.\mathbf{x} \geq 0$. If the label is $y = -1$ then $\mathbf{w}.\mathbf{x} < 0$. Show that this means that for a separating hyperplane, the product of y_i

and $\mathbf{w}.\mathbf{x}_i$ is always positive for each datapoint $(\mathbf{x}_i, y_i)$. Using this intuition, one might try to design hyperplanes by making this product as positive as possible. Consider the following problem:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} y_i(\mathbf{w}.\mathbf{x}_i)$$

Find a solution to this problem subject to the constraint that $|\mathbf{w}| = 1$. (**Hint:** Recall that $|\mathbf{z}| = (\sum_{i=1}^{k} z_i^2)^{1/2}$ where $\mathbf{z}$ is a k-dimensional vector and z_i are the components of that vector. You will need to use the result that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ the value of $\mathbf{a}$ that maximizes the product $\mathbf{a}.\mathbf{b}$ subject to the constraint that $|\mathbf{a}| = 1$ is simply $\mathbf{a}_{opt} = \frac{\mathbf{b}}{|b|}$.)

4. Let the conditional densities for a two-category one-dimensional problem be given by the Cauchy distribution

$$p(x|y=i) = \frac{1}{\pi b}\left(\frac{1}{1+(\frac{x-a_i}{b})^2}\right); i = 0, 1$$

If $P(y=0) = P(y=1)$, show that $P(y=0|x) = P(y=1|x)$ if $x = \frac{1}{2}(a_0 + a_1)$. Sketch $P(y=0|x)$ for the case $a_0 = 3$, $a_1 = 5$ and $b = 1$. How does $P(y=0|x)$ behave (i) as $x \to -\infty$? (ii) as $x \to +\infty$?

5. A multiclass (as opposed to two class) problem is one in which the data can belong to one of m classes where $m > 2$. Indicate by $P(x|y=i)$ the conditional probability of x given the ith class where i can take on values in $\{1, \ldots, m\}$. Indicate by $P(y=i)$ the prior probability of the ith class. A decision rule must map each data point X into one of m values. Therefore (analogous to the two class problem) it must map each $x \in X$ into $\{1, \ldots, m\}$. What is the optimal decision rule for this case?

6. The perceptron algorithm as discussed in class works for two class problems where the data is separable. Suppose we have four classes. How would you design a collection of perceptrons to solve such a four class problem?

7. In class, we considered *deterministic* decision rules given by functions $(g : X \to \{0, 1\})$ where for each point $x \in X$, the rule made a deterministic guess about the (unknown) label y. We derived the optimal decision rule for this case. Consider now the class of all *randomized* decision rules given by $\alpha(x) : X \to [0, 1]$. Thus, for each point x, the randomized rule α guesses the label to be $y = 1$ with probability $\alpha(x) \in [0, 1]$ and $y = 0$ with probabability $1 - \alpha(x)$ (i.e., for each x, $\alpha(x)$ is a number between 0 and 1 characterizing the probability of guessing). Show that the best randomized decision rule will not have a lower probability of error than the best deterministic rule. Therefore, randomization does not buy us any additional power.

8. Let the components of the vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$ (d odd) be binary valued (1 or 0). Further, let

$$p_{ij} = Prob[x_i = 1|y = j]; i = 1, \ldots, d; j = 1, 2$$

with the components x_i being statistically independent for each $y = j$. (This means simply that $Prob[x_i = m, x_k = n|y = j] = Prob[x_i = m|y = j]Prob[x_k = n|y = j]$ for $m, n \in \{0, 1\}$). Consider the special case when

$$p_{i1} = p > \frac{1}{2}$$

and

$$p_{i2} = 1 - p$$

Let the prior probabilities $P(y = 1) = P(y = 0) = \frac{1}{2}$.

(a) Show that the minimum error decision rule becomes
Decide $y = 1$ if $\sum_{i=1}^{d} x_i > \frac{d}{2}$.

(b) Show that the minimum probability of error is given by

$$P_e(d,p) = \sum_{k=0}^{(d-1)/2} C_k^d p^k (a-p)^{(d-k)}$$

where C_k^d is the number of ways of choosing k distinct objects from d distinct objects. $(= \frac{d!}{(d-k)!d!})$.

(c) What is the limiting value of $P_e(d,p)$ as $p \to \frac{1}{2}$?

9($*$; Optional). In class, we obtained an upper bound on the number of updates the perceptron learning algorithm can make if the data is linearly separable. This was seen to be in terms of the margin

$$\delta = \max_w \min_{i=1,\ldots,k} |w.x_i|$$

Suppose all the x_i's were on the sphere or radius R in n dimensions, can you provide upper or lower bounds on δ as a function of R, n, k.