The University of Connecticut **Spring 2021**
Dept. of ECE **KRP**

Problem Set # 4
(Due March 23, 2021)

1. Consider a binary classification problem with 4 patterns in a 2-dimensional feature space, with at least one sample from each class. Show that 7/8 of all the labeling possibilities into one of two classes are linearly separable.

   Indeed, there is a theorem called Cover's theorem that shows why linear separabiity is rare in practice unless the dimension of the feature space is very large. Consider a p-dimensional space and N patterns scattered randomly in this space. We assume that any (p+1) of these patterns will not fall in a (p-1) dimensional subspace (because this would be very coincidental, that is, with probability zero). In that case, we say that the patterns are in general position. Assume that these patterns are assigned randomly to either of two classes, C_1, C_2. Of all possible 2^N assignments, a certain fraction $f(N,p)$ is linearly separable. Cover's theorem shows that this fraction is given by the formula

$$f(N,p) = \begin{cases} 1 \ \textbf{if} \ N \le p+1 \\ \frac{2}{2^N}\sum_{i=0}^{p}\binom{N-1}{i} \ \textbf{if} \ N > p+1 \end{cases}$$

   Notice that for $N=p+1$ patterns or less, the fraction is 1. This simply means that a single hyperplane can always perform the separation correctly. Even for N=2(p+1), called the capacity of the hyperplane, half of the assignments are linearly separable. However, for large N, the fraction quickly goes to zero.

2. Consider the softmax function

$$\phi_m = \frac{e^{t_m}}{\sum_{k=1}^{C} e^{t_k}}$$

   Show that

$$\frac{\partial \phi_m}{\partial t_k} = \phi_m(\delta_{mk} - \phi_k); m = 1,2...,C; k = 1,2,..,C$$

   where

$$\delta_{mk} = \begin{cases} 1, if \ m = k \\ 0, otherwise \end{cases}$$

3. A potential function is used as the discriminant function

$$z(\mathbf{x}, \mathbf{w}) = \frac{1}{(1+ \| \mathbf{x} - \mathbf{w} \|^2)}$$

   Show that the corresponding update equation is:

$$\mathbf{w}^{(m+1)} = \mathbf{w}^{(m)} \pm \eta[z(\mathbf{x},\mathbf{w})]^2(\mathbf{x}-\mathbf{w})$$

4. The purpose of this exercise is to show that there is a great deal of flexibility in selecting step sizes when experimenting with stochastic approximation. As you know, stochastic approximation algorithms require the step size sequence $\{\eta_k\}$ to satisfy the following two conditions:

$$\sum_{k=1}^{\infty} \eta_k = \infty \textbf{ and}$$

$$\sum_{k=1}^{\infty} \eta_k^{\,2} < \infty$$

a) Show that the sequence $\eta_k = 1$ for $1 \le k \le 100$, $\eta_k = k/1000$ for $100 < k \le 500$ and $\eta_k = 1/(200+k)$ for $k > 500$ satisfies both conditions.

b) Show that the sequence $\eta_k = \eta_0 \dfrac{(k/K)+1}{(k/K)^2+1}$ where $0 < \eta_0 < 1$ and K is a positive integer, satisfies both conditions.

5. We want to formulate the cross entropy cost function used in logistic regression when there is a small probability ε that the class label on a training data point has been incorrectly set. Consider a binary classification problem in which the target values $z \in \{0,1\}$, with a network output $y(\underline{x}, \underline{w})$ that represents $P(z=1|\underline{x})$ and suppose that the probability that the label is correctly set in the training data is (1- ε). Assuming independent and identically distributed data, write down the error function corresponding to the negative log likelihood. Explain why this error function makes the model robust to incorrectly labeled data, in contrast to the usual error function.

6. Theodoridis, Problem 7.15.

7. Let $\hat{\theta}_u$ be an unbiased estimator of θ so that $E(\hat{\theta}_u) = \theta$. Recall that for an unbiased estimator $MSE(\hat{\theta}_u) = E\{(\theta - \hat{\theta}_u)^2\} = \sigma_u^2$. Now, define a biased estimator $\hat{\theta}_b = (1+\alpha)\hat{\theta}_u$.
Show that the range of α where the MSE of $\hat{\theta}_b$ is smaller than that of $\hat{\theta}_u$ is

$$-2 < -\frac{2\sigma_u^2}{\sigma_u^2+\theta^2} < \alpha < 0.$$

Show that the optimal value of α that minimizes $MSE(\hat{\theta}_b)$ is

$$\alpha^* = \frac{-1}{1+\dfrac{\theta^2}{\sigma_u^2}}$$

Suppose N data points are generated according to $y_n = \theta + v_n$, n=1,2,..,N, where $v_n \sim N(0,\sigma^2)$. What are $\hat{\theta}_u$ and σ_u^2 when the maximum likelihood (ML) method (in this case

least squares estimation) is used for estimating θ. Now, suppose instead of least squares, we use a regularized least squares method with a parameter λ, that is, we minimize

$$J(\theta,\lambda)=\frac{1}{2}\sum_{n=1}^{N}(y_n-\theta)^2+\frac{1}{2}\lambda\theta^2$$

Show that this results in a biased estimator $\hat{\theta}_b$. Express α in terms of λ and N. What is the optimal λ^*. Is MSE of $\hat{\theta}_b$ less than σ_u^2 at optimal λ^*?

8. Consider the two category problem and the following training patterns, each having four binary attributes:

z=1	z=0
1100	1100
0000	1111
1010	1110
0011	0111

.

(a) Use Information gain algorithm to create by hand an unpruned classifier for this data.
(b) Apply simple logical reduction methods to your tree in order to express each category with the fewest ANDs and ORs.
(c) Suppose it is known that during testing, the prior probabilities of the two categories are not equal, but instead P(z=1)=2P(z=0). Modify your training method and use the above data to from a new tree and new set of simplified rules.

9. Consider training a binary decision tree to classify two-component patterns from two categories using the following 6 samples from each class. The first component is binary, 0 or 1, while the second component has six possible values, A through F.

z=1: 1A 0E 0B 1B 1F 0D
z=0: 0A 0C 1C 0F 0B 1D

Compare splitting the root node based on the first feature with splitting on the second feature in the following way.

1. Use the information gain with a two way split on the first feature and a six way split on the second feature.
2. Repeat (a) using Gini index.

10. Consider the training set

z=1: (111), (100)
z=0: (110), (001)

Derive a decision tree of depth 2 that attains zero training error. Can the information gain algorithm attain zero error with a decision tree of depth 2? What about one with JMI criterion?

11. (Computational)
Experiment with the Perceptron, Random Forests, AdaBoost, and Gradient Boosting learning algorithms on the datasets you have selected. Use MATLAB or any NN software packages you have access to.