E1 213 Pattern Recognition and Neural Networks

Practice Problems: Set 3

1. Consider a two class problem with one dimensional feature space. Suppose we have six training samples: x_1, x_2, x_3 from one class and x_4, x_5, x_6 from the other class. Suppose we want to estimate the class conditional densities nonparametrically through a Parzen window estimate with Gaussian window with width parameter σ. Write an expression for the Bayes classifier (under 0–1 loss function) which uses these estimated densities.

2. Consider the kernel density estimate given by

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n}\phi\left(\frac{x - x_i}{h_n}\right)$$

Let the function ϕ be given by $\phi(x) = \exp(-x)$ for $x > 0$ and it is zero for $x \leq 0$. Suppose the true density (from which samples are drawn) is uniform over $[0, a]$. Show that the expectation of the density estimate is given by

$$E\hat{f}_n(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{a}\left(1 - e^{-x/h_n}\right) & \text{for } 0 \leq x \leq a \\ \frac{1}{a}\left(e^{a/h_n} - 1\right)e^{-x/h_n} & \text{for } x \geq a \end{cases}$$

Is this a good approximation to uniform density? Explain.

3. Consider 2-class PR problems with n Boolean features. Consider two specific classification tasks specified by the following: (i) a feature vector X should be in Class-I if the integer represented by it is divisible by 4, otherwise it should be in Class-II; (ii) a feature vector X should be in Class-I if it has odd number of 1's in it, otherwise it is in Class-II. In each of these two cases, state whether the classifier can be represented by a Perceptron; and, if so, show the Perceptron corresponding to it; if not, give reasons why it cannot be represented by a Perceptron.

4. Consider the incremental version of the Perceptron algorithm. The algorithm is: at iteration k, if $W(k)TX(k) \leq 0$ and thus we misclassified the next pattern then we correct the weight vector as: $W(k + 1) =$

$W(k) + X(k)$.
(i). By going over the proof presented in class, convince yourself that if we change the algorithm to $W(k+1) = W(k) + \eta X(k)$ for any positive step-size η, the proof is still valid.
(ii). In the perceptron algorithm, when we misclassify a pattern and hence correct the weight vector, the algorithm does not necessarily ensure that $W(k+1)$ will classify $X(k)$ correctly. Suppose we want to change the algorithm so that when we misclassify a pattern, we change the weight vector by an amount that ensures that after the correction, the weight vector correctly classifies this pattern. While this may seem like just a matter of choosing a 'step-size', note that if we want to choose η so that the above is ensured at every k then, the 'step-size' may have to vary from iteration to iteration and it may be a function of the feature vector. Hence, the earlier proof may not go through. Design a simple modified version of the Perceptron algorithm which effectively ensures the above property and for which the same convergence proof holds.

5. Consider the joint density of X, Y given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2\sigma^2(1-\rho^2)}(x^2+y^2-2\rho xy)\right), \quad -\infty < x, y < \infty$$

Find a function g that minimizes $E\left[(g(X) - Y)^2\right]$.

6. Suppose we have $y = \mathbf{a}^T X + \xi$ where ξ is a zero-mean random variable with variance σ^2. Under this model we have calculated in the class the variance of the least squares solution, W^*. Calculate the expected value of the least squares solution.

7. We can pose the problem of learning a linear classifier as minimizing

$$J(W) = \sum_{i=1}^{n} L(W^T X_i, y_i)$$

where L is a loss function. For least squares criterion, we take $L(a,b) = (a-b)^2$. If, instead we want to minimize absolute value of error, we can take $L(a,b) = |a-b|$. Show that logistic regression (in the 2-class case) can also be put in this framework with $L(W^T X, y) = \ln(1 + \exp(-yW^T X))$, where we assume that the class labels are $+1$ and -1.

What would be the loss function corresponding to mult-class logistic regression?

8. Consider a classification problem with K classes: $C_1, \cdots, C_K$. We say that the training set is linearly separable if there are K functions: $g_j(X) = W_j^T X + w_{j0}, \;\; j = 1, \cdots, K$, such that we have $g_i(X) \geq g_j(X), \forall j$, whenever $X \in C_i$. We say that a set of examples is totally linearly separable if given any C_i, there is a hyperplane that separates examples of C_i from the set of examples of all other claases. Show that totally linearly separable implies linearly separable but the converse need not be true.