ECEN 649 Pattern Recognition – Spring 2018

Problem Set 3

Due on: Apr 10

1. Consider a linear discriminant $g(x) = a^t x + b$.

   (a) Use the method of Lagrange multipliers to show that the distance of a point x_0 to the hyperplane $g(x) = 0$ is given by $|g(x_0)|/||a||$.

   (b) Use the previous result to show that the margin in a linear SVM $g(x) = a^t x + b = 0$ is given by $1/||a||$.

2. Consider the following training data consisting of 4 points:

   $$x_1 = (-1, 1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, -1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

   (a) Run manually the perceptron algorithm for these training data, considering the initial parameters to be $a(0) = (1, 0)$ and $a_0(0) = 0$, and a fixed step length $\ell = 1$. Plot the designed perceptron classifier.

   (b) By assuming the same initial parameters, find the condition on the fixed step length ℓ that allows the perceptron algorithm to find a solution after a single iteration.

   (c) By assuming the same initial parameters, and a fixed step length $\ell = 1$, show what happens with the perceptron algorithm if the training data are instead:

   $$x_1 = (-1, -1), y_1 = 1,\ x_2 = (1, 1), y_2 = 1,\ x_3 = (-1, 1), y_3 = 0,\ x_4 = (1, -1), y_4 = 0.$$

   Can you fix this by changing ℓ or the initial parameters?

3. Show that the polynomial kernel $K(x, y) = (1 + x^T y)^p$ satisfies Mercer's condition.

4. Consider a network with l and m neurons in two hidden layers (see Figure 30.3 in DGL). This network is specified by:

   $$\zeta(x) = c_0 + \sum_{i=1}^{l} c_i \xi_i(x)$$

   where $\xi_i(x) = \sigma(\phi_i(x))$, for $i = 1, \ldots, l$, and

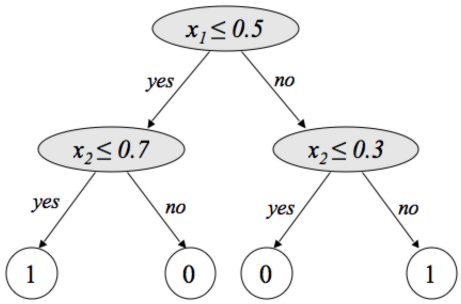
   $$\phi_i(x) = b_{i0} + \sum_{j=1}^{m} b_{ij} \upsilon_j(x)\,, \quad i = 1, \ldots, l$$

where $\upsilon_j(x) = \sigma(\chi_j(x))$, for $j = 1, \ldots, m$, and

$$\chi_j(x) = a_{j0} + \sum_{k=1}^{d} a_{jk} x_k \,, \quad j = 1, \ldots, m$$

Determine the backpropagation algorithm updates for the coefficients c_i, b_{ij}, and a_{jk}. Find the backpropagation equation(s) for this problem.

5. Consider the simple CART classifier in R^2 depicted below, consisting of three splitting nodes and four leaf nodes.



Design an equivalent two-hidden-layer neural network with threshold sigmoids, with three neurons in the first hidden layer and four neurons in the second hidden layer (note the correspondence with the numbers of splitting nodes and leaf nodes).

6. This problem concerns a parallel between discrete classification and Gaussian classification. Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a discrete feature vector, such that $\mathbf{X} \in \{0,1\}^d$, i.e., all features are binary. Assume furthermore that the features are conditionally independent given $Y = 0$ and given $Y = 1$, i.e., the features are independent "inside" each class — compare this to spherical class-conditional Gaussian densities, where the features are also conditionally independent given $Y = 0$ and given $Y = 1$.

   (a) As in the Gaussian case with equal covariance matrices, prove that the Bayes classifier is linear, i.e., show that $\psi^*(x) = I_{g(\mathbf{x})>0}$, for $\mathbf{x} \in \{0,1\}^d$, where the discriminant $g(\mathbf{x})$ is given by

$$g(\mathbf{x}) \;=\; a^T \mathbf{x} + b \,.$$

   Give the values of a and b in terms of the class-conditional distribution parameters $p_i = P(X_i = 1|Y = 0)$ and $q_i = P(X_i = 1|Y = 1)$, for $i = 1, \ldots, d$ (notice that these are different than the parameters p_i and q_i defined in the lecture), and the prior probability $c = P(Y = 1)$.

(b) Suppose that sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ is available, where $\mathbf{X}_j = (X_{j1}, \ldots, X_{jd})$, for $j = 1, \ldots, d$. Just as is done for LDA in the Gaussian case, obtain a sample discriminant $g_n(\mathbf{x})$ from $g(\mathbf{x})$ in the previous item, by plugging in maximum-likelihood (ML) estimators $\hat{p}_i$, $\hat{q}_i$, and $\hat{c}$ for the unknown parameters p_i, q_i, and c. The maximum-likelihood estimators in this case are the empirical frequencies (you can use this fact without showing it). As in LDA, show that the designed discrete classifier $\psi_n(\mathbf{x}) = I_{g_n(\mathbf{x})>0}$, for $\mathbf{x} \in \{0,1\}^d$, is linear, by showing that

$$g_n(\mathbf{x}) = \mathbf{a}_n^T \mathbf{x} + b_n \,.$$

Give the values of $\mathbf{a}_n$ and b_n in terms of $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.