EE5600: Introduction to AI & ML, Fall 2018 (12)

Indian Institute of Technology Hyderabad

HW 1, Assigned: Monday 20.08.2018.

Due: Monday 27.08.2018 at 11:59 pm.

The Force is with you, young Machine Learner. You are not a Jedi yet!
Don't be drawn to the Dark Side of online solutions!

1 Theory

1. Recall the statistical setting for supervised learning where the input $\mathbf{x}$ and the label y are random variables whose distribution is given by $p(\mathbf{x}, y)$. Under this assumption show that the optimal estimator for the average squared error criterion is given by $E[y|\mathbf{x}]$. (5)

2. In the statistical setting, show how the average squared error for an arbitrary model $\hat{y}(\mathbf{x})$ can be written as the sum of square of bias, variance and noise. (5)

3. Derive the least squares solution for a K-class linear discriminant classifier. (5)

4. Derive the Fisher's linear discriminant for the two-class classifier case. (5)

5. A *zero-one* loss function assigns a zero to a correct classification and a one to a misclassification. Find the optimal label estimator for this loss function under the statistical setting for supervised learning. (5)

2 Programming

1. Implement the two-class naive Bayes classifier assuming that the conditional distribution of the feature vector elements is Gaussian. Your program must accept as input the training data (observations and labels) along with its dimensions. Use the attached training data files to predict the class of the test vectors $\mathbf{x}_{N+1} = [1, 1]^T$, $\mathbf{x}_{N+2} = [1, -1]^T$, $\mathbf{x}_{N+3} = [-1, 1]^T$, $\mathbf{x}_{N+4} = [-1, -1]^T$. The training files contain 1000 2-D samples (*X.csv*) and associated labels (*Y.csv*). You will need to read these csv files. In *X.csv*, the first 1000 samples correspond to the first row of x and the second 1000 samples correspond to the second row of x. (10)

2. Implement the two-class k-nearest neighbor classifier on the same dataset (train and test) as in the previous problem. Experiment with k and report your performance. (10)