COEN 240 Machine Learning

Spring 2019

Homework #1

Assigned on April 12, 2019, Due on April 23, 2019

Guideline: Please complete the following problems and generate a PDF file. Please submit the PDF file and a separate zip file that contains all source code to Camino. Please refer to HomeworkFormat.pdf for the format of the submitted PDF file.

Problem 1 You have a set of N training inputs $\mathbf{x}_n \in \mathbb{R}^M, n = 1, 2, \dots, N, N \gg M$. The target outputs of the training inputs are $t_n \in \mathbb{R}, n = 1, 2, \dots, N$. Build a linear regression model to predict the target value by $\mathbf{w}^T\mathbf{x}_n$. Derive the closed-form solution for the weight vector $\mathbf{w} \in \mathbb{R}^M$ that minimizes the error function $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{w}^T\mathbf{x}_n - t_n\}^2$.

Problem 2 The Pima Indians diabetes data set (pima-indians-diabetes.xlsx) is a data set used to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage. The dataset consists of M = 8 attributes and one target variable, Outcome (1 represents diabetes, 0 represents no diabetes). The 8 attributes include Pregnancies, Glucose, BloodPressure, BMI, insulin level, age, and so on. There are N=768 data samples.

Randomly select n samples from the "diabetes" class and n samples from the "no diabetes" class, and use them as the training samples. The remaining data samples are the test samples. Build a linear regression model with the training set, and test your model on the test samples to predict whether or not a test patient has diabetes or not. Assume the predicted outcome of a test sample is $\hat{t}$, if $\hat{t} \geq 0.5$ (closer to 1), classify it as "diabetes"; if $\hat{t} < 0.5$ (closer to 0), classify it as "no diabetes". Run 1000 independent experiments, and calculate the prediction accuracy rate as $\frac{the\ number\ of\ correct\ predictions}{the\ total\ number\ of\ test\ cases}$ %. Let n=20, 40, 60, 80, 100, plot the accuracy rate versus n. Comment on the result. Attach the code at the end of the homework.

Problem 3 For the K-means clustering problem, when the binary indicators (responsibilities) r_{kn}'s are fixed for k=1, 2, …, K and n=1, 2, …, N, derive for the cluster centers $\mathbf{m}_k$, k=1, 2, …, K, such that the following objective function J is minimized:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \|\mathbf{m}_k - \mathbf{x}_n\|_2^2$$

Problem 4 Iris.xls contains 150 data samples of three Iris categories, labeled by outcome values 0, 1, and 2. Each data sample has four attributes: sepal length, sepal width, petal length, and petal width.

Implement the K-means clustering algorithm to group the samples into K=3 clusters. Randomly choose three samples as the initial cluster centers. Calculate the objective function value J as defined in **Problem 3** after the assignment step in each iteration. Exit the iterations if the following criterion is met: $J(\text{Iter}-1) - J(\text{Iter}) < \varepsilon$, where $\varepsilon = 10^{-5}$, and Iter is the iteration number. Plot the objective function value *J* versus the iteration number Iter. Comment on the result. Attach the code at the end of the homework.