COEN 240 Machine Learning

Homework #1

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. **Start with MicroSoft Word and then transform it into PDF**

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 $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{w}^T\mathbf{x}_n - t_n\}^2$. **It is like how we deduce the formula in class**

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 as described in **Problem 1** 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=40, 80, 120, 160, 200, plot the accuracy rate versus n. Comment on the result. Attach the code at the end of the homework.