Assignment 5

CS5824/ECE5424 – Fall 2020

Out: Nov 09, 2020
Due: Nov 19, 2020 (11:59pm)

For question 1, please submit PDF file saved from Latex. For other questions, submit both PDF and .ipynb file with all of your code and output. Both Colab and Jupyter are allowed. Please submit three files separately on Canvas, no need to submit a zip file.

Late submissions incur a 0.5% penalty for every rounded up hour past the deadline for the first 24 hours and 0.6% penalty for every rounded up hour past the deadline for the second 24 hours. For example, an assignment submitted 5 hours and 15 min late will receive a penalty of ceiling(5.25) * 0.5% = 3%. A grade of zero will be given on the third late day.

Be sure to include your name and student number with your assignment.

1. **[20 points] Theoretical Questions:**

   (a) **[10 points]** K-means Clustering
   Consider a clustering problem with K clusters and the corresponding means μ_k. For each data point x_n there is an associated indicator vector $\vec{r_n}$ which indicates which of the K clusters the data point belongs to. The kth element of the indicator vector $\vec{r_n} = [r_1, r_2, \cdots r_k, \cdots r_K]$ is 1 if x_n belongs to the kth class and the other elements are zeros. Hence for each data point x_n there are K indicator variables $r_{nk} \in \{0, 1\}$.
   Let us consider the following cost function for K means clustering:

   $$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{1}$$

   Derive a closed form estimate for the mean of the kth cluster μ_k

   (b) **[10 points]** Anomalies in Machine Learning
   Let us compare two estimators given by their corresponding cost functions given by:

   $$L1 = ||Y - XW||_2^2 \tag{2}$$

   $$L2 = |Y - XW| \tag{3}$$

   Which of the two estimators would you choose for a data corrupted with a portion $'\epsilon'$ of outliers, and only 1 outlier respectively, why?

2. **[80 points] Coding:**

   (a) **[40 points]** K-means Clustering
   K-means clustering is an unsupervised learning technique to group instances into clusters. An unlabeled

dataset "data1.mat" is provided in the assignment folder and 'X' stands for the training set. Please implement K-means clustering from scratch to group data points into 3 different clusters. Please randomly initialize the 3 centroids that are to be used in K-means on the dataset x. The iteration number is 10. You could implement k-means clustering based on the steps shown in course Clustering slides 10 to 15.
For submission,

i. **[20 points]** Please submit your code for k-means clustering from scratch.
ii. **[10 points]** Please report the centroids of the three clusters each iteration.
iii. **[10 points]** Please show the visualization for three clusters.

(b) **[40 points]** Anomaly Detection
Anomaly detection is used to detect outliers in the dataset. An unlabeled dataset "data2.mat" is provided in the assignment folder and please suspect that the vast majority of these examples are "normal" examples and some of them are outliers. There are three different variables in the dataset. 'X' stands for the training dataset, 'Xval' stands for the cross-validation set and 'yval' stands for the corresponding output for 'Xval'. Please implement anomaly detection algorithm to find the best threshold to use for selecting outliers and the number of outliers. You could implement anomaly detection based on the following steps:

i. Estimate Gaussian distribution for the training set and find mean and variance.

$$p\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_i^{(j)} \tag{5}$$

$$\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m}\left(x_i^{(j)} - \mu_i\right)^2 \tag{6}$$

ii. Find the probability of each data point by implementing multivariate Gaussian distribution. (http://cs229.stanford.edu/section/gaussians.pdf)
iii. Select a threshold using F1 score based on the cross validation set ('Xval' and 'yval') to flag outliers. Find the probabilities of 'Xval' and compare it with 'yval' for determining the threshold, where yval = 1 corresponds to an outlier and y = 0 corresponds to a normal data point. When p(x) $<$ threshold, it is considered as an outlier. Threshold is in range [4.5e-36, 0.09] with stepsize of 1000.

$$F_1 = \frac{2 \cdot prec \cdot rec}{prec + rec} \tag{7}$$

$$prec = \frac{tp}{tp + fp} \tag{8}$$

$$rec = \frac{tp}{tp + fn} \tag{9}$$

tp is the number of true positives: the ground truth label says it's an anomaly and our algorithm correctly classified it as an anomaly.
fp is the number of false positives: the ground truth label says it's not an anomaly, but our algorithm incorrectly classified it as an anomaly.
fn is the number of false negatives: the ground truth label says it's an anomaly, but our algorithm incorrectly classified it as not being anomalous.

For submission,

i. **[30 points]** Please submit your code for anomaly detection from scratch including the calculation of Guassian distribution, F1-score, precision and recall.
ii. **[10 points]** Please report the best threshold you found which could achieve the highest F1-score, the best F1-score and the number of outliers.