Assignment #2 — due Tuesday, Oct. 15, 2024, at 11:59PM

*Submit your homework on Canvas.

*No late homework will be accepted for credit.

*Append the codes you used to your submission.

*Using AI tool is allowed for coding problems, but needs to be declared in your submission.

Problem 1: Basic Knowledge *(200 points)*

1. Show that the log-likelihood function is concave in logistic regression.
2. Besides the gradient descent algorithm and Newton algorithm, state another algorithm which can be used to find MLE for logistic regression. Explain the details and its pros and cons compared to Newton algorithm.
3. In a logistic regression model, how would you interpret the coefficient β_i associated with a continuous predictor variable X_i?
4. Suppose you can access the LogisticRegression() function in python library. Write the pseudocodes to draw ROC curve of logistic regression on the training data $\mathcal{D}_{\mathsf{train}}$.
5. Explain the primary difference between LDA and Logistic Regression in the context of classification tasks.
6. Derive the discrimination function $\delta_k(x)$ in LDA using Bayes' Theorem.
7. Show that the expected pooled sample covariance matrix $\widehat{\Sigma} := \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}\big(X_i-\widehat{\mu}_k\big)\big(X_i-\widehat{\mu}_k\big)^\top$ in LDA is an unbiased estimator for the population covariance matrix Σ.
8. Derive the discrimination function $\delta_k(x)$ in QDA using Bayes' Theorem.
9. Explain how confusion matrix can be extended to multiclass classification problem? How about ROC?
10. Explain how the number of folds K affects the variance and bias of cross validation error.

Problem 2: Predicting Breast Cancer *(100 points)*

Dataset: BreastCancer_train.csv and BreastCancer_test.csv
Description: The objective is to identify each of a number of benign or malignant classes. Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself. Each variable except for the first was converted into 11 primitive numerical attributes with values ranging from 0 through 10. There are 16 missing attribute values. See cited below for more details. The meaning of these predictors are as follows.
ID: Sample code number
Cl.thickness: Clump Thickness
Cell.size: Uniformity of Cell Size
Cell.shape: Uniformity of Cell Shape

Marg.adhesion: Marginal Adhesion
Epith.c.size: Single Epithelial Cell Size
Bare.nuclei: Bare Nuclei
Bl.cromatin: Bland Chromatin
Normal.nucleoli: Normal Nucleoli
Mitoses: Mitoses
Class: Class

Goal: Learn a classifier based on the training dataset and test its performance on test dataset.

1. Use all the predictors to fit a logistic regression model and report the summary. Plot the ROC curve on the test dataset.
2. Use the predictors *Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin* to fit a logistic model and report the summary. Plot the ROC curve on the test dataset.
3. Use all the predictors to fit an LDA model and report the summary. Plot the ROC curve on the test dataset.
4. Use the predictors *Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin* to fit an LDA model and report the summary. Plot the ROC curve on the test dataset.
5. Use all the predictors to fit a QDA model and report the summary. Plot the ROC curve on the test dataset.
6. Compare all the above models by AUC.

Problem 3: Speed and Stopping Distances of Cars *(100 points)*

Dataset: Cars.csv
Description: The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s. The dataset has 50 observations, each observation with 2 variables.
speed: numerical value describing the speed of car
dist: numerical value describing the stopping distance

Goal: Use Cross-Validation to find the relation between stopping distance and speed.

1. Assume that dist is a polynomial function of speed. Use leave-one-out cross validation, and plot the CV errors *versus* degree of polynomial. Report your finding and conclusion.
2. Continue from Step 1: use 5-fold cross validation, and plot the CV errors *versus* degree of polynomial. Report your finding and conclusion.
3. Fit a non-parametric model by KNN with Gaussian kernel smoothing where the bandwidth h is the tuning parameter. Apply leave-one-out cross validation and 5-fold cross validation to choose the best bandwidth, and plot the CV errors *versus* bandwidth.
4. Compare the KNN-Gaussian kernel and polynomial regression, and report your findings.

Problem 4: Titanic – Survival or Not *(100 points)*

Dataset: titanic.csv
Description: The dataset contains information of passengers on the famous Titanic. It has 891 observations and each observation has 12 variables. The meaning of these variables are as follows.
survival: whether this passenger died or survived, 0=No, 1=Yes.
pclass: ticker class, 1=1st, 2=2nd, 3=3rd
sex: male or female
Age: age in years
sibsp: # of siblings/spouses aboard the Titanic
parch: # of parents/children aboard the Titanic
ticket: Ticket number
fare: Passenger fare
cabin: Cabin number
embarked: Port of Embarkation, C=Cherbourg, Q=Queenstown, S=Southampton

Goal: Apply Bootstrap for the statistical inference of how these variables affect the probability of survival.

1. Treat *survival* as response, fit a logistic regression model using predictors *pclass, sex, age, sibsp* and *fare*. Report the estimated coefficients for *Sex/male* and *pclass/3rd*, and report their 95% confidence intervals.

2. Now, apply Bootstrap with 1000 repetitions to obtain the 95% confidence intervals for the above coefficients. How do they compare with the reported confidence intervals from above.

3. Explore the dataset as you like and report some of your findings.

4. Keep the 1st observation as a test point and other observations as training. Train a logistic regression model and predict the probability that the test point will survive. Then, ust Bootstrap to construct a 95% prediction interval of the probability that the test point will survive.

5. Similar as Step 4, but now train a QDA and predict the probability that the test point will survive. Then, use Bootstrap to construct a 95% prediction interval of the probability that the test point will survive.