Midterm Project — due Wednesday, 25 Oct.

*Submit your homework as a report on Canvas. *Answer each question explicitly.*

*No late homework will be accepted for credit.

*Append the codes you used to your submission.

Problem 1: Speed and Stopping Distances of Cars

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 2: Titanic – Survival or Not

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.

Problem 3: Predicting First-Year College Students' GPA

Dataset: FirstYearGPA.csv
Description: The dataset contains information of 219 first-year college students. Each students has 10 variables including his/her GPA in the first year of college. The meaning of these variables are as follows.
GPA: First-year college GPA on a 0.0 to 4.0 scale.
HSGPA: High school GPA on a 0.0 to 4.0 scale
SATV: Verbal/critical reading SAT score
SATM: Math SAT score
Male: 1=male, 0=female
HU: Number of credit hours earned in humanities courses in high school
SS: Number of credit hours earned in social science courses in high school
FirstGen: 1= student is the first in her or his family to attend college; 0=otherwise
White: 1=white students; 0=others
CollegeBound: 1=attended a high school where $\geq 50\%$ students intended to go on to college; 0=otherwise

Goal: select the best linear model to predict first-year college students' GPA.

1. Use all variables to predict students' first-year GPA, by best subset selection up to the size of 8. Report the 8 best linear models (best model for each model size $k = 1, \cdots, 8$) and plot the R-square *versus* model size. Which is the best model using adjusted R-square?
2. Repeat Step 1 but choose the best model by 5-fold CV. Report the summary of that model.
3. Use all variables to predict students' first-year GPA, by forward stepwise selection up to the size of 8. Report the 8 (best model for each model size $k = 1, \cdots, 8$) best linear models and plot the adjusted R-square *versus* model size. Which is the best model using BIC?
4. Repeat Step 3 but choose the best model by 5-fold CV. Report the summary of that model.
5. A student is said in *good* position if his/her first-year GPA is not lower than 3. Fit a logistic regression model to predict whether a student will be in *good* position or not. Choose and report the best model by forward stepwise selection up to size of 8 and use 5-fold CV to choose the best model.
6. Similar as Step 5, but fit an LDA model to predict whether a student will be in *good* position or not. Choose and report the best model by forward stepwise selection up to size of 8 and use 5-fold CV to choose the best model.

Problem 4: Prediction of the Progression of Diabetes

Dataset: diabetes_train.csv and diabetes_test.csv
Description: The dataset contains information of 442 patients of diabetes. For each patient, the dataset provides some measure of 10 baseline variables. The meaning of these variables are as follows.
Y: a quantitive measure of the progression of diabetes.
age: age of the patient (note that the column is standardized)
sex: male or female (note that they are originally coded by number 1 and 2, then the column isstandardized)
bmi: body mass index (column is standardized)
map: some numerical measure of blood
tc: some numerical measure of blood
ldl: some numerical measure of blood
hdl: some numerical measure of blood
tch: some numerical measure of blood
ltg: some numerical measure of blood
glu: some numerical measure of blood
⋮
age.tch: the interaction of baseline variables *age* and *tch*
⋮

Goal: apply regularized regression to find which baseline variables or their interactions affect the progression of diabetes.

1. Use the train dataset to fit LASSO estimators with regularization parameter λ chosen from the grid $10^{seq(4,-2,length=100)}$. Plot the coefficients *versus* the ℓ_1 norm. Report any findings you think interesting.

2. Use train dataset to fit LASSO and apply 10-fold cross validation. Plot the CV error *versus* the values of lambda. What is the best lambda value according to CV error? Report the linear model using the best lambda value, how many variables are included in the model, what are they?

3. Use the best model in Step 2 to predict the progression of diabetes on the test dataset, and report the mean test error.

4. For the best model reported in Step 2, how do you construct 95% confidence intervals for the estimated coefficients?

5. LASSO-type method can also be applied for classification problems such as ℓ_1-norm penalized logistic regression:
$$\arg\min_{\beta} - \text{log-likelihood} + \lambda\|\beta\|_{\ell_1},$$
called the LASSO-logistic. A patient is said in *stable* condition if the Y variable is smaller than 150. Fit the LASSO-logistic on the training dataset and apply 10 -fold cross validation. Plot the CV error *versus* the values of lambda. What is the best lambda value according to CV error? Report the best model, how many variables are included in the model, what are they?