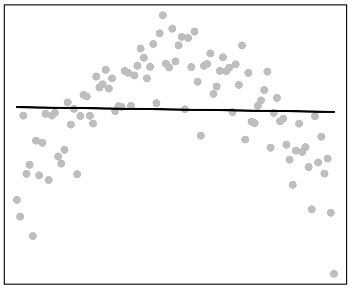
BUS 41204 Machine Learning

Midterm

- **This is an INDIVIDUAL exam. You cannot work in groups**.
- The exam must be submitted on gradescope before 11.59pm on Sunday, February 12. You should submit a pdf document.
- Ask coding questions on Piazza. Do not reveal answers when formulating questions.
- When answering questions, provide plots and supporting analysis. Label plots and axes.
- Be concise.
- Provide supporting R code to help us understand your answers.

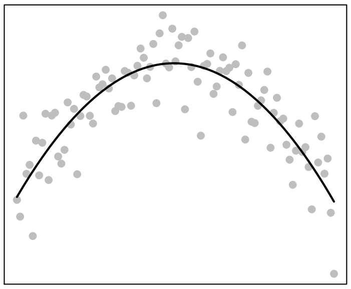
1 True or False [20 points]

1. As one increases k, the number of nearest neighbor, in a kNN classifier,
   (a) the bias of the classifier will increase;
   (b) the variance of the classifier will increase;
   (c) the misclassification rate on the training dataset will increase;
   (d) the misclassification rate on a test dataset will increase.
2. Consider the three line regression fits to the gray points plotted below.



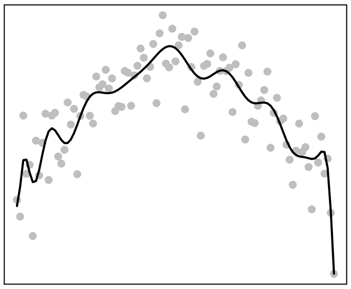
$\hat{y} = \beta_0 + \beta_1 x$

(1)



$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$

(2)



$\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_{20} x^{20}$

(3)

   (a) The estimate in (2) has a higher variance than the estimate in (1).
   (b) The estimate in (2) has a higher bias than the estimate in (3).
   (c) The estimate in (3) has the smallest training error.
   (d) The estimate in (1) has the smallest test error.
3. Misclassification rate of a classifier evaluated on a validation set will never be smaller than the one evaluated on the training set that is used to build the classifier.
4. k-fold cross-validation provides an unbiased estimate of the predictive error of the models.

2 Question [40 points]

Crimes in Philly. The dataset `PhillyCrime.csv` contains information of two types of crime incidents (`Vandalism` and `Thefts`) occurred in Philadelphia in 2016.[1] In this problem, we will use k nearest neighbor to predict the category of the crimes.

1. Make two separate scatterplots of the crime incidents using their latitudes (`X`) and longitudes (`Y`), one for `Vandalism` and one for `Thefts`. Describe any pattern or difference you find from the scatterplots.
   (*You might want to decrease the size of the points so that they don't overlap too much.*)
2. Split the data, with 50% into a training set and the other 50% reserved for a validation set. Fit a set of k-nearest neighbors classifiers for $k = 1, 2, \ldots, 100$ on the training set to predict the crime category using latitude (`X`) and longitude (`Y`) as predictors.
   (a) Make a scatterplot of the misclassification rate on the validation set against the parameter k.
   (b) Report the optimal k selected by the validation-set approach and the minimum misclassification rate on the validation set.
   (c) Make a single scatterplot of the crime incidents using their latitudes (`X`) and longitudes (`Y`). Color the points according to their *predicted* class given by the optimal k-nearest neighbors selected above.
3. Repeat (2) for 20 times (with 20 random splits of the dataset).
   (a) Report the 20 optimal k's selected by the validation sets.
   (b) Report the average of the minimum out-of-sample misclassification rates as well as its standard error.
4. Repeat (2) and (3), this time with a 90/10 split of the data. Make sure you include in your solutions
   (a) (For a particular split,) a scatterplot of the misclassification rate on the validation set against the parameter k.
   (b) (For a particular split,) a scatterplot of the crime incidents (`Y` vs `X`) with points colored according to predicted class given by the optimal k.
   (c) 20 optimal k's selected by the validation sets.
   (d) Average minimum out-of-sample misclassification rates and its standard error.
5. Comment on the difference between the results obtained in (2), (3) and (4).
6. For a 25-nearest neighbors classifier trained on 50% of the data in `PhillyCrime.csv`, plot the ROC curve calculated on a validation set with the other 50% of the data.

[1] The whole dataset is available at https://www.opendataphilly.org/dataset/crime-incidents

3 Question [30 points]

A university has extensive dataset on its alumni, including past studies, demographic information by zip code, and past donations. The university is planning to send a deluxe brochure and a donation request to some of the alumni. The university has hired you to help with creating a targeting model under the assumptions:

- Donation amount may vary.
- Alumni may spontaneously make a donation (even when not targeted).
- Targeting cost is fixed $15 per individual.
- Other than the targeting cost, there are no additional costs for alumni who are targeted and decide not to donate.

You have been asked to build several data mining models that would suggest which alumni should be targeted. Use the expected value framework to determine which models should be used to address the problem.

Note: It is sufficient to write down the correct expected value equations to identify the models that should be constructed. There is no need to further solve/develop the equations.

4 Question [60 points]

Use Tayko.csv dataset for this question.

4.1 Background

Tayko is a software catalog firm that sells games and educational software. It started out as a software manufacturer, and added third party titles to its offerings. It has recently put together a revised collection of items in a new catalog, which it is preparing to roll out in a mailing.

In addition to its own software titles, Tayko's customer list is a key asset. In an attempt to expand its customer base, it has recently joined a consortium of catalog firms that specialize in computer and software products.

The consortium affords members the opportunity to mail catalogs to names drawn from a pooled list of customers. Members supply their own customer lists to the pool, and can "withdraw" an equivalent number of names each quarter. Members are allowed to do predictive modeling on the records in the pool so they can do a better job of selecting names from the pool.

4.2 The Mailing Experiment

Tayko has supplied its customer list of 200,000 names to the pool, which totals over 5,000,000 names, so it is now entitled to draw 200,000 names for a mailing. Tayko would like to select the names that have the best chance of performing well, so it conducts a test - it draws 20,000 names from the pool and does a test mailing of the new catalog to them.

This mailing yielded 1065 purchasers - a response rate of 0.053. Average spending was $103 for each of the purchasers, or $5.46 per catalog mailed. To optimize the performance of the data mining techniques, it was decided to work with a stratified sample that contained equal numbers of purchasers and non-purchasers. For ease of presentation, the dataset for this case includes just 1000 purchasers and 1000 non-purchasers, an apparent response rate of 0.5. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case's "probability of purchase" by 0.053/0.5 or 0.107.

4.3 Data

There are two response variables in this case. "Purchase" indicates whether or not a prospect responded to the test mailing and purchased something. "Spending" indicates, for those who made a purchase, how much they spent. The overall procedure in this case will be to develop two models. One will be used to classify records as "purchase" or "no purchase." The second will be used for those cases that are classified as "purchase," and will predict the amount they will spend.

Table below provides a description of the variables available in this case. A partition variable is used because we will be developing two different models in this case and want to preserve the same partition structure for assessing each model.

- `US`: Is it a US address? binary 1: yes 0: no
- `Source_*`: Source catalog for the record, binary 1: yes 0: no (15 possible sources)
- `Freq`: Number of transactions in last year at source catalog,numeric
- `last_update_days_ago`: How many days ago was last update to cust. record, numeric
- `first_update_days_ago`: How many days ago was 1st update to cust. record, numeric

- `Web_order`: Customer placed at least 1 order via web, binary 1: yes 0: no
- `Gender_is_male`: Customer is male, binary 1: yes 0: no
- `Address_is_res`: Address is a residence, binary 1: yes 0: no
- `Purchase`: Person made purchase in test mailing, binary 1: yes 0: no
- `Spending`: Amount spent by customer in test mailing ($), numeric

- `Partition`: Variable indicating which partition the record will be assigned to, categorical t:training v:validation s:test

4.4 Assignment

1. Each catalog costs approximately $2 to mail (including printing, postage and mailing costs). Estimate the gross profit that the firm could expect from the remaining 180,000 names if it randomly selected them from the pool.

2. Develop a model for classification a customer as a purchaser or non-purchaser. Partition the data into training on the basis of the partition variable, which has 800 "t's," 700 "v's" and 500 "s's" (training data, validation data and test data, respectively) randomly assigned to cases. Choose one model on the basis of its performance on the validation data.

3. Plot an ROC curve for the chosen model in the previous step on the test data.

4. Develop a model for predicting spending among the purchasers. Again, partition the data using the partition variable as before. Choose one model on the basis of its performance on the validation data.

5. For every case in the test data, compute the expected spending amount. This is obtained by multiplying predicted spending by adjusted probability of purchase and adjusted probability of purchase is obtained by multiplying the probability outputed by your model by 0.107 to adjust for oversampling the purchasers.

   $$\text{Expected spending}(x) = \text{predicted\_spending}(x) \cdot p(\text{purchase} \mid x) \cdot 0.107.$$

   Order cases according to the expected spending and plot cummulative actual spending divided by the avearage spending that would result from random selection.

6. **BONUS QUESTION**: Using this cumulative lift curve, estimate the gross profit that would result from mailing to the 180,000 on the basis of your data mining models.