Problem Set

The assignment is worth 10% of your final grade.

Why?

Now it's time to check your understanding in machine learning, theory and Bayesian statistics.

The Problems Given to You

Question 1 [40pt]:

You are collaborating with a fellow Data Scientist on a business-critical wireless account retention project. The project is concerned with modelling contract terminations (i.e., churn) to help maintain and grow the revenue generating base. The goals for the model are to predict wireless churn, and take action, such as special offers, to keep accounts.

On your first day our colleague provides you with the following materials that they prepared ahead of time:

- Datasets to use for developing the model (Link)
- A brief overview on the project's objectives (under the heading **Project Objectives** below).

After some Q&A with your colleague, you get some added details about the materials your colleague provided. This detail can be found below, in the section **Notes on the Materials Provided**.

Notes on the Materials Provided

The datasets provided are anonymized product, billing and demographic data for a set of wireless accounts. The data contains one account identifier (Customer_ID).

Three datasets

- wls_churn_master_target_t1.csv,
- wls_customer_demographics_t1.csv, and
- wls_billing.csv

were extracted from our team's cloud database, which means that this data was initially prepared into features by our team's feature creation pipeline. The master table (wls_churn_master_target_t1.csv) contains the target field ('churn'). All other tables can be joined to the master table before modeling.

Your colleague has provided **NO** data dictionary, but recommends conducting some exploratory data analysis, and selecting useful features with filters (i.e., use a metric like correlation/chi-square, and based on that filter features). Lastly, your colleague points out that there is no separate test dataset, and that you will need to split the sample data into training and test data.

Project Objectives

As above, the aim of this project is to predict if a wireless account will churn or not. The business would like you to train and assess machine learning models using the provided sample data. As a secondary objective, the business would also like to understand which features are driving churn.

Assessment Questions

1. [10 pts] Review the datasets provided by your colleague. What feedback would you give to your colleague to help them improve any future modelling work?
2. [10 pts] Using the data provided, construct an MVP ML code solution (i.e., the model with the best performance, and the data processing/engineering steps required to build it). Prioritize the most important solution elements. Explain the reasoning behind decisions to implement or not implement.
3. [10 pts] Which features are driving churn?
4. [10 pts] Include your Jupyter Notebook with comprehensive analysis in markdown format.

Question 2 [10pt].

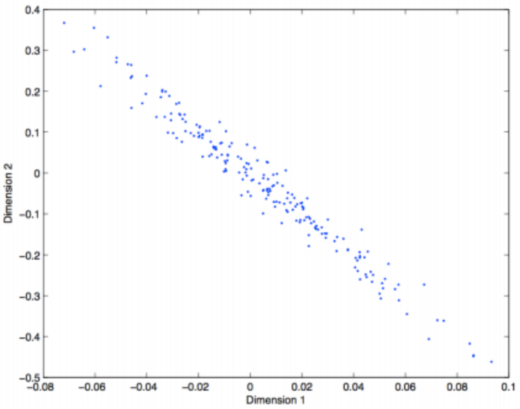
Imagine you had a learning problem with an instance space of points on the plane and a target function that you knew took the form of a line on the plane where all points on one side of the line are positive and all those on the other are negative. If you were constrained to only use decision tree or nearest-neighbor learning, which would you use? Why?

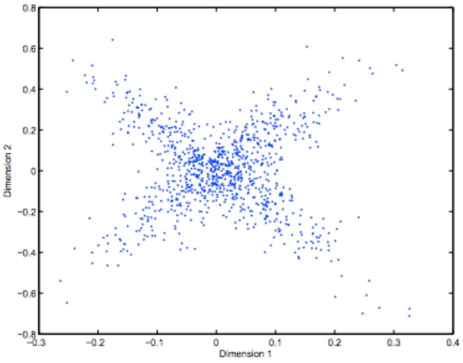
Question 3 [10pt].

1. Compare L1 and L2 regularization. Why does L1 regularization result in sparse models? [4pt]
2. Describe the trade-off between bias and variance using two examples. [3pt]
3. Describe the curse of dimensionality with at least one example. [3pt]

Question 4 [10pt].

Plot the direction of the first and second PCA components in the figures given



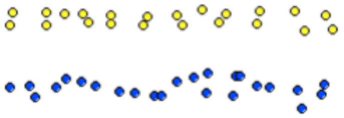


Question 5 [10pt].

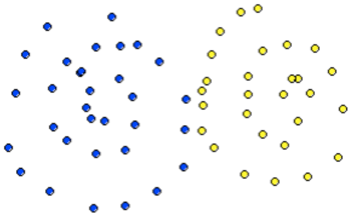
Which clustering method(s) is most likely to produce the following results at k = 2? Choose the most likely method(s) and briefly explain why it/they will work better where others will not in at most 3 sentences. Here are the five clustering methods you can choose from:

1. Hierarchical clustering with single link
2. Hierarchical clustering with complete link
3. Hierarchical clustering with average link
4. Kmeans
5. EM

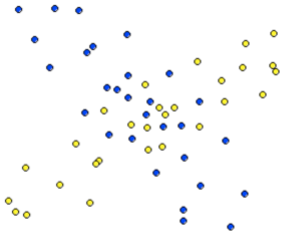
a.



b .



c.



Question 6 [10pt].

Explain how you can use Decision Trees to perform regression? Show that when the error function is squared error, then the expected value at any leaf is the mean. Take the Boston Housing dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/) and use Decision Trees to perform regression.

Question 7 [10pt].

Design a two-input perceptron that implements the boolean function $A \wedge \neg B$. Design a two-layer network of perceptrons that implements $A \oplus B$ ($\oplus$ is XOR).

What to Turn In

You must submit a tar or zip file named *firstname_lastname_NUID*.{zip,tar,tar.gz} that contains a single folder or directory named *firstname_lastname_NUID* that in turn contains: -->

1. A file named *README.txt* that contains instructions for running your code
2. Your code
3. A file named firstname_lastname_NUID-*classifier.pdf* that contains your writeup.
4. Another file named firstname_lastname_NUID-*solutions.pdf* that answers questions 2 to 7.
5. Any supporting files you need (for example, your datasets).