CompSci 526: Intro To Data Science **Fall 2025**

Homework 1

Due: Tuesday 9/23 at 11:59 pm ET (submit via Gradescope)

Submission: Please submit your responses in a single PDF file. Ensure that all written content is typed; hand-drawn figures are acceptable. Code snippets or outputs should also be included within the PDF. When submitting on Gradescope, you will be prompted to label pages for each question. Make sure to complete this labeling accurately to expedite the grading process.

Overview: This homework has 4 questions, for a total of 35 points.

Policy: Collaboration is encouraged, but each student must submit their own individual work. If you collaborate with others, please acknowledge your collaborators by explicitly listing their names below.

Names of your collaborators:

1. Let us consider the Pima Indians Diabetes Database, which can be downloaded from Canvas as a CSV file. We will practice data quality assessment and data cleaning with this dataset. It originally comes from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes based on certain diagnostic measurements included in the dataset.

   Several constraints were applied to the selection of these instances from a larger database. All patients are females at least 21 years old of Pima Indian heritage. The dataset consists of several medical predictor variables (also known as attributes) and one target variable, `Outcome`. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

   For the attributes `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`, a value of 0 is used as an indicator of missing data. We refer to these attributes as *the attributes with missing values.*

   When working with a dataset that contains missing values, an important first step is to understand how the missing values are distributed. This can be achieved by collecting statistics about the missing values. You will conduct this exercise by answering the following questions.

   (a) (4 points) We will investigate how missing values are distributed across records. Calculate the number of missing values in each record for the attributes with missing values. Summarize your findings by plotting a figure where the x-axis represents the number of missing values in a record and the y-axis represents the number of records with that many missing values.

   (b) (4 points) Since the dataset contains data from two classes indicated by the target variable `Outcome`, it is interesting to observe how the missing values are distributed across these two classes. To report the results, please draw two figures: one for each class. Use the same configuration as in the previous question, i.e., the x-axis represents the number of missing values in a record, and the y-axis represents the number of records with that many missing values, for each class separately.

   (c) (4 points) Compute the conditional probability $P(X = 0 \mid Y = 0)$, where X and Y are attributes with missing values. To report your results, create a table where:

   - Each row and each column represent one of the attributes with missing values.
   - The entry in the table at position $[X, Y]$ shows the conditional probability $P(X = 0 \mid Y = 0)$.

   This table will provide the conditional probability of one attribute having a missing value given that another attribute has a missing value.

   (d) (0 points) **Food for Thought:** Consider whether the attributes with missing values are independent of each other. How might this information affect your understanding of the dataset?

2. For each of the following four scenarios, determine whether a flexible statistical learning method is expected to perform better or worse than an inflexible method. Provide a rationale for your answer.

   (a) (2 points) When the sample size is extremely large and the number of predictors is small.

   (b) (2 points) When the number of predictors is extremely large but the number of observations is small.

   (c) (2 points) In cases where the relationship between the predictors and the response is highly non-linear.

   (d) (2 points) When the error terms in the data have very high variance (i.e., there is a lot of noise in the data).

3. (3 points) Explain the practical applications of regression in two real-life scenarios. In one scenario, emphasize the primary objective of inference, while in the other, focus on prediction. For each case, clearly elucidate the response variable and the predictor variables, and provide a detailed rationale for why each application is categorized as either inference or prediction.

4. (12 points) You have the dataset below of customer profiles and purchases. Using Naive Bayes, predict whether a **new customer** will **make a purchase** or **not make a purchase** given their profile. Be sure to compute results for **both classes** (`Purchase` = Yes and `Purchase` = No).

   The dataset is as follows:

`Age Group`	`Income Level`	`Gender`	`Previous Purchases`	`Purchase`
Young	High	Female	Yes	Yes
Middle-aged	Medium	Male	No	No
Senior	Low	Female	Yes	No
Young	Medium	Male	Yes	Yes
Middle-aged	High	Female	Yes	Yes
Senior	Medium	Male	No	No
Young	High	Female	No	Yes
Middle-aged	Low	Female	No	No
Senior	High	Male	Yes	Yes
Young	Medium	Male	No	Yes

   The new customer has profile:

   - `Age Group`: Young
   - `Income Level`: High
   - `Gender`: Female
   - `Previous Purchases`: Yes

   For this case, compute the priors, the class-conditional likelihoods, and the *unnormalized* posterior scores for both classes, then state the final prediction. If any count is zero, apply Laplace smoothing ($k = 1$) and indicate that you used it.