NORTHEASTERN UNIVERSITY, KHOURY COLLEGE OF COMPUTER SCIENCE

CS 6220 Data Mining — Assignment 3
Due: February 15, 2023(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

Principle Components Analysis

Italy is home to over 2000 grape varieties. Even within a single region, wines exhibit distinct attributes from different cultivators that can be measured with objective and numerical features. Notably, in the dataset we are exploring today, there are thirteen different measurements taken for different constituents found in the three types of wine. We would like to visualize how well-separated the data is for the different wineries in a 2D scatter plot.

We will be using the UCI Wine's dataset. Please review `sklearn`'s description of wine data, and load it in with the following code:

```
from sklearn.datasets import load_wine
wine = load_wine()
```

Question 1 [25 pts]

Preprocess the the data with **z-score normalization** and scatter the data that's been projected onto the first two principle components with different colors for each target/class of wine. Include your code (linked or inline).

The below scatter plot is an example of displaying multiple classes with different colors on a single plot.

PCA with Raw Unscaled Penguins Data
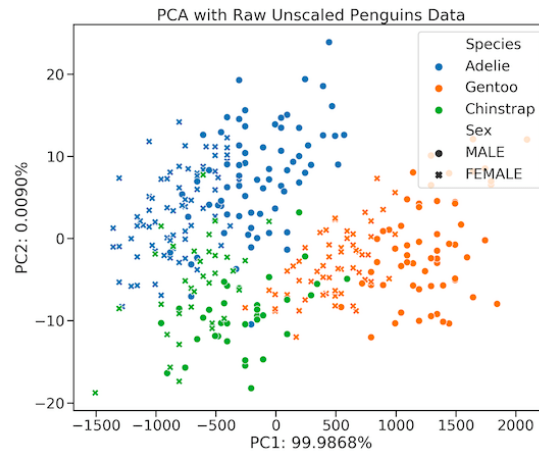
Parameter Estimation

It is well-known that light bulbs commonly go out according to a Poisson distribution, and are independent regardless of whether or not they're made in the same factory. The Poisson distribution has the form:

$$p(X|\lambda) = \frac{\exp^{-\lambda} \lambda^{x_i}}{x_i!}$$

An architect has outfitted a building with 32,000 of the same lightbulb. The factory has provided him with data on when N of these lightbulbs have gone out over their lifetimes. They've been measured with $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$

Question 2 [25 pts]

Derive the maximum likelihood estimate of the parameter λ in terms of x_i. Please show your work.

Submission Instructions

When you have finished, follow the instructions on the homework main page. Commit your code, outputs, and PDF writeup to your repository and provide the repository link to Gradescope.