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

Question 1 [50 pts]

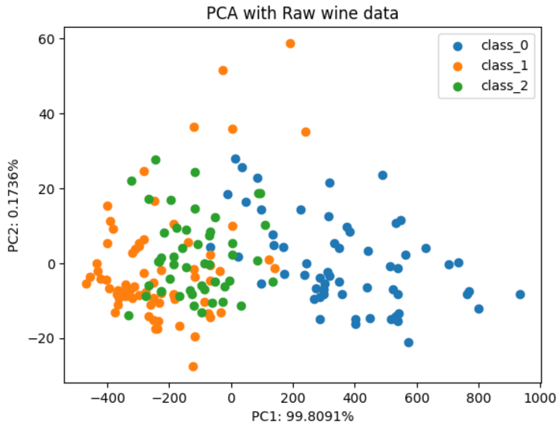
Preprocess the the data with **z-score normalization** and scatter the data that's been projected onto the first two principle components with different colors for each target/class of wine. Include your code (linked or inline).

```
import numpy as np
from sklearn.datasets import load_wine
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

wine = load_wine()
x = wine.data
y = wine.target
scale = StandardScaler()
x_normalized = scale.fit_transform(x)
pca = PCA(n_components=2)
pca.fit(x)
x_pca = pca.transform(x)
ratio = pca.explained_variance_ratio_

for item in np.unique(y):
    idx = (y==item)
    plt.scatter(x_pca[idx,0],x_pca[idx,1],label=f'class_{item}')
```

```
plt.title('PCA with Raw wine data')
plt.xlabel(f'PC1: {np.round(100*ratio[0],4)}%')
plt.ylabel(f'PC2: {np.round(100*ratio[1],4)}%')
plt.legend()
plt.show()
```



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 [50 pts]

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

$$\begin{aligned} f(\mathcal{D}) &= \prod_{i=1}^{N} p(x_i|\lambda) \\ &= \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \end{aligned}$$

Maximize $f(\mathcal{D})$ is equivalent to maximizing $\log f(\mathcal{D})$.

$$\begin{aligned}
\log f(\mathcal{D}) &= \sum_{i=1}^{N} \log p(x_i|\lambda) \\
&= \sum_{i=1}^{N} (-\lambda + x_i \log \lambda - \log(x_i!)) \\
&= -N\lambda + (\log \lambda) \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \log(x_i!)
\end{aligned}$$

Then we obtain derivative of $\log f(\mathcal{D})$ as:

$$\frac{\partial \log f(\mathcal{D})}{\partial \lambda} = -N + \frac{\sum_{i=1}^{N} x_i}{\lambda}$$

Let the derivative to be 0. We can obtain the estimated value of parameter λ.

$$\lambda^* = \frac{\sum_{i=1}^{N} x_i}{N}$$