6220 Homework3 Writeup

Overview

There are two questions in this assignment. Both of them are basic data mining techniques in industry.

Question 1 focuses on calculating the first two principal components with z-score normalization and hope we can scatter the data for a given data set. The expectation plot should be a scattered dot in one picture with different colors.

Question 2 focuses on the Poisson distribution rule and hope we can derive the maximum likelihood estimate of the parameter lambda (λ) which is the mean number of events within a given interval of time or space. Also this question can let us understand how to draw method functions with the maximum likelihood in the data mining area.

Algorithm Explanation

Question1:

Step1

Load data and use zero-score normalization to format the data and create a dataframe.

Step2

Check the necessary library and use PCA after z-score, the key part is df4 = pca2.fit_transform(df2)

Step3

Add the feature for the plot part, such as color. Then plot the dotted picture, compare after z-score normalization with before zero score normalization.

Question2:

Step1

Setting up the environment and importing the necessary library.

Step2

Generate 1000 data sizes to randomize an array in order to produce random data points.

Step3

Calculate pois with the lambda and size parameter. Then computing PMF, I need to understand the necessary library which provides the function for this computing.

Step4

Define a method called likelihood to calculate sum and adjust lambda to plot PMF.

Algorithm Diagram

Diagram 3.1 for question1

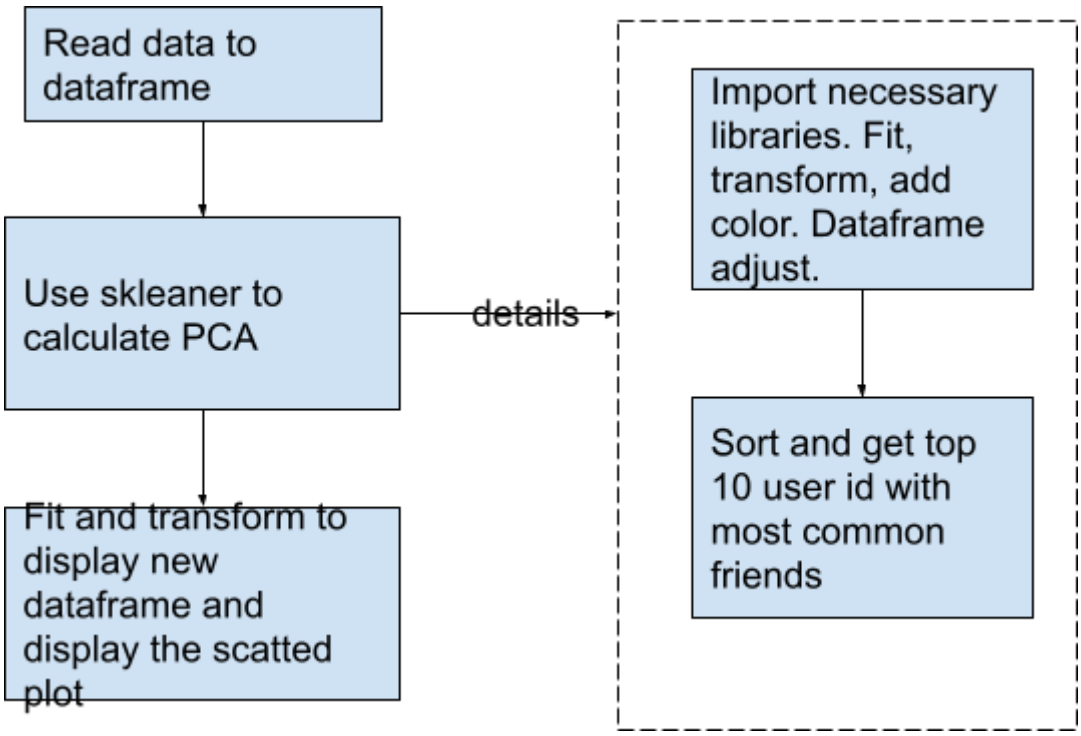
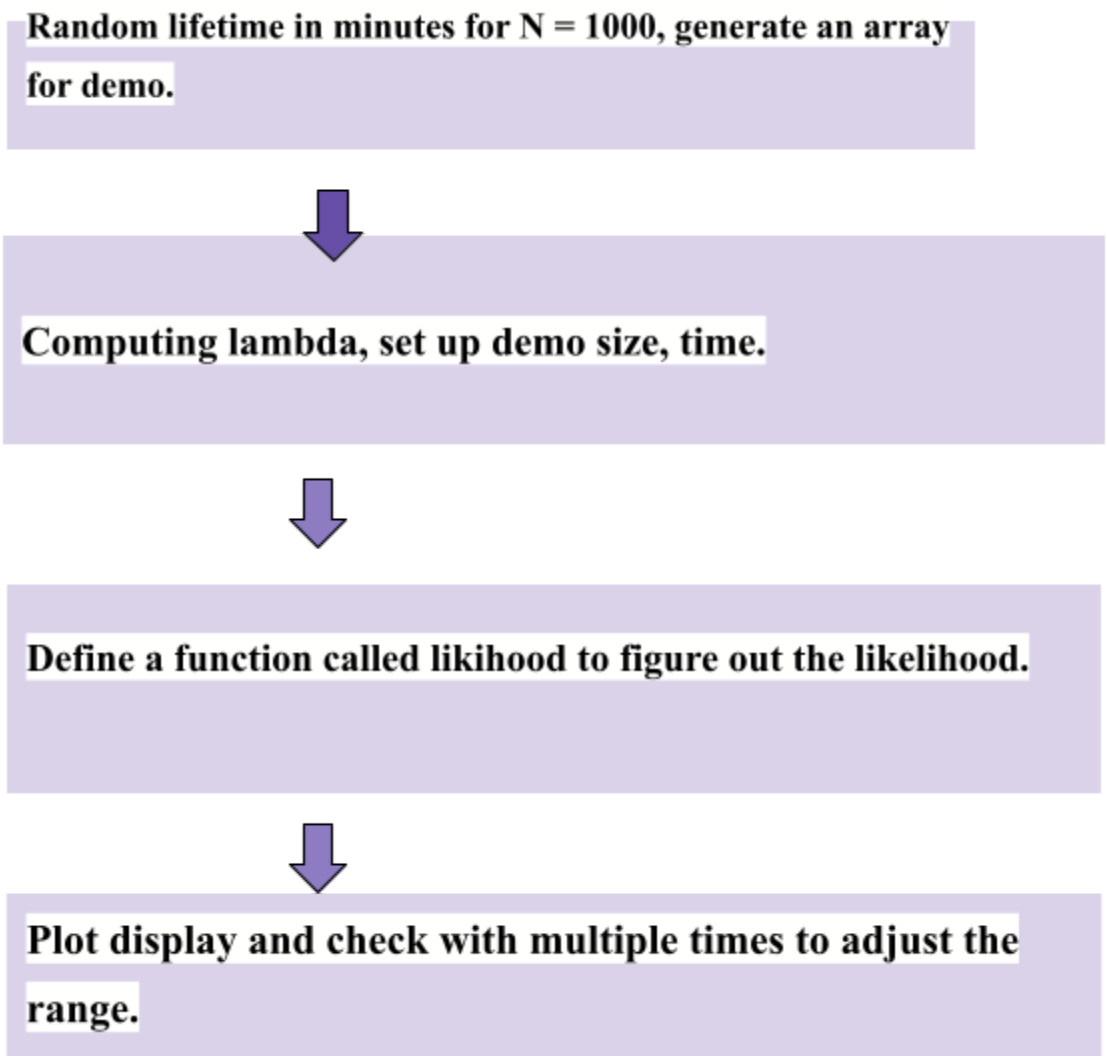


Diagram 3.2 for question 2



Time Complexity

Question 1:

We use some library methods to find the first two principles in this approach. The time complexity can be treated as a linear algorithm.

Question 2:

We evaluate the probability mass function (PMF) for each possible value of the random variable to compute the Poisson distribution. So far, we know the formula is:

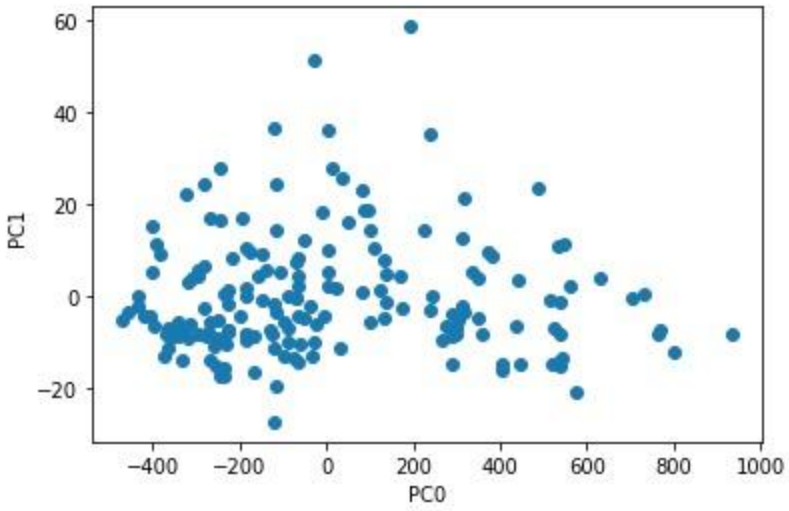
P(X = k) = (lambda^k / k!) * e^(-lambda)

There is an important parameter lambda which is the mean of the Poisson distribution and k is the number of occurrences.

The time complexity of computing the PMF of the Poisson distribution using this method is O(n), where n is the maximum number of occurrences that we want to compute. Because we need to compute the factorials and exponential function for each value of k.

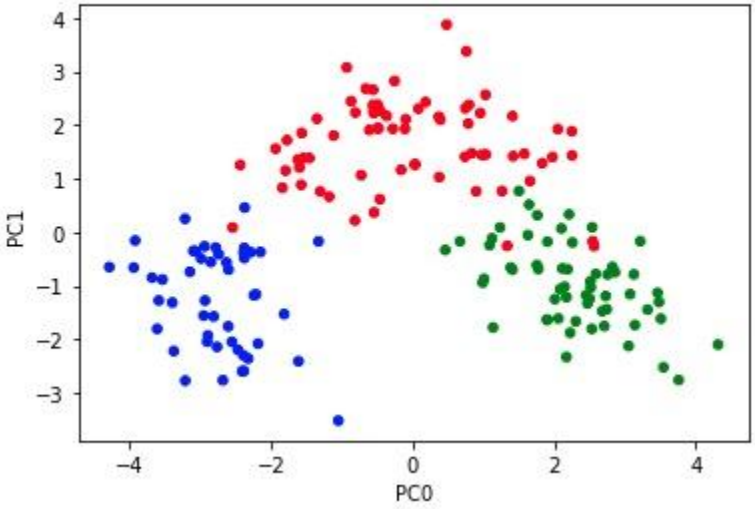
Demo Output

```
plot.scatter(df3[:,0], df3[:, 1])
plot.xlabel('PC0')
plot.ylabel('PC1')
plot.show()
```



178 rows × 4 columns

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc0b0c4e370>
```



```
mle_arr = make_array()
for lam in lambs:
  #print(lam)
  mle_arr = np.append(mle_arr, likihood(lam))

# plot the PMF for each possible lambda value depend on the simulated data
plots.plot(lambs, mle_arr)

# from the plot of the curv, looks the PMF lambda value is little bit different than the
# than the mean value 50 and maybe due to the quality of simulated data
```

```
[<matplotlib.lines.Line2D at 0x7f10e40936d0>]
```