Vincent Purcell

Professor Cohen

ECE-487

19 November 2019

Homework 8

# Problem 8.1 on Pg. 513

**8.1) Find the Edit distance between the word "poem" and its misspelled version "poten." Draw the optimal path.**

Using the following the following rules I was able to find the edit distance to be '2' for the editing of "poten" to "poem". The rules were based on Levenshtein's method for edit distance algorithm:

- Inserting, and Deleting operations cost 1
- Substitution cost 2

I was able to calculate an edit distance matrix using those rules which allowed for the optimal path to be found and plotted.

$$D(i,j) = min\begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+\begin{cases} 2; if\ S_1(i) \neq S_2(j) \\ 0; if\ S_1(i) = S_2(j) \end{cases} \end{cases}$$

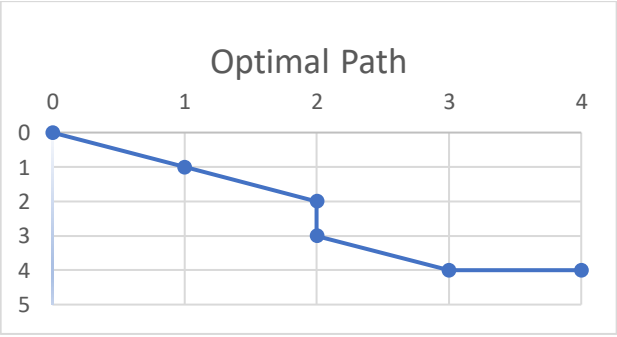
**Edit Distance Cost Matrix**

| | # | P | O | E | M |
|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 |
| P | 1 | 0 | 1 | 2 | 3 |
| O | 2 | 1 | 0 | 1 | 2 |
| T | 3 | 2 | 1 | 2 | 3 |
| E | 4 | 3 | 2 | 1 | 2 |
| N | 5 | 4 | 3 | 2 | 3 |

After that Edit Distance Cost Matrix was found it was quite easy to use 'backtracing' in order to find the optimal path and then simply plot that path.

**Optimal Path Decision Matrix**

| | # | P | O | E | M |
|---|---|---|---|---|---|
| # | 0 | ←1 | ←2 | ←3 | ←4 |
| P | ↑1 | ↖0 | ←↖1 | ←↖2 | ←↖3 |
| O | ↑2 | ↑↖1 | ↑←↖0 | ←↖1 | ←↖2 |
| T | ↑3 | ↑↖2 | ↑←↖1 | ↑←↖2 | ↑←↖3 |
| E | ↑4 | ↑↖3 | ↑↖2 | ↑←↖1 | ←↖2 |
| N | ↑5 | ↑↖4 | ↑↖3 | ↑↖2 | ↑←↖3 |



# *Estimating the Number of Clusters in a Dataset via the Gap Statistic*

**What are the two choices for reference distribution?**

The reference distribution can be defined by either of the two Theorems listed in the paper, "Estimating the Number of Clusters in a Dataset via the Gap Statistic." The first theorem says that if p is equal to 1 (p is the amount of features), then for all k greater than or equal to 1:

$$\inf_{X \in S^p} \left\{ \frac{\mathrm{MSE}_X(k)}{\mathrm{MSE}_X(1)} \right\} = \frac{\mathrm{MSE}_U(k)}{\mathrm{MSE}_U(1)}.$$

The above equation states that "among all unimodal distributions, the uniform distribution is the most likely to produce spurious clusters by the gap test." (R.T.,G.W.,T.H)

The second theorem says that if p is greater than 1 then no distribution $U \in S^p$ can satisfy the above equation unless its support is degenerate to a subset of a line.

There are two types of reference distributions, Empirical and Theoretical. Empirical reference distributions are observed, real data sets that are significantly large. A theoretical reference distribution allows for particular significant values within statistics to be evaluated. This theoretical distribution in many cases can be defined by a normal distribution.

**Testing the Gap Statistic**

Below attached is published MATLAB code where I used the 'evalclusters' function and evaluated its effectiveness. I ran 4 tests where I varied the amount of gaussian distributions within the data. Some distributions were clearly farther away from each other which led to easily clustered data. Some distributions were placed close to each other which led to some distributions being clustered together or broken up in between clusters. It seems that the gap statistic will fail if the clusters have varying densities and they aren't clear clusters and have significant distance between their means. For example, in the fifth figure there are three gaussian distributions but two of them have small densities, so they are clustered with the larger distribution.

There is also a point where the distributions are simply too close, and they fail to differentiate them. I was running this test with distributions of variance between 0.5 and 1. The clusters begin to be clustered together when their means are within 1 to 2 Euclidean distances of each other. That can be seen in figure 3 where there are 4 distributions but only 2 clusters.

# References


- Minimum Edit distance
  - https://web.stanford.edu/class/cs124/lec/med.pdf
- Estimating the Number of Clusters in a Dataset via the Gap Statistic (R.T.,G.W.,T.H)
  - https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00293
- GapEvalution MATLAB Documentation
  - https://www.mathworks.com/help/stats/clustering.evaluation.gapevaluation-class.html
- Reference Distributions
  - http://geog.uoregon.edu/bartlein/courses/geog495/lec09.html#empirical-reference-distributions

## Contents



## Vincent Purcell - HW 8 - ECE487

```
clear; clc; close all;
```

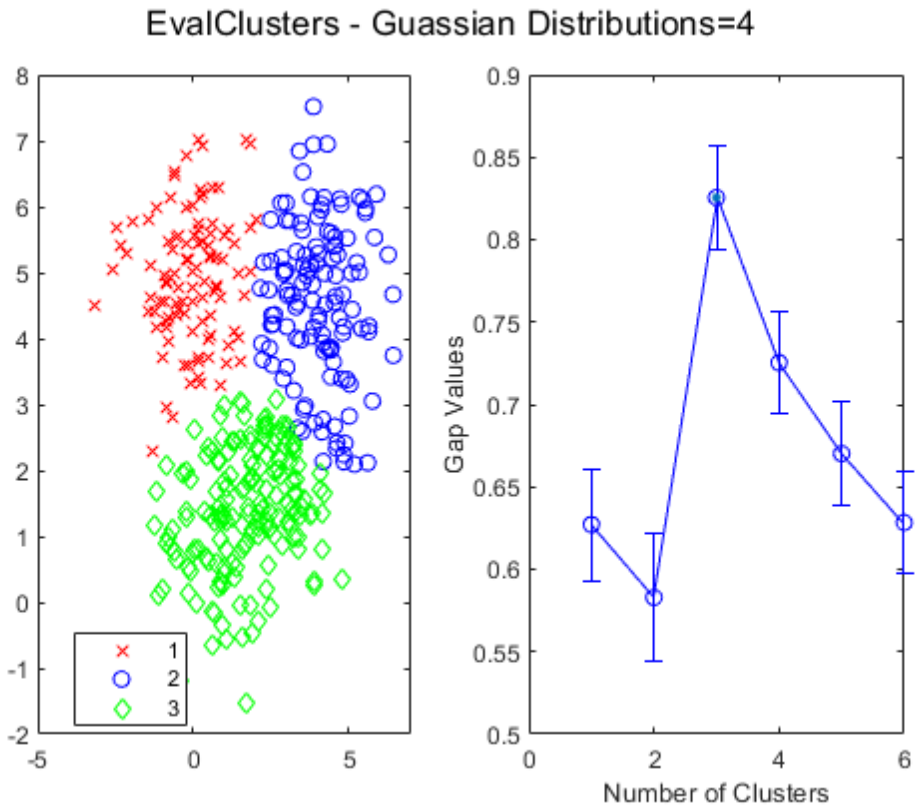
## EVAL Clusters

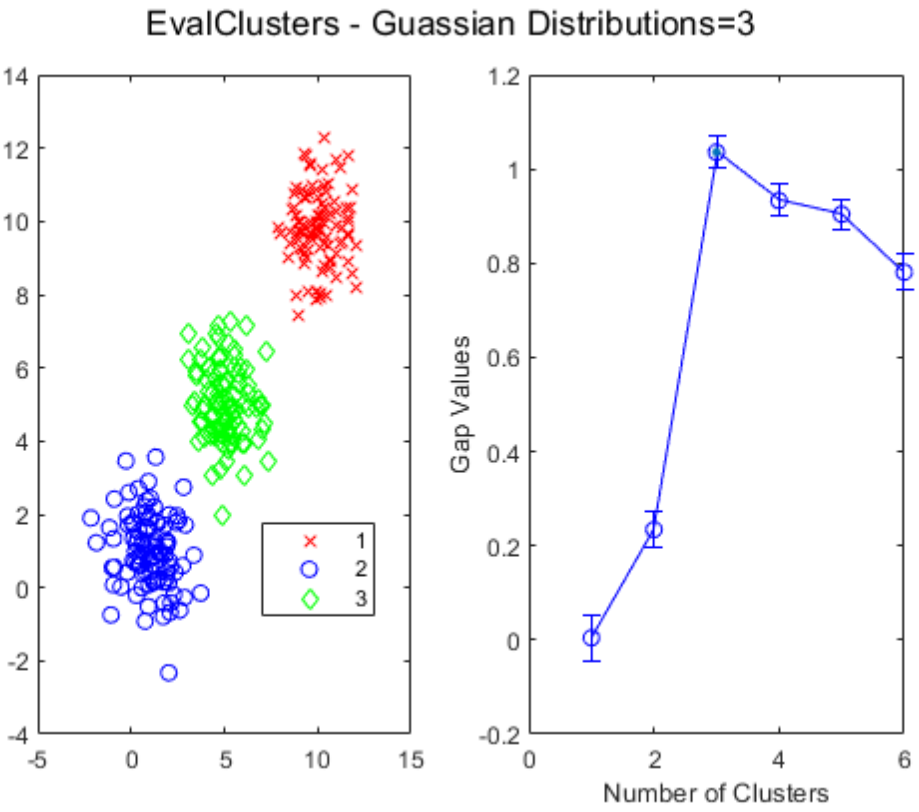
```
X = [mvnrnd([1 1], 1.*eye(2), 100); ...
     mvnrnd([3 2], 1.*eye(2), 100); ...
     mvnrnd([4 5], 1.*eye(2), 100); ...
     mvnrnd([0 5], 1.*eye(2), 100)];
evalFunc(X,4)

X = [mvnrnd([1 1], 1.*eye(2), 100); ...
     mvnrnd([5 5], 1.*eye(2), 100); ...
     mvnrnd([10 10], 1.*eye(2), 100)];
evalFunc(X,3)

X = [mvnrnd([2 6], 1.*eye(2), 100); ...
     mvnrnd([1 6], 1.*eye(2), 100); ...
     mvnrnd([1 0], 1.*eye(2), 100); ...
     mvnrnd([3 0], 1.*eye(2), 100)];
evalFunc(X,4)

X = [mvnrnd([0 0], 0.5.*eye(2), 100); ...
     mvnrnd([2 2], 0.5.*eye(2), 100); ...
     mvnrnd([4 4], 0.5.*eye(2), 100); ...
     mvnrnd([6 6], 0.5.*eye(2), 100); ...
     mvnrnd([8 8], 0.5.*eye(2), 100); ...
     mvnrnd([10 10], 0.5.*eye(2), 100)];
evalFunc(X,6)

X = [mvnrnd([3 6], 0.5.*eye(2), 100); ...
     mvnrnd([0 6], 1.*eye(2), 10); ...
     mvnrnd([2 4], 1.*eye(2), 10)];
evalFunc(X,3)

X = [mvnrnd([3 4], 0.5.*eye(2), 50); ...
     mvnrnd([0 4], 1.*eye(2), 15); ...
     mvnrnd([2 1], 1.*eye(2), 15)];
evalFunc(X,3)
```

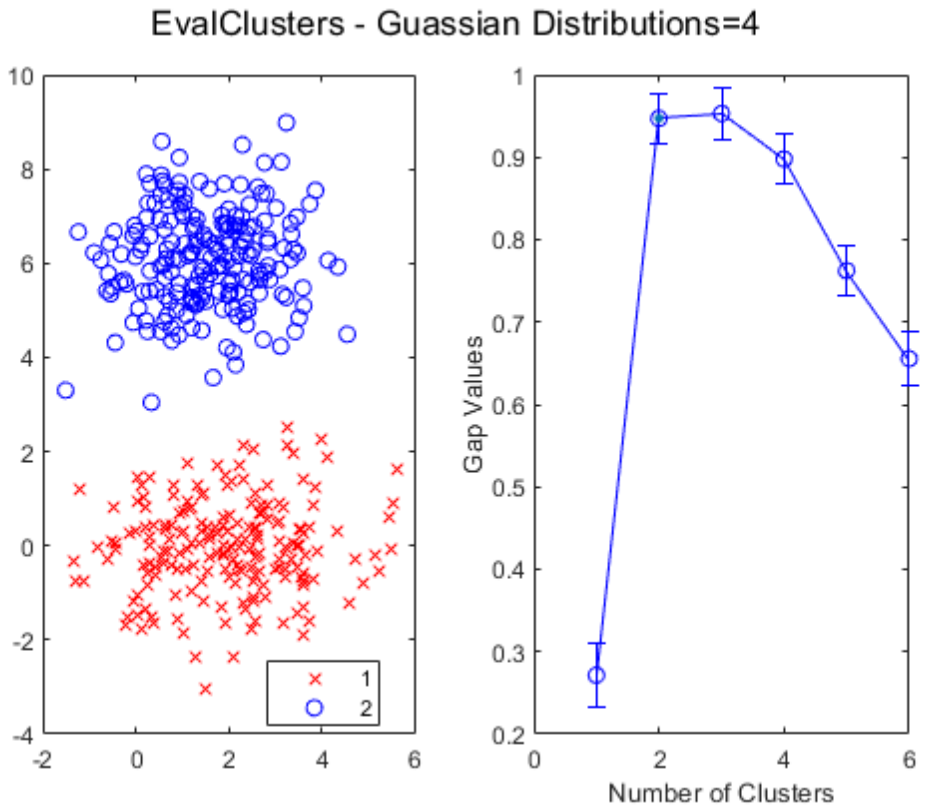
## Eval Clusters Function

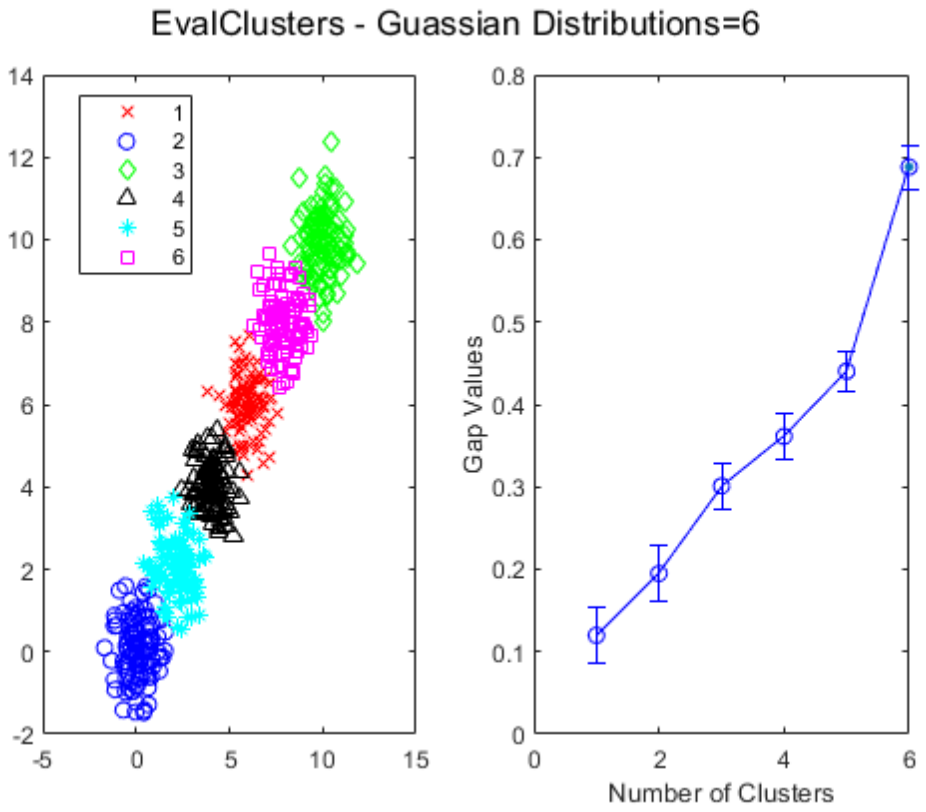
```
function evalFunc(X,k)
    eva = evalclusters(X,'kmeans','gap','KList',1:6); %eval clusters
    figure;
    hold on
    subplot(2,2,[2 4]);plot(eva)
    ClusterGroup = eva.OptimalY; % optimal clusters
```

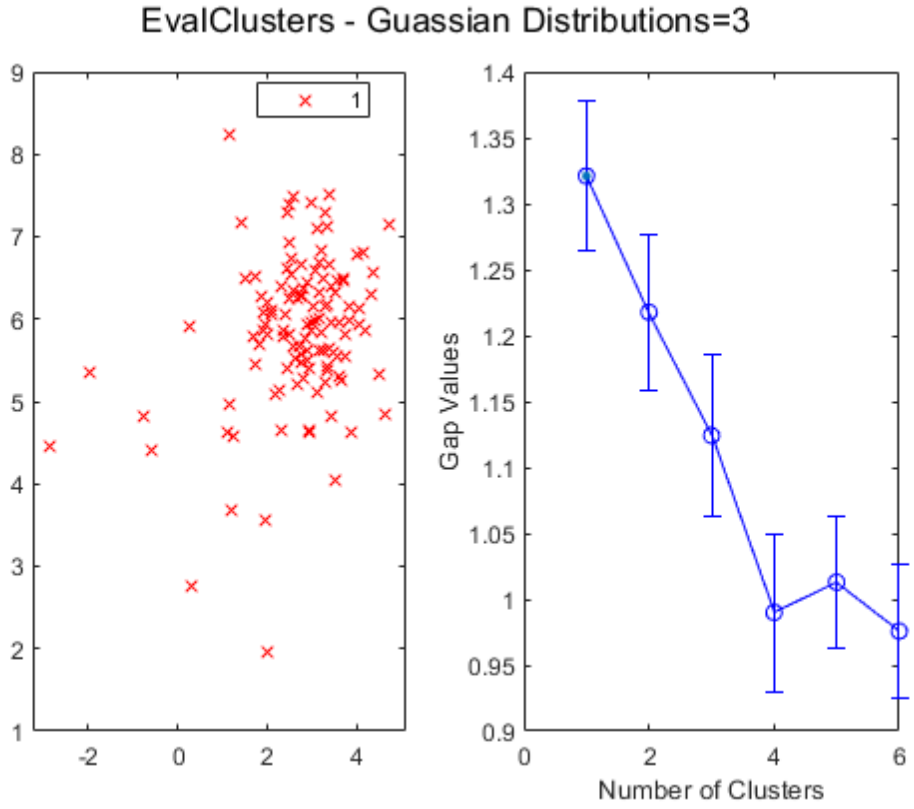
```
    subplot(2,2,[1 3]);gscatter(X(:,1),X(:,2),ClusterGroup,'rbgkcmy','xod^*s+h');
    title_str = "EvalClusters - Guassian Distributions=" + num2str(k);
    sgtitle(title_str)
    hold off
end
```

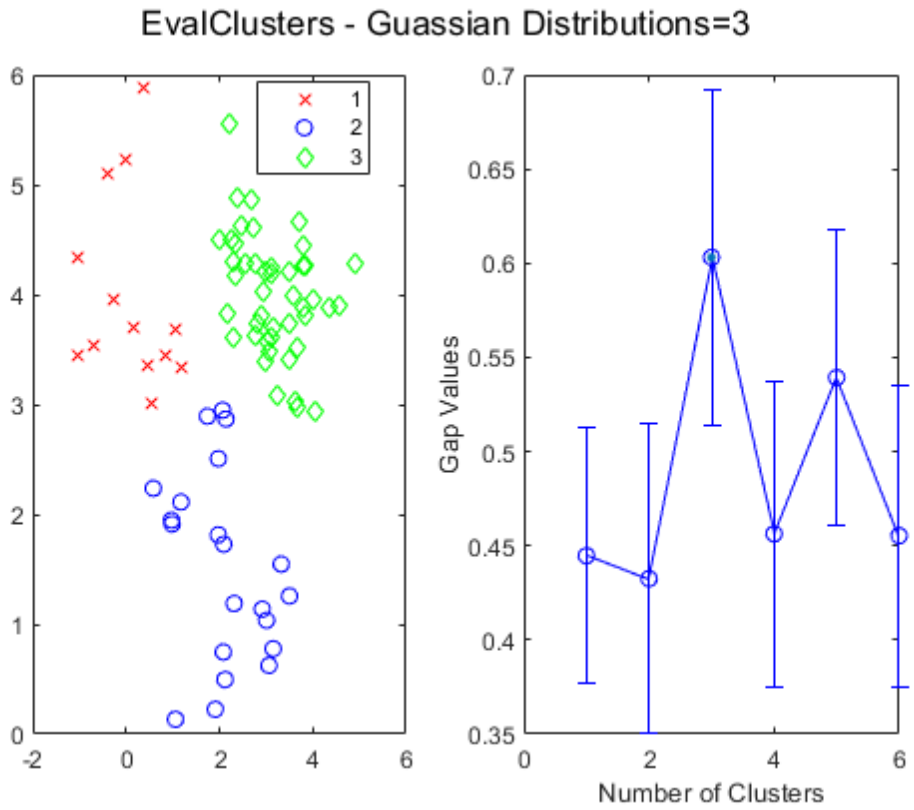

EvalClusters - Guassian Distributions=4
1
2
3
Gap Values
Number of Clusters


EvalClusters - Guassian Distributions=3
1
2
3
Gap Values
Number of Clusters


EvalClusters - Guassian Distributions=4
1
2
Gap Values
Number of Clusters


EvalClusters - Guassian Distributions=6
1
2
3
4
5
6
Gap Values
Number of Clusters




*Published with MATLAB® R2019b*