Vincent Purcell

Professor Cohen

ECE 487 – Pattern Recognition

5 November 2019

Homework 6

"Spectral learning": Kamvar, S. D., D. Klein, C. D. Manning (2003)

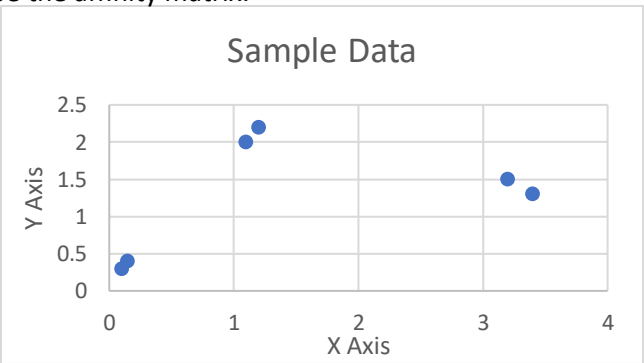
Where does the kernel enter the approach?

The 'kernel' trick is a method that is used to solve a non-linear issue with a linear classifier. Spectral learning is a type of clustering method that uses chains of data points to group them together. Two data points may be not closely related to each other but if there are many similar data points in between them they most likely belong to the same cluster based on this algorithm. Referring to the 'kernel' trick, this algorithm is not linearly separable. I believe this kernel function would be the Markov chain method they detail in the paper. A Markov chain is mathematical concept that uses probabilities or weights in between different states. Within spectral learning it is the weighted distances between data points which helps to cluster the data.

How is the affinity matrix calculated?

An affinity matrix is used in spectral clustering in order to determine the similarity between data points. The affinity matrix is calculated by calculating the quadratic distance between every data point and determining which data points have the highest affinity and thus should be clustered together. Given the following data points this would be the affinity matrix:

	X	Y
x1	1.1	2
x2	1.2	2.2
x3	3.2	1.5
x4	3.4	1.3
x5	0.1	0.3
x6	0.15	0.4



Affinity Matrix						
	x1	x2	x3	x4	x5	x6
x1	0	0.224	2.159	2.404	1.972	1.861
x2	0.224	0	2.119	2.377	2.195	2.084
x3	2.159	2.119	0	0.283	3.324	3.242
x4	2.404	2.377	0.283	0	3.448	3.372
x5	1.972	2.195	3.324	3.448	0	0.112
x6	1.861	2.084	3.242	3.372	0.112	0

What is the relationship between the input data and the data that is presented to the classification (e.g. clustering) algorithm?

The input data is given class descriptors but that data that goes into the classification algorithm is all considered to be all one class, or there are no classes at all. The clustering algorithm bases its decisions on the affinity matrix which is calculated with no knowledge of the classes.

Textbook Problems

Table 5.1 Acceptance Intervals $[-x_\rho, x_\rho]$ Corresponding to Various Probabilities for an $\mathcal{N}(0, 1)$ Normal Distribution

$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.282	1.440	1.645	1.967	2.326	2.576	3.090	3.291

Table 5.1 above is from page 271 of the textbook. The first row '1-p' is the confidence percentage which is 1 minus the significance level. So, for *computer experiment 5.1* our confidence percentage was 0.95 or 95%. The second row is the acceptance interval bounds, x_ρ. This acceptance interval is a set of numbers that allows someone to either accept the null hypothesis or accept an alternate hypothesis.

To calculate x_ρ you use the following equation:

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{N(\bar{x}-\hat{\mu})^2}{\sigma^2}\right)$$

In the above equation N is the number of samples, $\bar{x}$ is the sample mean, σ is the standard deviation, and $\hat{\mu}$ is the hypothesized mean.

Below you can see the code and output figures for *computer experiment 5.1*. There are 4 tests that were run with varying sample sizes and sample means. Each part has 3 subplots and a conclusion textbox. The top two subplots are the two sample data sets with a gaussian fit overlaid. The larger bottom subplot is the two data sets combined with a gaussian fit. The conclusion textbox details the rejection or acceptance of the null hypothesis. The value 'h' that is an output of the *ttest2()* function in MATLAB details that hypothesis conclusion. When h = 0 you can conclude that the means of the data sets are the same with a confidence of 1 minus the significance level. If h = 1 you can conclude that the means are different.

Contents

- Vincent Purcell - HW 6 - ECE487
- Problem 5.1
- Part A
- Part B
- Part C
- Generate Data and Run T-Test

Vincent Purcell - HW 6 - ECE487

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

Problem 5.1

Problem 5.1 from the Text on page 316.

```
rng(10);
```

Part A

mu1=0, mu2=2, N1=100, N2=100

```
N1 = 100;
N2 = 100;
mu1 = 0;
mu2 = 2;
var = 1;
runTtest(mu1,mu2,var,N1,N2,"Part A");
```

Part B

mu1=0, mu2=0.2, N1=100, N2=100

```
runTtest(mu1,0.2,var,N1,N2,"Part B");
```

Part C

mu1=0, mu2=2, N1=150, N2=250 for part 1 and mu1=0, mu2=0.2, N1=150, N2=250 for part 2

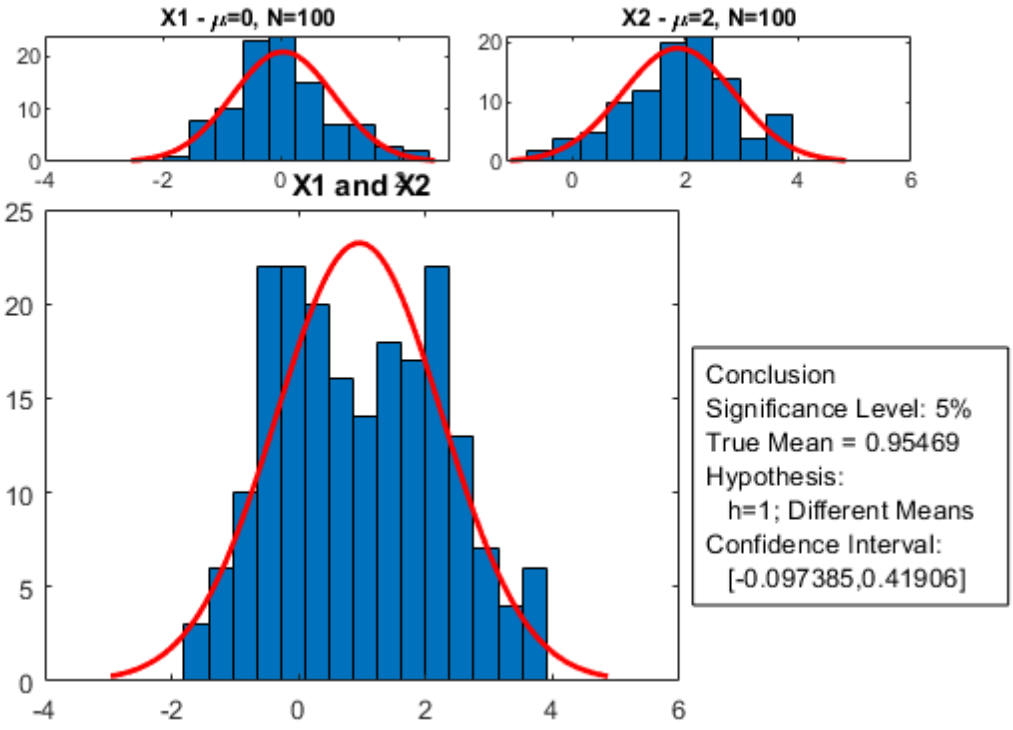
```
runTtest(mu1,mu2,var,150,250,"Part C - 1");
runTtest(mu1,0.2,var,150,250,"Part C - 2");
```

Generate Data and Run T-Test

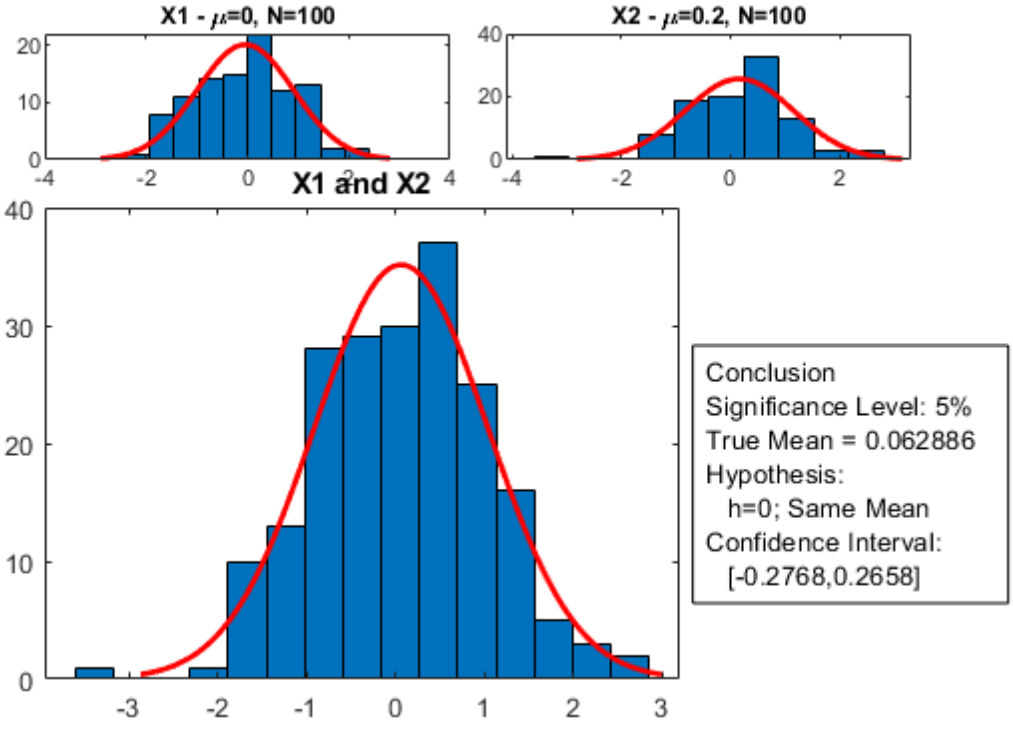
This function will generate two random data sets of size N1 and N2 centered around mu1 and mu2 with a variance of var. This funciton then runs a ttest and plots the two data sets with gaussian fits and the two data sets together. It also displays the results of the ttest on the plot.

```
function runTtest(mu1,mu2,var,N1,N2,sub_title)
    x1 = normrnd(mu1,var,N1,1);
    x2 = normrnd(mu2,var,N2,1);
    [h,~,ci,~] = ttest2(x1,x2);

    %Plot subplots
    figure;
    subplot(4,4,[1 2]); histfit(x1); %X1
    title("X1 - \mu=" + num2str(mu1) + ", N=" + num2str(N1));
    subplot(4,4,[3 4]); histfit(x2); %X2
    title("X2 - \mu=" + num2str(mu2) + ", N=" + num2str(N2));
    subplot(4,4,[5 15]); histfit([x1; x2]); %X1 and X2
    title("X1 and X2");
    sgtitle(sub_title);

    true_mean="True Mean = " + num2str(mean([x1;x2])); %True mean of data
    %Results of null hypothesis rejection/acceptance
    if h==0
        hypothesis="   h=0; Same Mean";
    else
        hypothesis="   h=1; Different Means";
    end
    %confidence interval
    con_int = "   [" + num2str(ci(1)+(mu2-mu1)) + "," + num2str(ci(2)+(mu2-mu1)) + "]";

    %Text that displays results of ttest
    text = {"Conclusion","Significance Level: 5%",true_mean,...
        "Hypothesis:", hypothesis,"Confidence Interval:", con_int};
    annotation('textbox',[0.71 0 0 .5],'String',text,'FitBoxToText','on')
end
```

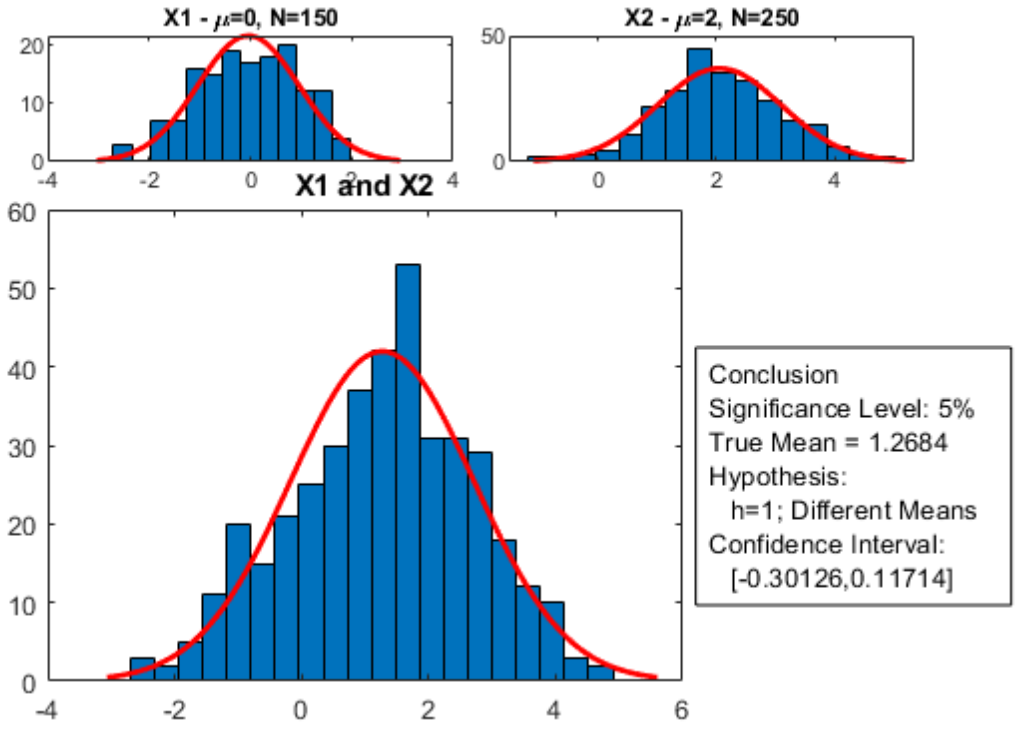
Part A


X1 - μ=0, N=100
X2 - μ=2, N=100
X1 and X2
Conclusion
Significance Level: 5%
True Mean = 0.95469
Hypothesis:
h=1; Different Means
Confidence Interval:
[-0.097385,0.41906]

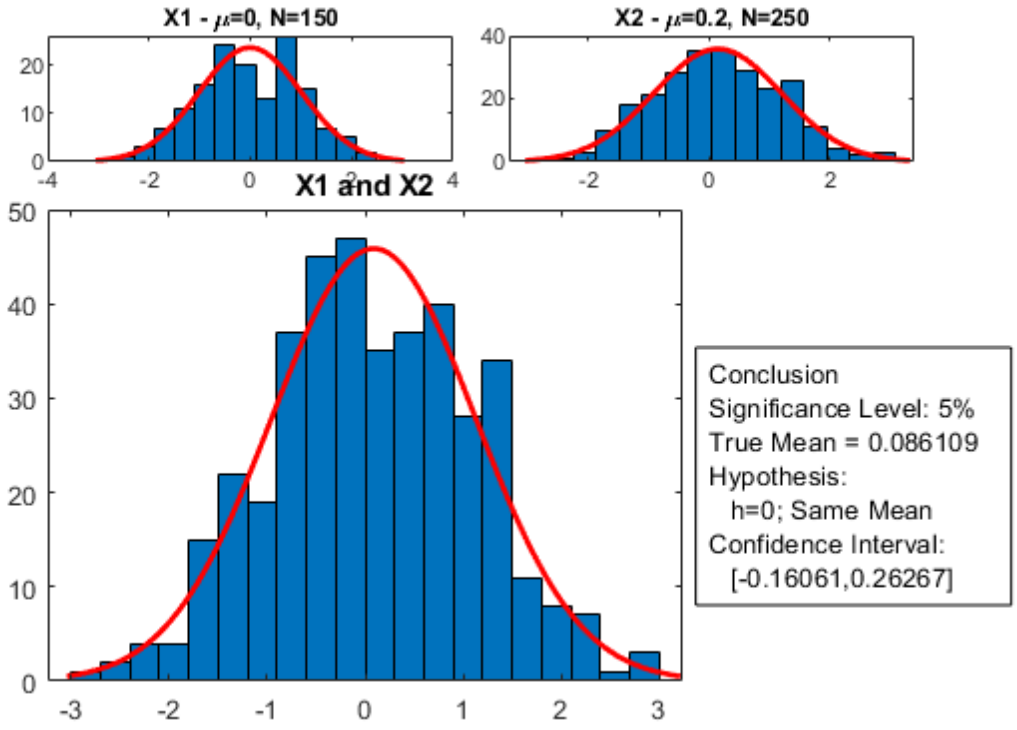
Part B


X1 - μ=0, N=100
X2 - μ=0.2, N=100
X1 and X2
Conclusion
Significance Level: 5%
True Mean = 0.062886
Hypothesis:
h=0; Same Mean
Confidence Interval:
[-0.2768,0.2658]

Part C - 1



Part C - 2



Published with MATLAB® R2019b