Assignment 5

CS422 - Data Mining
Amit Nikam (A20470263)

Recitation Exercises

Exercise 7:

1. Grubb's test as n approaches infinity.

   $\lim_{m\to\infty} m\text{-}1/\sqrt{m} * \sqrt{(t_C^2 / m\text{-}2+t_C^2)} = \lim_{m\to\infty} m\text{-}1 / \sqrt{m(m\text{-}2+t_C^2)} * t_C = t_C$

2. As m limits to infinity, where m is the number of values, the outcome directly depends on t_C which is a parameter which is dependent on significance such that the probability of the sample mean is greater than significance value. Significance α, for our case is 0.05. So the outcome will be a value t_C such that we get more than 5% significance. So distribution of g is defined by t distribution as m increases.

Exercise 8:

1. 3 Standard deviations from the mean in a normal distribution cover about 99.73% of the data. Thus 0.27% of the data is beyond the 3 standard deviations. Thus for our case, in 1,000,000 values, at least 2,700 will be outliers.

2. Outliers are defined by their distance from the normal, i.e. central object data in comparison to others in the dataset. So depending on the size of the dataset, for a large dataset (like millions) there can be thousands of outliers, so the threshold can be increased to have fewer outliers.

Exercise 9:

Mahalanobis distance $= (x-\mu)\Sigma^{-1}(x-\mu)^T$

Proof:

$$f(x) = \underbrace{\frac{1}{(\sqrt{2\pi})^m |\Sigma|^{1/2}} e^{-\frac{(x-\mu)\Sigma^{-1}(x-\mu)^T}{2}}}_{\text{RHS}}$$

taking log on both sides,

$$\log(f(x)) = \log(\text{RHS})$$

$$= \log\left(\frac{1}{(\sqrt{2\pi})^m |\Sigma|^{1/2}}\right) + \log\left(e^{-\frac{(x-\mu)\Sigma^{-1}(x-\mu)^T}{2}}\right)$$

$$= \underbrace{\log(1) - \log\left((\sqrt{2\pi})^m |\Sigma|^{1/2}\right) - \frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T}$$

The constant part doesnot affect the ordering

Thus,

$$\log(f(x)) = (x-\mu)\Sigma^{-1}(x-\mu)^T$$

i.e. $\log(f(x))$ = Mahalanobis distance

Exercise 11:

1. Point D affects the center of the compact cluster by pulling it off the center of the compact cluster.

2. No it won't as the point will still be a cluster in itself making added clusters ineffective on this matter.

3. It would be incorrect in cases where absolute distance is important. For example, consider temperature monitoring of unstable elements. If the temperature goes above or below a specified range of values, then this has a physical meaning. It would be incorrect to not identify such special abnormalities, even though there are cases of relatively similar abnormal temperatures.

Exercise 12:

The detection rate would be 99%.

The false alarm rate would be = (0.99m * 0.01) / (0.99m x 0.01 + 0.01m * 0.99) = 50%

Exercise 16:

The statistical notion of an outlier relies on the idea that an object with low probability is a suspect. But since the distribution we are given is uniform in nature, there would be no outliers thus the statistical notion of an outlier as an infrequently observed value would be meaningless for this.
