Homework 5

Recitation Exercise

Exercise 7

Grubbs Approach for outlier examination.

```
1: Input the values and α.
    - m = number of values
    - α = parameter
    - tc = Value chosen, so that α = prob($|z| \geq g_c$) for a t distribution with m-2 degree of freedom.
2: repeat
3: compute the sample mean ($\overline{x}$) and standard deviation ($s_x$).
4: compute a value $g_c$ so that prob($|z| \geq g_c$) = α.
    - In terms of $t_c$ and m, $g_c = \frac{m-1}{\sqrt{m}}\sqrt{\frac{t_c^2}{m-2+t_c^2}}$
5: compute the z-score of each value, i.e. $z = (x - \overline{x})/s_x$.
6: Let $g = max|z|$ i.e., find the z-score of the largest magnitude and call it g.
7: if $g > g_c$ then
8: Eliminate the value corresponding to $g$.
9: $m \leftarrow m - 1$
10: end if
11: until No objects are eliminated.
```

a) What is the limit of the value $\frac{m-1}{\sqrt{m}}\sqrt{\frac{tc^2}{m-2+tc^2}}$ used for Grubb's test as m approaches infinity? Use a significance level of 0.05.

- $\lim\limits_{m\to\infty} \frac{m-1}{\sqrt{m}} \sqrt{\frac{tc^2}{m-2+tc^2}}$

$$= \lim_{m\to\infty} \frac{m-1}{\sqrt{m(m-2+\ t_c^2}} * t_c$$
$$= 1 * t_c$$
$$= t_c$$

- As value of m increases, value of t_c will also increase gradually.

b) Describe, in words, the meaning of the previous results.

- According to the above result, I realize that the Grubbs test identifies the maximum value from m as an outlier, restricting the value above demonstrates this. The likelihood of the sample mean being greater than equal to t_c is equal to the significance amount, which in this case is 0.05, as can be seen from the algorithm**. As per test distribution of m increases and distribution of g becomes t.**

Exercise 8

Many statistical tests for outliers were developed in an environment in which a few hundred observations was a large data set. We explore the limitations of such approaches.

a) For a set of 1,000,000 values, how likely are we to have outliers according to the test that says a value is an outlier if it is more than three standard deviations from the average? (Assume a normal distribution.)

- Since the aim of this question is to demonstrate that even a small likelihood of an outlier results in a large number of outliers for a large data set. As 0.14 percent on either side of the bell curve is considered to be greater than three standard deviation. As a result, 0.28 percent is greater than three standard deviations. Therefore, the number of outliers will be either 1,400 or 2,800.

b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted when dealing with large data sets? If so, how?

- If outliers are thought to have a negative impact, you should try to reduce the number of outliers. If the outliers are the only thing that people care about, you might expect to see a little more. We may consider these artifacts as outliers or raise the threshold to reduce the number of outliers.

Exercise 9

The probability density of a point x with respect to a multivariate normal distribution having a mean μ and covariance matrix Σ is given by the equation.

f(x)=1(2π)m|Σ|1/2e−(x−μ)Σ−1(x−μ)2.

- Distance between a data point x and the sample mean x plus a constant that does not depend on x.

$$\log \mathrm{prob}(\mathrm{x}) = -\log\left(\left(\sqrt{2\pi}\right)^{m} |\Sigma|^{\frac{1}{2}}\right) - \frac{1}{2}\,(\mathrm{x} - \mu)\Sigma^{-1}(\mathrm{x} - \mu)^{T}\ .$$

If the sample mean and covariance are used as estimates of and, respectively, then.

$$\log \mathrm{prob}(\mathrm{x}) = -\log\left(\left(\sqrt{2\pi)^{m}}\ |S|\text{^}(1/2)\ \right) - \frac{1}{2}\,(\mathrm{x} - \mathrm{x})S^{-1}(x - \overline{x})^{T}$$

We can only keep the variable part, which is the Mahalanobis distance, if we want to base a distance on this quantity. Only the magnitude of the constant and constant factor affects the ordering of this quantity.

Exercise 11

Consider the (relative distance) K-means scheme for outlier detection described in Section 9.5 and the accompanying figure, Figure 9.10.

a) The points at the bottom of the compact cluster shown in Figure 9.10 have a somewhat higher outlier score than those points at the top of the compact cluster. Why?

- Point D pulls the mean of the points away from the compact cluster's center a little.

b) Suppose that we choose the number of clusters to be much larger, e.g., 10. Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

- No, it is not true. This point will form its own cluster.

c) The use of relative distance adjusts for differences in density. Give an example of where such an approach might lead to the wrong conclusion.

- Temperature sensors in nuclear power plants are a good example. When the temperature goes above or below a certain range of values, it has a significant impact. It would be erroneous not to consider any temperature reading outside of that range to be abnormal.

Exercise 12

If the probability that a normal object is classified as an anomaly is 0.01 and the probability that an anomalous object is classified as anomalous is 0.99, then what is the false alarm rate and detection rate if 99% of the objects are normal? (Use the definitions given below.)

$$\text{detection rate} = \frac{\text{number of anomalies detected}}{\text{total number of anomalies}}$$

$$\text{false alarm rate} = \frac{\text{number of false anomalies}}{\text{number of objects classified as anomalies}}$$

- The detection rate = 99%.

The false alarm rate = $\frac{(0.99m \times 0.01)}{(0.99m \times 0.01 + 0.01m \times 0.99)}$

$= 0.50$

$= 50\%$.

Exercise 16

Consider a set of points that are uniformly distributed on the interval [0,1]. Is the statistical notion of an outlier as an infrequently observed value meaningful for this data?

- That is not true. The standard mathematical concept of an outlier assumes that something with a low likelihood is suspicious. No such differentiation can be made in the case of a uniform distribution.