Assignment 7 (Sol.)

Introduction to Machine Learning
Prof. B. Ravindran

1. Which of the following constitute Type I errors?

   (a) the null hypothesis is rejected when it is true.

   (b) the null hypothesis is accepted when it is false.

   (c) the null hypothesis is accepted when it is true.

   (d) the alternate hypothesis is accepted when it is true.

   Solution: A
   By definition of Type I errors.

2. Suppose you are an online advertiser (like Google Ads), which accepts advertisements (consisting of short text and a link) from your customers (companies, such as say Samsung or Hindustan Unilever). You needed to build a system, which on submitting an ad-page to it, classifies it as spam or not spam, and immediately adds it to your corpus of ads if it is not spam. Your development team has come up with two systems - system A and system B, to perform this task. You need to evaluate which system is better for the task using hypothesis testing based methods. Which of these variables are likely to be extraneous to the task? (Note that multiple answers may be correct)

   (a) Classification Accuracy of the system

   (b) Average Click-Through Rate (fraction of users who open the link in the ad) for ads which you classify as non-spam.

   (c) Month of the year

   (d) Regional market in which the system will be deployed (India, Canada or USA)

   Solution: C, D
   A, B represent the variables whose outcome is which we want to monitor and influence the choice of the system. C, D also influence the quality, but they are not being modelled hence they are extraneous.

3. Suppose that a psychologist wants to evaluate the effectiveness of a new learning strategy. She randomly assigns students to two groups and assigns each student the same passage on a particular topic to study for half an hour. Subsequently each student participates in an individual assessment on the topic, where students of the one group use the new learning strategy, and students of the other group use any strategy they prefer. Which among the following is an extraneous variable in the above experiment.

   (a) The choice of using two groups

   (b) The amount of time given to study the passage

   (c) Existing knowledge about the passage among the students

   (d) The amount of time given to complete the assessment

Solution: C
The number of groups is part of the setup which doesn't effect the effectiveness of the learning strategy. The amount of time, is constant for both the groups, hence it is not a factor. The amount of knowledge of the students can hurt influence the results.

4. In the previous question, what step has the experimenter taken to reduce the effect of extraneous variables?

   (a) Split the students into two groups

   (b) Assigned students to the two groups randomly

   (c) Ensured same amount of time is given to each student to read the passage

   (d) Ensured same amount of time is given to each student to complete the assessment

   Solution: B
   Assigning students randomly attempts to prevent any unfair advantage to either group arising due to existing knowledge among the students.

5. I have sampled 20 points from an unknown probability distribution. The sample mean is 5.0 and the standard deviation of the sample is 2.3. Estimate a 95% confidence interval for the mean of the distribution. (You might need to round your answer a little bit to agree with the right option. You can use the t-table available here)

   (a) (4.53, 5.4703)

   (b) (3.948, 6.0517)

   (c) (3.923, 6.076)

   (d) None of the above

   Solution: C

   $$\mu = 5$$
   $$\sigma = 2.3$$

   Compute the Standard Error,

   $$\text{SE} = \frac{\sigma}{\sqrt{n}}$$
   $$\text{SE} = \frac{2.3}{\sqrt{20}}$$

   Now look up the in the two tailed t-table for 0.975 and degree of freedom as 19.

   $$\text{Margin of Error} = \text{SE} \times 2.093 = 1.076$$

   Hence answer would be 5 ± 1.076

6. I have trained a classifier, and to evaluate it's performance I perform a 10 fold validation. I have obtained the following accuracies on the validation set in each of the runs - 0.90, 0.98, 0.95, 0.98, 0.97, 0.96, 0.94, 0.99, 0.96, 0.96. Use this data to answer the next three questions. What is the mean accuracy?

(a) 0.93

(b) 0.959

(c) 0.98

(d) 0.97

Solution: B

$$\mu = \frac{\sum_{i=0}^{N} x_i}{N}$$

7. What is the sample standard deviation for the accuracies?

   (a) 0.0243

   (b) 0.0256

   (c) 6.5444e-04

   (d) 5.8900e-04

   Solution: B

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N} (x_i - \mu)^2}{N - 1}}$$

8. Estimate a 95% confidence interval for the true accuracy of the classifier.

   (a) (0.9407, 0.9773)

   (b) (0.9397, 0.9783)

   (c) (0.9442, 0.9738)

   (d) None of the above.

   Solution: A

$$\text{SE} = \frac{0.0256}{\sqrt{10}} = 0.0081$$

   From the t-table, the critical value for cumulative probability of 0.975 with 9 degrees of freedom. Thus making the Margin of error 0.0183.

9. Which of the following statements is/are true?

   (a) T-test is used when the number of samples is small.

   (b) Z-test is used when the number of samples is small.

   (c) T-test assumes the underlying distribution is a normal distribution.

   (d) T-test assumes the underlying distribution is a beta distribution.

   Solution: A, C

10. If a test of hypothesis has a Type I error probability (α) of 0.01, we mean

    (a) If the null hypothesis is true, we don't reject it 1% of the time.

    (b) If the null hypothesis is true, we reject it 1% of the time.

(c) If the null hypothesis is false, we dont́ reject it 1% of the time.

(d) If the null hypothesis is false, we reject it 1% of the time.

Solution: B