Kavisha Arora

SET 1

Set 1- Macine Learning part 1

MCQ

1. B) In hierarchical clustering you don’t need to assign number of clusters in beginning
2. A) max_depth
3. B) RandomOverSampler
4. C) 1 and 3
5. D) 1-3-2
6. C) K-Nearest Neighbors
7. C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node

Choose the correct options:

8. A and D
   Ridge will lead to some of the coefficients to be very close to 0 and Lasso will cause some of the coefficients to become 0.
9. B, C, D
   remove only one of the features or Use ridge regularization or use Lasso regularization
10. A and D
    Overfitting and Outliers

Subjective answer type questions:

11. **One-hot encoding** may not be suitable in situations where the categorical variable has a large number of categories or levels. One-hot encoding creates a binary variable for each category. It takes only numerical categorical values. Suppose male and female, male is marked as 0 and female as 1. It improves model performance by providing more information to the model. It can also lead to over fitting if there are many categories in the variable and sample size is small.
    Alternative encoding techniques is label encoding or target encoding**. Label encoding** assigns a unique integer label to each category (0,1,2,3,4) which can be more efficient than one-hot encoding. **Target encoding**, on the other hand, replaces each category with the mean or median of the target variable within that category, which can capture the relationship between the categorical variable and the target variable but may be susceptible to overfitting or bias. The choice of encoding technique depends on the specific problem and the characteristics of the categorical variable, and should be evaluated based on the performance of the resulting model.

12. Imbalanced datasets occur when one class is represented by significantly fewer samples than the other classes, which can lead to biased or inaccurate models that prioritize the majority class.
    SMOTE (Synthetic Minority Over-sampling Technique): This technique involves generating synthetic samples of the minority class by interpolating between existing samples and their nearest neighbors in feature space. This method can address the problem of overfitting in random oversampling by generating more diverse and representative synthetic samples, but may also introduce noise or bias if the nearest neighbors are not appropriate or if the data has complex or nonlinear distributions.

    Ensemble methods: This technique involves using a combination of classifiers or models that are trained on different subsets or representations of the data to improve the overall classification performance. Ensemble methods such as bagging, boosting, or stacking can be used to balance the dataset by assigning different weights or penalties to different samples or classes, or by generating diverse and complementary models that can compensate for the weaknesses or biases of individual models.

    Random under sampling: This technique involves randomly selecting a subset of the majority class samples such that the ratio of minority class to majority class samples is closer to the desired balance. This method can be fast and simple, but may result in loss of information and reduced performance if important features or patterns in the data are underrepresented in the selected subset.

13. **SMOTE** (Synthetic Minority Over-sampling Technique) and **ADASYN** (Adaptive Synthetic Sampling) are two popular sampling techniques used to address data imbalance in classification. Both techniques are designed to generate synthetic samples of the minority class in order to balance the class distribution, but they differ in how they generate these samples.

    SMOTE generates synthetic samples by interpolating between existing minority class samples and their nearest neighbors in feature space. Specifically, for each minority class sample, SMOTE selects one or more nearest neighbors and creates a synthetic sample by randomly selecting a point along the line connecting the original sample and the nearest neighbor(s). This allows SMOTE to create new samples that lie within the boundaries of the original class distribution, but may not capture the full complexity or diversity of the minority class.

    ADASYN, on the other hand, is designed to adaptively generate synthetic samples based on the density of the minority class in feature space. Specifically, ADASYN focuses on the samples that are difficult to classify correctly by assigning higher weights to those samples and generating more synthetic samples in the regions where the density of the minority class is low. This allows ADASYN to address the problem of SMOTE generating too many synthetic samples in regions where the density of the minority class is already high, while also capturing the underlying structure and diversity of the minority class more effectively.

In summary, SMOTE generates synthetic samples based on the nearest neighbors of existing minority class samples, while ADASYN adaptively generates synthetic samples based on the density of the minority class in feature space.

14. **GridSearchCV** is a hyperparameter tuning technique used to find the optimal set of hyperparameters for a given machine learning algorithm. The purpose of using GridSearchCV is to systematically search through a specified set of hyperparameters and evaluate their performance using cross-validation to find the best combination of hyperparameters that yields the best performance on the given dataset. GridSearchCV is a useful hyperparameter tuning technique that can be used for both small and large datasets, but its performance may decrease as the dataset and number of hyperparameters increase. it can be computationally expensive, especially for large datasets with a large number of hyperparameters to tune. Other hyperparameter tuning techniques such as RandomizedSearchCV, which searches for the optimal hyperparameters randomly within a specified range, which uses a probabilistic model to optimize the hyperparameters, may be more suitable. These techniques can be faster and more efficient than GridSearchCV, especially for large datasets with many hyperparameters.

15. There are several evaluation metrics that can be used to assess the performance of a regression model. Here are some of the most commonly used ones:

    a) **Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted and actual values. It gives a measure of the overall model accuracy, with higher values indicating a poorer model fit.
    b) **Root Mean Squared Error (RMSE):** RMSE is similar to MSE, but the square root is taken to make the metric more interpretable. It represents the average deviation of the predicted values from the actual values, with lower values indicating a better model fit.
    c) **Mean Absolute Error (MAE):** MAE measures the absolute difference between the predicted and actual values, without squaring the errors. It gives an indication of the average magnitude of the errors, with lower values indicating a better model fit.
    d) **R-squared (R^2):** R-squared is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating a better model fit.
    e) **Adjusted R-squared:** Adjusted R-squared is a modified version of R-squared that adjusts for the number of independent variables in the model. It penalizes the model for including unnecessary variables and provides a more accurate measure of the model's predictive power.

Set 1 part 2 - Python assignment

1. C) %
2. B) 0
3. C) 24

4. A)2
5. D) 6
6. C) the finally block will be executed no matter if the try block raises an error or not.
7. A) It is used to raise an exception
8. C) in defining a generator
9. A) _abc and C) abc2

Question 11 to 15 in jupyter notebook!

Set 1 part 3 – Statistics

Mcq

1. b. The probability of failing to reject H0 when H1 is true
2. b. null hypothesis
3. d. Type I error
4. b. the t distribution with n - 1 degrees of freedom
5. a. accepting Ho when it is false
6. d. a two-tailed test
7. b. the probability of committing a Type I error
8. a. the probability of committing a Type II error
9. a. z > zα
10. a. knowledge of whether the test is one-tailed or two-tail
11. a. level of significance
12. a. Degrees of Freedom

Subjective

13. **ANOVA (Analysis of Variance) in SPSS** is a statistical method used to compare the means of two or more groups of data to determine if there is a significant difference between them. It is used to test the null hypothesis that the means of all groups are equal against the alternative hypothesis that at least one group has a different mean. SPSS provides various types of ANOVA such as one-way ANOVA, factorial ANOVA, and repeated measures ANOVA, among others. The ANOVA output in SPSS includes various statistics such as F-ratio, p-value, and effect size measures, which help to interpret the results and draw conclusions about the differences between the groups.

14. **The assumptions of ANOVA** (Analysis of Variance) include:

a) Normality: The dependent variable should be normally distributed for each group or factor level.

b) Homogeneity of variances: The variances of the dependent variable should be equal across all groups or factor levels.

c) Independence: The observations in each group should be independent of each other.

d) Random sampling: The data should be obtained through a random sampling process.
e) No significant outliers: There should not be any significant outliers in the data.
f) Homogeneity of regression slopes: If ANOVA involves a regression model, the regression slopes should be equal across all groups or factor levels.

15. **One-way ANOVA** (analysis of variance) is a statistical technique used to compare means between two or more independent groups. It is called "one-way" because there is only one independent variable being tested.
    Two-way ANOVA, on the other hand, involves two independent variables, and tests how they interact to influence a dependent variable. In other words, it examines whether the effect of one independent variable on the dependent variable is the same for different levels of the other independent variable. The key difference between one-way and two-way ANOVA is the number of independent variables being tested. One-way ANOVA has one independent variable, while two-way ANOVA has two independent variables.

SET 2

SET 2 PART1 – MACHINE LEARNING

MCQ

1. B) They cannot be used when the data is not completely linearly separable while allowing no errors.
2. B) It’s the classifier for which the margin length or the distance between the closest data-point on either side of the classifier and the classifier is maximized.
3. A , C and D) They are less sensitive to outliers and can be used even in their presence, They allow some degree of errors or misclassification, They can be used in case data is not completely linearly separable.
4. A and B. They take the data from lower dimensional space to some higher dimensional space in case the data is not likely to be linearly separable, They use the kernel tricks to escape the complex computations required to transform the data.
5. A and c. These functions give value of the dot product of pairs of data-points in the desired higher. dimensional space without even explicitly converting the

whole data in to higher dimensional space, The data product values given by the kernel functions are used to find the classifier in the higher dimensional space.

6. D) It is a model trained using supervised learning. It can be used for classification not for regression.
7. D) all of the above. Selection of Kernel , Kernel Parameters ,Soft Margin Parameter
8. D) None of these
9. A) Misclassification would happen.
10. B) How accurately the SVM can predict outcomes for unseen data.

SET 2 part 2- Python Assignment

MCQ

1. B) struct
2. C) 1_no
3. A) in
4. A) left to right
5. B) iii – iv – i – ii
6. C) 0.3333...
7. B) string
8. A) Division and multiplication have same precedence in python and D) In case of operators' having the same precedence, the one on the left side is executed first.
9. A) abc = 1,000,000 and D) a b c = 1000 2000 3000
10. C) x^16

Subjective

11. In Python, a list, tuple, set, and dictionary are all different data structures with different properties and uses. Here are some differences between them:
    a) **List**: Lists are ordered, mutable (modifiable) data structures. They can contain any data type, including other lists or nested data structures. The elements in a list can be accessed using an index, and they are separated by commas and enclosed in square brackets []. Some methods available for lists include. append(), .extend(), .insert(), and .remove().Example: my_list = [1, 2, 3, 'four', True]
    b) **Tuple**: Tuples are ordered, immutable (non-modifiable) data structures. They can contain any data type, including other tuples or nested data structures. The elements in a tuple can be accessed using an index, and they are separated by commas and enclosed in parentheses (). Tuples are useful when you need to store a collection of related values that should not be changed. Example: my_tuple = (1, 2, 3, 'four', True)
    c) **Set**: Sets are unordered, mutable (modifiable) data structures. They can contain any data type, but each element in a set must be unique. Elements in a set are unordered and cannot be accessed using an index. Instead, you can use the in keyword to check if

an element is in the set. Some methods available for sets include .add(), .remove(), .union(), and .intersection().Example: my_set = {1, 2, 3, 'four', True}

d) **Dictionary**: Dictionaries are unordered, mutable (modifiable) data structures. They are composed of key-value pairs, where each key is unique and maps to a corresponding value. Keys in a dictionary are unordered and cannot be accessed using an index. Instead, you can use the key to access its corresponding value. Some methods available for dictionaries include .get(), .keys(), .values(), and .items().Example: my_dict = {'name': 'John', 'age': 30, 'student': True}

12. No, **strings are immutable in Python**, which means that you cannot change individual characters in a string once it has been created. However, you can create a new string with the desired changes. To replace the + character with a space in the string "I+Love+Python", you can use the replace() method. my_string = "I+Love+Python

```
new_string = my_string.replace("+", " ")

print(new_string)
```

OUTPUT = I Love python

13. The **ord() function** in Python returns the Unicode code point of a given character. In other words, it returns the integer value that represents the Unicode character.

Ord()

```
uni_code=ord('A')
print(uni_code) output = 65
```

For data type

```
num=34
data_type =type(num)
print(data_type)
output=int
```

question 14 and 15 done in jupyter notebook!

set 2 part 3 – statistics

Mcq

1. C. Type I; Type II
2. C. We have made a Type II error
3. b. critical value
4. d. A correct decision was made.
5. a. x = 23 s , = 3

6. c. reject H0
7. c. At α = 0.05, reject the null hypothesis.
8. a. 0.100 and c. 0.055
9. d. 0.042
10. b. Two tail
11. a. Less than the significance level
12. b. 0.375

Subjective

13. **T distribution and Z distribution** are both probability distributions that are used in statistical inference, specifically in hypothesis testing and confidence interval estimation. **Z distribution**, also known as the standard normal distribution, is a continuous probability distribution with a mean of 0 and a standard deviation of 1. It is often used when the sample size is large, or when the population standard deviation is known. **Z distribution** is used to test hypotheses and construct confidence intervals about the population mean. **T distribution**, also known as the student's t-distribution, is a continuous probability distribution that resembles the normal distribution, but with heavier tails. **T distribution** is used when the sample size is small and the population standard deviation is unknown. T distribution is also used when the population distribution is not normal or approximately normal. T distribution is used to test hypotheses and construct confidence intervals about the population mean.
    Z distribution is used when the sample size is large and/or the population standard deviation is known, while **T distribution** is used when the sample size is small and/or the population standard deviation is unknown.

14. **The t-distribution is a type of normal distribution** that is used for smaller sample sizes. Normally-distributed data form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tails. The T distribution is similar in shape to the normal distribution, but it has heavier tails, which means that it has more probability in the tails than the normal distribution does. The exact shape of the T distribution depends on its degrees of freedom (df), which is related to the sample size used to estimate the population standard deviation. When the degrees of freedom are large (e.g., above 30), the T distribution becomes very similar to the normal distribution. In fact, as the degrees of freedom approach infinity, the T distribution becomes identical to the normal distribution. Therefore, for large sample sizes, the T distribution can be approximated by the normal distribution.

15. The **T distribution tells us the probability** of obtaining a certain sample mean from a population with an unknown population mean and an unknown population standard deviation, given a sample size and a sample mean. The T distribution is used in statistical inference, specifically in hypothesis testing and confidence interval estimation, to make inferences about the population mean when the population standard deviation is unknown and/or the sample size is small. The T distribution allows us to estimate the variability in the sample mean that is due to chance, and to test hypotheses about the population mean based on the sample mean. The T distribution is used to calculate confidence intervals for the population mean and to perform hypothesis tests about the population mean. Specifically, it allows us to calculate the probability of obtaining a

sample mean that is as extreme or more extreme than the observed one, given that the null hypothesis is true.

SET 3

Set 3 part 1 – Machine learning

Mcq

1. C) y intercept
2. A) True
3. B) the dependent variable
4. B) Linear Regression
5. C) the correlation coefficient squared
6. B) y increases as x increases
7. A) linear data
8. A) 0 to 1
9. B) RMSE and D) MAE
10. A) Linear regression is a supervised learning algorithm.
11. A) Ridge B) Lasso and D) Elastic Net
12. A) Large amount of training samples with small number of features.
13. A) Linearity B) Homoscedasticity

Subjective

14. **Linear regression** is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the dependent variable and the independent variable(s). In other words, linear regression attempts to draw a straight line through the data points that best represents the relationship between the variables. The objective of linear regression is to estimate the values of these coefficients, such that the line drawn through the data points fits the data as closely as possible. This is achieved by minimizing the sum of the squared errors between the predicted values and the actual values of the dependent variable. Linear regression can be used for both simple linear regression, which involves only one independent variable, and multiple linear regression, which involves more than one independent variable.

15. **In simple linear regression**, there is only one independent variable and one dependent variable. The relationship between the two variables is modelled using a straight line. Simple linear regression is used when we want to predict a continuous dependent variable based on a single independent variable.

    In multiple linear regression, there are two or more independent variables and one dependent variable. The relationship between the dependent variable and the independent variables is modelled using a linear equation. Multiple linear regression is used when we want to predict a continuous dependent variable based on multiple independent variables. Multiple linear regression models tend to be more complex than simple linear regression models, as they involve more independent variables and interactions between them.

Set 3 part 2 – Python assignment

Mcq

1. D) int('32') will raise a value error in python.
2. C) 4
3. B) (a**b)%c
4. A) <class 'type'>
5. C) 65
6. D) Method
7. B) False
8. B) Sometimes
9. A) -68.7e100 B) 42e3 and C) C) 4.2038
10. You can call a function with positional and keyword arguments. D) Positional arguments must be before keyword arguments in a function call.

Question 11-15 done in jupyter notebook!

Part 3 set 3- statistics

Mcq

1. a. True
2. b. The underlying distribution
3. a. True
4. C. We are 95% confident that the results have occurred by chance

5. c. If the region of rejection is located in one or two tails of the distribution
6. c. We accept a null hypothesis when it is not true
7. a. It is a sample proportion.
8. a. .013
9. b. 0.745
10. a. –3.75
11. b. At least 16% of American women belong to sports clubs and d)There is no conclusive evidence of a gender difference in the proportion belonging to sports clubs.
12. The FALSE statement is (b) "It is reasonable to say that more than 40% of Americans exercise regularly."

<u>Subjective></u>

13. **T test :**

- Calculate the sample mean for each group.
- Calculate the sample standard deviation for each group.
- Calculate the standard error of the difference between the means of the two groups. This is given by the formula: standard error = sqrt((s1^2 / n1) + (s2^2 / n2)) where s1 and s2 are the sample standard deviations, n1 and n2 are the sample sizes for each group.
- Calculate the test statistic using the formula for the specific test you are using. For example, if you are using a t-test, the formula for the test statistic is: t = (x1 - x2) / standard error where x1 and x2 are the sample means for each group.
- Compare the calculated test statistic to the critical value from the appropriate table (e.g. t-table or z-table) using your chosen significance level (usually 0.05). If the calculated test statistic is greater than the critical value, you can reject the null hypothesis and conclude that there is a significant difference between the two groups.

14. To find t**he sample mean difference** between two groups or samples, you need to calculate the difference between the means of the two samples. The sample mean difference, denoted as (x̄1 - x̄2), is the numerical value that represents the average difference between the two-sample means. To calculate the sample, mean difference, follow these steps:

- Calculate the sample mean for each group or sample. This is done by adding up all of the values in the sample and dividing by the total number of values.
- x̄1 = (sum of values in sample 1) / (number of values in sample 1)
- x̄2 = (sum of values in sample 2) / (number of values in sample 2)
- Subtract the second sample mean from the first sample mean.
- (x̄1 - x̄2)
  The resulting value is the sample mean difference. This value tells you how much, on average, the sample means differ from each other.

15. A two-sample t-test is a statistical test used to compare the means of two independent samples. Here is an example scenario where a two-sample t-test could be used: Suppose a company wants to compare the average daily sales of two different stores located in different cities. The company selects a random sample of 30 days from each store and records the daily sales for each day in dollars. The data is as follows:
Store 1: {150, 200, 175, 180, 190, 225, 250, 225, 220, 215, 175, 160, 200, 195, 230, 240, 250, 200, 185, 240, 190, 180, 220, 230, 195, 175, 190, 205, 220, 235}
Store 2: {175, 185, 160, 140, 190, 220, 230, 200, 195, 180, 190, 175, 200, 185, 215, 240, 225, 190, 205, 220, 215, 225, 205, 190, 200, 175, 195, 210, 220, 230}
The company wants to know if there is a significant difference between the average daily sales of the two stores. To perform a two-sample t-test, we can follow these steps:

- State the null and alternative hypotheses:
- Null hypothesis: The mean daily sales of Store 1 and Store 2 are equal.
- Alternative hypothesis: The mean daily sales of Store 1 and Store 2 are not equal.
- Set the significance level, usually 0.05.
- Calculate the sample mean and sample standard deviation for each group:
- Store 1: x̄1 = 205.33, s1 = 29.08
- Store 2: x̄2 = 197.67, s2 = 27.63
- Calculate the pooled standard deviation:
- s_p = sqrt(((n1-1)s1^2 + (n2-1)s2^2) / (n1+n2-2))
- where n1 and n2 are the sample sizes.
- s_p = sqrt(((30-1)29.08^2 + (30-1)27.63^2) / (30+30-2)) = 28.36
- Calculate the t-statistic:
- t = (x̄1 - x̄2) / (s_p * sqrt(1/n1 + 1/n2))
- t = (205.33 - 197.67) / (28.36 * sqrt(1/30 + 1/30)) = 1.34
- Determine the critical value from the t-distribution table with degrees of freedom = (n1 + n2 - 2) = 58 and significance level = 0.05. The critical value is 2.00.
- Compare the calculated t-statistic with the critical value. Since 1.34 < 2.00, we fail to reject the null hypothesis. We conclude that there is no significant difference between the mean daily sales of Store 1 and Store 2 at the 0.05 significance level.

16.