Mayank Singh Deo

SET 1

Set 1- Macine Learning part 1

MCQ

B) In hierarchical clustering you don’t need to assign number of clusters in beginning

A) max\_depth

B) RandomOverSampler

C) 1 and 3

D) 1-3-2

C) K-Nearest Neighbors

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:

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.

B, C, D remove only one of the features or Use ridge regularization or use Lasso regularization

A and D

Overfitting and Outliers

Subjective answer type questions:

When a categorical variable includes a lot of categories or levels, one-hot encoding might not be the best option. For each category, one-hot encoding generates a binary variable. It only accepts categorical numerical values. Assume that male and female are indicated with a 0 for male and a 1 for female. By giving the model extra data, it enhances model performance. If the variable has a large number of categories while the sample size is limited, it may also result in overfitting. Alternative encoding techniques is label encoding or target encoding. Label encoding, which can be more effective than one-hot encoding, gives each category a distinct integer label (0,1,2,3,4). The mean or median of the target variable within each category is substituted for each category in target encoding, which can represent the link between the categorical variable and the target variable but may be biassed or overfit. The categorical variable's features and the particular problem being addressed determine the best encoding approach, which should then be assessed based on how well the model performs.

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: In order to enhance the overall classification performance, this strategy includes combining classifiers or models that have been trained on several subsets or representations of the data. The dataset can be balanced using ensemble approaches like bagging, boosting, or stacking by applying various weights or penalties to various samples or classes, or by creating a variety of complimentary models that can make up for the shortcomings or biases of individual models.

Random under sampling: In order to achieve the necessary balance, this method includes randomly picking a portion of the majority class samples from the whole population. Although this strategy may be quick and easy, if significant features or patterns in the data are underrepresented in the chosen subset, it could lead to information loss and decreased performance.

To address data imbalance in classification, two common sampling strategies are SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling). Both strategies aim to provide artificial samples of the minority class in order to balance the distribution of classes, but their methods for doing so vary.

By interpolating between existing minority class samples and their closes neighbours in feature space, SMOTE creates synthetic samples. In particular, SMOTE chooses one or more nearest neighbours for each minority class sample and then generates a synthetic sample by randomly choosing a point along the line between the original sample and the nearest neighbour (s). SMOTE is able to do this by producing new samples that are within the bounds of the original class distribution, however they might not completely 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.

The hyperparameter tuning method known as GridSearchCV is used to identify the ideal collection of hyperparameters for a particular machine learning algorithm. In order to determine the optimal set of hyperparameters that produces the highest performance on the given dataset, GridSearchCV is used to systematically search through a given set of hyperparameters and evaluate their performance using cross-validation. Although GridSearchCV is a practical method for hyperparameter tuning that works with both small and large datasets, its performance may suffer as the size of the dataset and the number of hyperparameters rise. When there are many hyperparameters to tune and large datasets, it might be computationally expensive. more methods for fine-tuning hyperparameters, includeRandomizedSearchCV, which searches for the optimal hyperparameters randomly within a specified range, which uses a probabilistic model to optimize the hyperparameters,

It could be better to use RandomizedSearchCV, which employs a probabilistic model to optimise the hyperparameters and randomly searches for the best hyperparameters within a certain range. For big datasets with several hyperparameters, these strategies may be quicker and more effective than GridSearchCV.may be more suitable. These techniques can be faster and more efficient than GridSearchCV, especially for large datasets with many hyperparameters.

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:

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.

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.

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.

R-squared (R²): 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.

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

C) %

B) 0

C) 24

A)2

D) 6

C) the finally block will be executed no matter if the try block raises an error or not.

A) It is used to raise an exception

C) in defining a generator

A) \_abc and C) abc2

Set 1 part 3 – Statistics

Mcq

b. The probability of failing to reject H0 when H1 is true

b. null hypothesis

d. Type I error

b. the t distribution with n - 1 degrees of freedom

a. accepting Ho when it is false

d. a two-tailed test

b. the probability of committing a Type I error

a. the probability of committing a Type II error

a. z > zα

a. knowledge of whether the test is one-tailed or two-tail

a. level of significance

a. Degrees of Freedom

Subjective

In SPSS, a statistical procedure called ANOVA (Analysis of Variance) is used to compare the means of two or more groups of data to see if there is a statistically significant difference between them. It is used to compare the alternative hypothesis, that at least one group has a different mean, to the null hypothesis, that the means of all groups are equal. One-way ANOVA, factorial ANOVA, repeated measures ANOVA, among others, are just a few of the different ANOVA variants offered by SPSS. Many statistics, including the F-ratio, p-value, and effect size measurements, are included in the ANOVA output in SPSS. These statistics aid in interpreting the data and identifying group differences.

The assumptions of ANOVA (Analysis of Variance) include:

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

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

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

Random sampling: The data should be obtained through a random sampling process.

No significant outliers: There should not be any significant outliers in the data.

Homogeneity of regression slopes: If ANOVA involves a regression model, the regression slopes should be equal across all groups or factor levels.

15. A statistical method called one-way ANOVA (analysis of variance) is used to compare the means of two or more independent groups. Due to the fact that just one independent variable is being examined, it is referred to as "one-way".

The two-way ANOVA, on the other hand, examines the interactions between two independent factors and a dependent variable. It investigates, in other words, whether the impact of one independent variable on the dependent variable is the same at various values of the other independent variable. The number of independent variables being examined is the primary distinction between one-way and two-way ANOVAs. Two-way ANOVA has two independent variables, whereas one-way ANOVA only has one.

SET 2

SET 2 PART1 – MACHINE LEARNING

MCQ

B) They cannot be used when the data is not completely linearly separable while allowing no errors.

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.

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.

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.

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.

D) It is a model trained using supervised learning. It can be used for classification not for regression.

D) all of the above. Selection of Kernel , Kernel Parameters ,Soft Margin Parameter

D) None of these

A) Misclassification would happen.

B) How accurately the SVM can predict outcomes for unseen data.

SET 2 part 2- Python Assignment

MCQ

B) struct

C) 1\_no

A) in

A) left to right

B) iii – iv – i – ii

C) 0.3333…

B) string

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.

A) abc = 1,000,000 and D) a b c = 1000 2000 3000

C) x^16

Subjective

In Python, a list, tuple, set, and dictionary are all different data structures with different properties and uses. Here are some differences between them:

List: Ordered mutable (modifiable) data structures include lists. Any type of data, such as nested data structures or additional lists, may be contained within them. A list's elements are separated by commas and contained in square brackets, and an index can be used to retrieve them. Lists can be manipulated using a variety of methods, such as append(), extend(), insert(), and remove() (). An illustration is my list = [1, 2, 3, 'four, True]

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)

Set: Sets are mutable (modifiable), unordered data structures. Any data type may be contained in them, but each element of a set must be distinct. A set's elements are not ordered and cannot be found using an index. Instead, you can check whether an element is in the set by using the in keyword. The following are some of the methods for sets:.add(),.remove(),.union(), and.intersection (). For instance, "my set = 1, 2, 3, "four," True"

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}

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

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

C. Type I; Type II

C. We have made a Type II error

b. critical value

d. A correct decision was made.

a. x = 23 s , = 3

c. reject H0

c. At α = 0.05, reject the null hypothesis.

a. 0.100 and c. 0.055

d. 0.042

b. Two tail

a. Less than the significance level

b. 0.375

Subjective

In statistical inference, specifically in hypothesis testing and confidence interval estimation, the T distribution and Z distribution are both probability distributions. The standard normal distribution, or Z distribution, is a continuous probability distribution with a mean and standard deviation of 0 and 1, respectively. When the sample size is large or the population standard deviation is known, it is frequently utilised. To test theories and create confidence intervals around the population mean, the Z distribution is used. The T distribution, commonly referred to as the student's t-distribution, is a continuous probability distribution with heavier tails than the normal distribution. When the sample size is small and the population standard deviation is unknown, the T distribution is utilised. When the population distribution is not normal or roughly normal, the T distribution is also utilised. To test theories and create confidence intervals around the population mean, one uses the T distribution.

Whereas T distribution is utilised when the sample size is small and/or the population standard deviation is unknown, Z distribution is employed when both of these conditions are met.

For smaller sample sizes, the t-distribution, a kind of normal distribution, is utilised. When shown on a graph, normally distributed data take the shape of a bell, with more observations located close to the mean and fewer in the tails. The T distribution resembles the normal distribution in terms of shape, but it has heavier tails, meaning that there is a greater probability in the tails than there is in the centre. The T distribution's degrees of freedom (df), which are correlated with the sample size used to calculate the population standard deviation, determine the exact form of the distribution. The T distribution closely resembles the normal distribution when the degrees of freedom are high (i.e., above 30). In actuality, the T distribution resembles the normal distribution as the degrees of freedom go closer to infinity. Hence, the normal distribution can roughly resemble the T distribution for large sample sizes.

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

C) y intercept

A) True

B) the dependent variable

B) Linear Regression

C) the correlation coefficient squared

B) y increases as x increases

A) linear data

A) 0 to 1

B) RMSE and D) MAE

A) Linear regression is a supervised learning algorithm.

A) Ridge B) Lasso and D) Elastic Net

A) Large amount of training samples with small number of features.

A) Linearity B) Homoscedasticity

Subjective

The statistical technique of linear regression is used to simulate the association between a dependent variable and one or more independent variables. It presupposes that the dependent variable and the independent variable have a linear relationship (s). In other words, the goal of linear regression is to find the best straight line that can be drawn between the data points to depict the relationship between the variables. In order to create a line that matches the data as closely as feasible, these coefficients' values must be estimated using linear regression. This is accomplished by reducing the sum of squared errors between the dependent variable's actual values and its predicted values.

There is only one independent variable and one dependent variable in simple linear regression. Using a straight line, the relationship between the two variables is modelled. When we wish to forecast a continuous dependent variable based on a single independent variable, we utilise simple linear regression.

One dependent variable and two or more independent variables make up multiple linear regression. A linear equation is used to model the relationship between the dependent variable and the independent variables. When we wish to forecast a continuous dependent variable based on several independent factors, we use multiple linear regression. As they include additional independent variables and interactions between them, multiple linear regression models are typically more complex than simple linear regression models.

Set 3 part 2 – Python assignment

Mcq

D) int('32') will raise a value error in python.

C) 4

B) (a\*\*b)%c

A) <class ‘type’>

C) 65

D) Method

B) False

B) Sometimes

A) -68.7e100 B) 42e3 and C) C) 4.2038

You can call a function with positional and keyword arguments. D) Positional arguments must be before keyword arguments in a function call.

Part 3 set 3- statistics

Mcq

a. True

b. The underlying distribution

a. True

C. We are 95% confident that the results have occurred by chance

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

a. It is a sample proportion.

a. .013

b. 0.745

a. –3.75

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.

The FALSE statement is (b) "It is reasonable to say that more than 40% of Americans exercise regularly."

Subjective>

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. ttable 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 the 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 (x1̄ - 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.

x1̄ = (sum of values in sample 1) / (number of values in sample 1)

x2̄ = (sum of values in sample 2) / (number of values in sample 2)

Subtract the second sample mean from the first sample mean.

(x1̄ - x2)̄

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: x1̄ = 205.33, s1 = 29.08

Store 2: x2̄ = 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 = (x1̄ - x2)̄ / (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.