Assignment 6

STATISTICS WORKSHEET- 6

1. Which of the following can be considered as random variable?
a) The outcome from the roll of a die
b) The outcome of flip of a coin
c) The outcome of exam
Ans : d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?
Ans : a) Discrete
b) Non Discrete
c) Continuous
d) All of the mentioned

3. Which of the following function is associated with a continuous random variable?
Ans: a) pdf :-probability density function
b) pmv
c) pmf
d) all of the mentioned
cumulative distribution function, CDF

4. The expected value or _______ of a random variable is the center of its distribution.
a) mode
b) median
Ans :- c) mean
d) bayesian inference

5. Which of the following of a random variable is not a measure of spread?
Ans : a) variance
b) standard deviation
c) empirical mean
d) all of the mentioned

6. The _________ of the Chi-squared distribution is twice the degrees of freedom.
a) variance
Ans : b) standard deviation
c) mode
d) none of the mentioned

mean is degree of freedom

7. The beta distribution is the default prior for parameters between ____________
a) 0 and 10
b) 1 and 2
And : c) 0 and 1
d) None of the mentioned

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
a) baggyer
Ans : b) bootstrap
c) jacknife
d) none of the mentioned

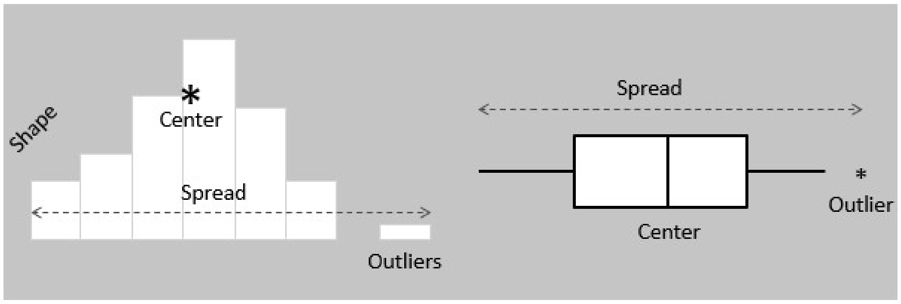
9. Data that summarize all observations in a category are called __________ data.
a) frequency
Ans b) summarized
c) raw
d) none of the mentioned

10. What is the difference between a boxplot and histogram?

Ans : Data visualization is the process of transforming data into an interpretable graphic representation. Different types of data visualization can be utilized to emphasize different key takeaways. For instance, a line graph may be an excellent choice to show a positive trend between two types of quantitative data while a pie chart may be helpful at illustrating percentages.

A **box plot** is a form of data visualization that is used to display several aspects of a data set. A segmented box represents the **median**, **upper quartile (Q3)**, and **lower quartile (Q1)**. A quartile is a segment of data that composes 25% of observed responses, while the median is the data value that represents the middle data value. Lower and higher limits are displayed at the end of **whiskers**, which are lone segments that extend from the main box. Boxplots may also depict values that are far outside of the normal range of responses (referred to as outliers).

A **histogram** is a graphical representation of the spread of data points. Values are broken up into ranges on the x-axis, and the number of responses is denoted on the y-axis. The total number of responses that occur within each range is depicted as a bar graph. Histograms are helpful in determining how data is distributed.



11. How to select metrics?

Ans : First of all, metrics which we optimise tweaking a model and performance evaluation metrics in machine learning are not typically the same. Below, we discuss metrics used to optimise Machine Learning models. For performance evaluation, initial business metrics can be used.

Understanding the task

Based on prerequisites, we need to understand what kind of problems we are trying to solve. Here is a list of some common problems in machine learning:

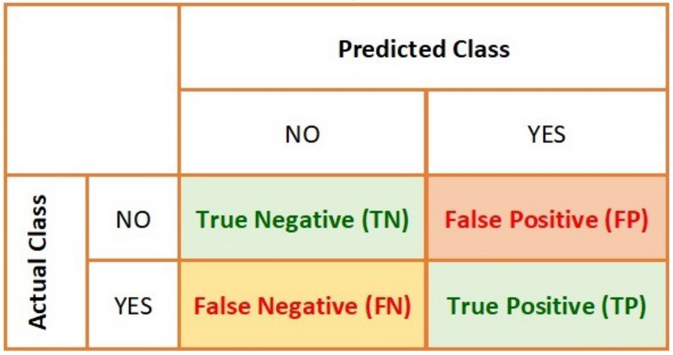
- Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.
- Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.
- Ranking. The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.

In our case, we are solving the problem of finding mathematical metrics which will also optimize the initial business problem. Below we list basic metrics to start with.

CLASSIFICATION of **performance metrics**

Confusion Matrix

Confusion matrix is a very **intuitive cross tab** of actual class values and predicted class values. It contains the count of observations that fall in each category.
Build model → make class predictions on test data using the model → create a confusion matrix for each model. Use one of the following ratios to compare any two models.



		Predicted Class	
		NO	YES
Actual Class	NO	True Negative (TN)	False Positive (FP)
	YES	False Negative (FN)	True Positive (TP)

Let us see all the metrics that can be derived from confusion matrix and when to use them:
1. **Accuracy** — Ratio of correct predictions to total predictions.
Important when: you have symmetric datasets (FN & FP counts are close)
Used when: false negatives & false positives have similar costs.
Accuracy = (TP+TN)/(TP+FP+FN+TN)
2. **Sensitivity/Recall** — Ratio of true positives to total (actual) positives in the data.
Important when: identifying the positives is crucial.
Used when: the occurrence of false negatives is unacceptable/intolerable. You'd rather have some extra false positives (false alarms) over saving some false negatives. For example, when predicting financial default or a deadly disease.
Sensitivity or Recall = TP/(TP+FN)
3. **Precision** — Ratio of true positives to total predicted positives.
Important when: you want to be more confident of your predicted positives.

Used when: the occurrence of false positives is unacceptable/intolerable. For example, Spam emails. You'd rather have some spam emails in your inbox than miss out some regular emails that were incorrectly sent to your spam box.

Precision = TP/(TP+FP)

4. **Specificity** — Ratio of true negatives to total negatives in the data.

Important when: you want to cover all true negatives.

Used when: you don't want to raise false alarms. For example, you're running a drug test in which all people who test positive will immediately go to jail.

Specificity = TN/(TN+FP)

12. How do you assess the statistical significance of an insight?

Ans : To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

13. **Give examples of data that doesnot have a Gaussian distribution, nor log-normal.**

Any type of categorical data won't have a gaussian distribution or lognormal distribution. Exponential distributions-eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

14. Give an example where the median is a better measure than the mean.

15. What is the Likelihood?

MACHINE LEARNING

Assignment 6

1. In which of the following you can say that the model is overfitting?

A) High R-squared value for train-set and High R-squared value for test-set.

B) Low R-squared value for train-set and High R-squared value for test-set.

C) High R-squared value for train-set and Low R-squared value for test-set.

D) None of the above

2. Which among the following is a disadvantage of decision trees?
A) Decision trees are prone to outliers.
B) Decision trees are highly prone to overfitting.
C) Decision trees are not easy to interpret
Ans : D) None of the above.

3. Which of the following is an ensemble technique?
A) SVM B) Logistic Regression
Ans : C) Random Forest D) Decision tree

4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?
A) Accuracy B) Sensitivity
Ans : C) Precision D) None of the above.

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?
A) Model A B) Model B
C) both are performing equal D) Data Insufficient

In Q6 to Q9, more than one options are correct, Choose all the correct options:

6. Which of the following are the regularization technique in Linear Regression?
A) **Ridge** B) R-squared
C) MSE D) **Lasso**

7. Which of the following is not an example of boosting technique?
A) **Adaboost** B) Decision Tree
C) Random Forest D) **Xgboost**.

8. Which of the techniques are used for regularization of Decision Trees?
A) **Pruning** B) L2 regularization
C) Restricting the max depth of the tree D) All of the above

9. Which of the following statements is true regarding the Adaboost technique?
A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points
B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
C) It is example of bagging technique
D) None of the above

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

11. Differentiate between Ridge and Lasso Regression.

Ans : When we talk about regression, we often end up discussing Linear and Logistic Regression, as they are the most popular of the 7 types of regressions. In this article, we'll focus on Ridge and Lasso regression, which are powerful techniques generally used for creating parsimonious models in the presence of a 'large' number of features. Here 'large' can typically mean either of two things:

1. Large enough to enhance the *tendency of a model to overfit* (as low as 10 variables might cause overfitting)

2. Large enough to *cause computational challenges*. With modern systems, this situation might arise in the case of millions or billions of features.

Though Ridge and Lasso might appear to work towards a common goal, the inherent properties and practical use cases differ substantially. If you've heard of them before, you must know that they work by penalizing the magnitude of coefficients of features and minimizing the error between predicted and actual observations. These are called 'regularization' techniques. The key difference is in how they assign penalties to the coefficients:

1. **Ridge Regression:**
   - Performs L2 regularization, i.e., adds penalty equivalent to the **square of the magnitude** of coefficients
   - Minimization objective = LS Obj + α * (sum of square of coefficients)
2. **Lasso Regression:**
   - Performs L1 regularization, i.e., adds penalty equivalent to the **absolute value of the magnitude** of coefficients
   - Minimization objective = LS Obj + α * (sum of the absolute value of coefficients)

Here, LS Obj refers to the 'least squares objective,' i.e., the linear regression objective without regularization.
If terms like 'penalty' and 'regularization' seem very unfamiliar to you, don't worry; we'll discuss these in more detail throughout this article. Before digging further into how they work, let's try to understand why penalizing the magnitude of coefficients should work in the first place.

https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans :

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

Summary

- Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
- Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.
- VIF measures the number of inflated variances caused by multicollinearity.

Variance Inflation Factor and Multicollinearity

In ordinary least square (OLS) regression analysis, multicollinearity exists when two or more of the independent variables demonstrate a linear relationship between them. For example, to analyze the relationship of company sizes and revenues to stock prices in a regression model, market capitalizations and revenues are the independent variables.

A company's market capitalization and its total revenue is strongly correlated. As a company earns increasing revenues, it also grows in size. It leads to a multicollinearity problem in the OLS regression analysis. If the independent variables in a regression model show a perfectly predictable linear relationship, it is known as perfect multicollinearity.

With multicollinearity, the regression coefficients are still consistent but are no longer reliable since the standard errors are inflated. It means that the model's predictive power is not reduced, but the coefficients may not be statistically significant with a Type II error.

Therefore, if the coefficients of variables are not individually significant – cannot be rejected in the t-test, respectively – but can jointly explain the variance of the dependent variable with rejection in the F-test and a high coefficient of determination (R^2), multicollinearity might exist. It is one of the methods to detect multicollinearity.

VIF is another commonly used tool to detect whether multicollinearity exists in a regression model. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity.

Use of Variance Inflation Factor

VIF can be calculated by the formula below:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} = \frac{1}{\mathrm{Tolerance}}$$

Where $\mathbf{R_i^2}$ represents the unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones. The reciprocal of VIF is known as **tolerance**. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.

If R_i^2 is equal to 0, the variance of the remaining independent variables cannot be predicted from the i^{th} independent variable. Therefore, when VIF or tolerance is equal to 1, the i^{th} independent variable is not correlated to the remaining ones, which means multicollinearity does not exist in this regression model. In this case, the variance of the i^{th} regression coefficient is not inflated.

Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

However, there are also situations where high VFIs can be safely ignored without suffering from multicollinearity. The following are three such situations:

1. High VIFs only exist in control variables but not in variables of interest. In this case, the variables of interest are not collinear to each other or the control variables. The regression coefficients are not impacted.
2. When high VIFs are caused as a result of the inclusion of the products or powers of other variables, multicollinearity does not cause negative impacts. For example, a regression model includes both x and x^2 as its independent variables.
3. When a dummy variable that represents more than two categories has a high VIF, multicollinearity does not necessarily exist. The variables will always have high VIFs if there is a small portion of cases in the category, regardless of whether the categorical variables are correlated to other variables.

Correction of Multicollinearity

Since multicollinearity inflates the variance of coefficients and causes type II errors, it is essential to detect and correct it. There are two simple and commonly used ways to correct multicollinearity, as listed below:

1. The first one is to remove one (or more) of the highly correlated variables. Since the information provided by the variables is redundant, the coefficient of determination will not be greatly impaired by the removal.
2. The second method is to use principal components analysis (PCA) or partial least square regression (PLS) instead of OLS regression. PLS regression can reduce the variables to a smaller set with no correlation among them. In PCA, new uncorrelated variables are created. It minimizes information loss and improves the predictability of a model.

https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/

13. Why do we need to scale the data before feeding it to the train the model?

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Ans : Model evaluation is very important in data science. It helps you to understand the performance of your model and makes it easy to present your model to other people. There are many different evaluation metrics out there but only some of them are suitable to be used for regression. This article will cover the different metrics for the regression model and the difference between them. Hopefully, after you read this post, you are clear on which metrics to apply to your future regression model.

Every time when I tell my friends: “Hey, I have built a machine learning model to predict XXX.” Their first reaction would be: “Cool, so what is the accuracy of your model prediction?” Well, unlike classification, accuracy in a regression model is slightly harder to illustrate. It is impossible for you to predict the exact value but rather **how close your prediction is against the real value**.

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square
2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
3. Mean Absolute Error(MAE)

R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R square formula

R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

R Square is a good measure to determine how well the model fits the dependent variables. **However, it does not take into consideration of overfitting problem**. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Mean Square Error formula

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Mean Absolute Error formula

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. **MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same**.

Adjusted R square is the only metric here that considers the overfitting problem. R Square has a direct library in Python to calculate but I did not find a direct library to calculate Adjusted R square except using the statsmodel results.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

Measure	**Value**	**Derivations**
Sensitivity	0.8000	TPR = TP / (TP + FN)
Specificity	0.9600	SPC = TN / (FP + TN)
Precision	0.9524	PPV = TP / (TP + FP)
Negative Predictive Value	0.8276	NPV = TN / (TN + FN)
False Positive Rate	0.0400	FPR = FP / (FP + TN)
False Discovery Rate	0.0476	FDR = FP / (FP + TP)
False Negative Rate	0.2000	FNR = FN / (FN + TP)
Accuracy	0.8800	ACC = (TP + TN) / (P + N)
F1 Score	0.8696	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.7699	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))