Machine learning

Work sheet-1

Ans(1)>>	**A**
Ans(2)>>	**A**
Ans(3)>>	**A**
Ans(4)>>	**B**
Ans(5)>>	**C**
Ans(6)>>	**A**
Ans(7)>>	**D**
Ans(8)>>	**A**
Ans(9)>>	**A**
Ans(10)>>	**B**
Ans(11)>>	**B**
Ans(12)>>	**A,B**
Ans(13)>>	

Regularization

Regularisation is a technique that discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

If model is overfitting, model is trying too hard to capture the noise in your training dataset. By noise mean the data points that don’t really represent the true properties of data, but random chance.

The concept of balancing bias and variance, is helpful in understanding the phenomenon of overfitting.

A simple relation for linear regression-

Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

Y ≈ β0 + β1X1 + β2X2 + …+ βpXp

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

This will adjust the coefficients based on training data. If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

Ridge Regression

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

RSS - modified by adding the shrinkage quantity.

The coefficients are estimated by minimizing this function.

λ -tuning parameter that decides how much want to penalize the flexibility of our model.

Lasso

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Lasso, in which the above function is minimized. It’s differs from ridge regression only in penalizing the high coefficients. It is known as the L1 norm.

Ans (14)>>>

Algorithms are used for regularization is

1. Ridge Regression (L2 Norm)

2. Lasso (L1 Norm)

3. Dropout

Ridge and Lasso can be used for any algorithms involving weight parameters, including neural nets. Dropout is primarily used in any kind of neural networks e.g. ANN, DNN, CNN or RNN to moderate the learning. Let’s take a closer look at each of the techniques.

Ridge Regression (L2 Regularization)

Ridge regression is also called L2 norm or regularization.

When using this technique, we add the sum of weight’s square to a loss function and thus create a new loss function which is denoted thus:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} Wi^2$$

As above, the original loss function is modified by adding normalized weights. Normalized weights are in the form of squares.

λ - parameter that needs to be tuned using a cross-validation dataset.

When λ=0, it returns the residual sum of square as loss function which is initially chosen. For a very high value of λ, loss will ignore core loss function and minimize weight’s square and will end up taking the parameters’ value as zero.

The parameters are learned using a modified loss function. To minimize the above function, parameters need to be as small as possible. Thus, L2 norm prevents weights from rising too high.

Lasso Regression (L1 Regularization)

Also called lasso regression and denoted as below:

$$\text{Loss} = \sum_{j=1}^{m} \left(Yi - Wo - \sum_{i=1}^{n} Wi\, Xji \right)^2 + \lambda \sum_{i=1}^{n} |Wi|$$

Lasso technique is different from ridge regression as it uses absolute weight values for normalization. λ is again a tuning parameter and behaves in the same as it does when use ridge regression.

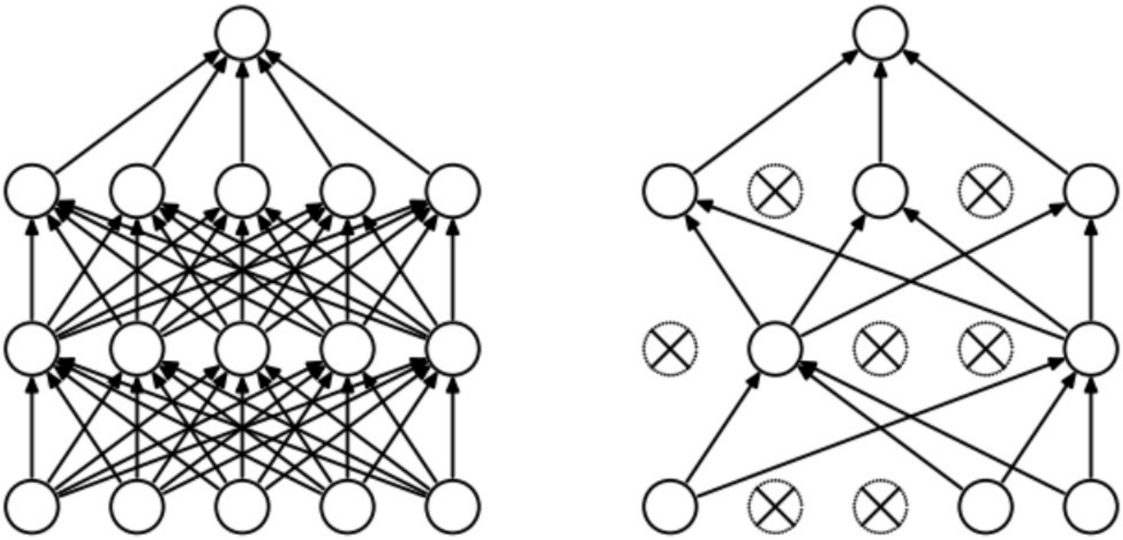
As loss function only considers absolute weights, optimization algorithms penalize higher weight values.

In ridge regression, loss function along with the optimization algorithm brings parameters near to zero but not actually zero, while lasso eliminates less important features and sets respective weight values to zero. Thus, lasso also performs feature selection along with regularization.

Dropout

Dropout is a regularization technique used in neural networks. It prevents complex co-adaptations from other neurons.

In neural nets, fully connected layers are more prone to overfit on training data. Using dropout, you can drop connections with *1-p* probability for each of the specified layers. Where *p* = **keep probability parameter** and which needs to be tuned.



(a) Standard Neural Net (b) After applying dropout.

With dropout, left with a reduced network as dropped out neurons are left out during that training iteration.

Dropout decreases overfitting by avoiding training all the neurons on the complete training data in one go. It also improves training speed and learns more robust internal functions that generalize better on unseen data. However, it is important to note that Dropout takes more epochs to train compared to training without Dropout (If you have 10000 observations in your training data, then using 10000 examples for training is considered as 1 epoch).

Ans(15)>>

A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e

$$\sum e_i{}^2 = \sum (Y_i - \hat{Y}_i)^2$$

For an in-depth understanding of the Maths behind Linear Regression,

PYTHON WORKSHEET-1

Answer sheet

1) Ans.>> (c)

2) Ans.>> (b)

3) Ans.>> (c)

4) Ans.>> (a)

5) Ans.>> (d)

6) Ans.>> (b)

7) Ans.>> (a)

8) Ans.>> (c)

9) Ans.>> (a,c)

10)Ans.>> (a,b)

STATISTICS WORKSHEET -1

Answer sheet

1) Ans.>> (a)

2) Ans.>> (a)

3) Ans.>> (b)

4) Ans.>> (d)

5) Ans.>> (c)

6) Ans.>> (b)

7) Ans.>> (b)

8) Ans.>> (a)

9) Ans.>> (a)

10) Ans.>> The term normal distribution is a most important probability distribution. It is also known as the Gaussian distribution .It is symmetric about the mean and perfectly symmetrical around its center. In graph it is in

well shape because of its probability density looks like a well. The well curve is symmetrical,half of the data will fall to the left of the mean whereas the half of the data will fall to the right mean.

Properties of a normal distribution are as :-

The curve is symmetric at the center (around the mean) with no screw.

The mean median and mode are all equal.

The total area under the curve is 1

Height, birth, weight etc. Are just a few example of it.

This cab be described by two variable :the mean (location parameter) and standard deviation (scale parameter).

EMPERICAL RULE:

Mean + / - standered Deviation	**Percentage of data**
1	**68%**
2	**95%**
3	**99%**

11) Ans.>> The data often found missed data even if it is well designed. Missing data can reduce the statical power of study.

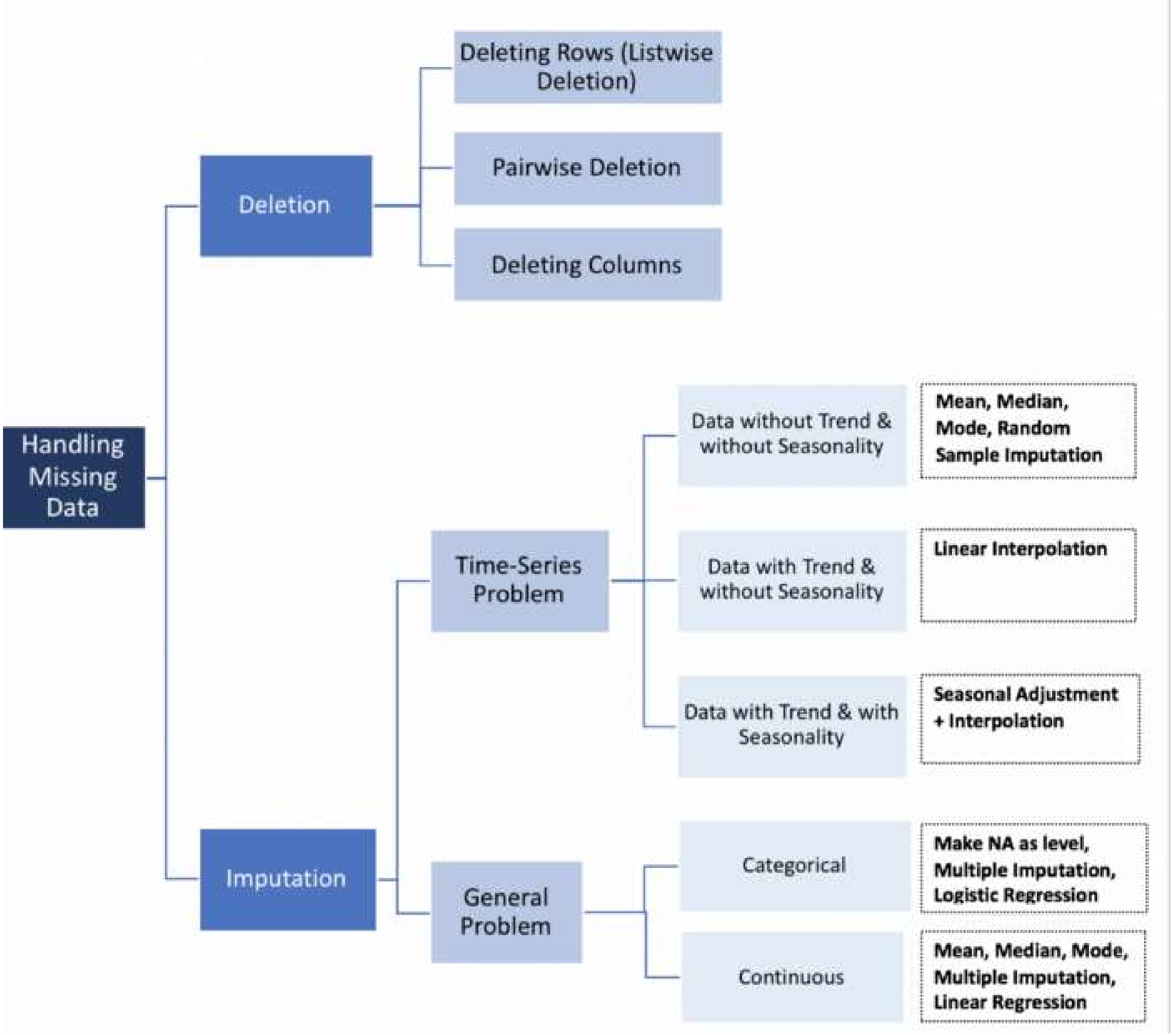
TYPES OF MISSING DATA

Missing completely at random

Missing at random

Missing not at random

Techniques for handling missing data-


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

I prefer the Complete Case Analysis(CCA):-

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. This method is also popularly known as “Listwise deletion”.

Assumptions:-

- Data is Missing At Random(MAR).
- Missing data is completely removed from the table.

Advantages:-

- Easy to implement.
- No Data manipulation required.

Limitations:-

- Deleted data can be informative.
- Can lead to the deletion of a large part of the data.
- Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.
- The production model will not know what to do with Missing data.

When to Use:-

- Data is MAR(Missing At Random).
- Good for Mixed, Numerical, and Categorical data.
- Missing data is not more than 5% – 6% of the dataset.
- Data doesn’t contain much information and will not bias the dataset.

12) Ans.>>

A/B testing refers to the experiments where two or more variations of the same webpage are compared against each other by displaying them to real-time visitors to determine which one performs better for a given goal.

A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample

A/B testing process can be simplified as follows:

1. start the A/B testing process by making a claim (hypothesis).

2. launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.

3.The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

hypothesis breaks down into:

Null hypothesis

Alternative hypothesis

The null hypothesis states the default position to be tested or the situation as it is (assumed to be) now, i.e. the status quo.

The alternative hypothesis challenges the status quo (the null hypothesis) and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that A/B test will prove to be true.

A/B Testing Errors

Hypothesis testing (A/B testing) is a decision-making method. You can make the right decision or you can make a mistake.

In hypothesis testing there are three possible outcomes of the test:

No error-everything is clear

Type I error-occurs when you incorrectly reject the null hypothesis and conclude that there is actually a difference between the original page and the variation when there really isn't.

Type II error-occurs when you fail to reject the null hypothesis at the right moment, obtaining this time false negative test results.

13) Ans.>>

It's a popular solution to missing data, despite its drawbacks.

But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Problem #1: Mean imputation does not preserve the relationships among variables.

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

This is the original logic involved in mean imputation.

Problem #2: Mean Imputation Leads to An Underestimate of Standard Errors

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

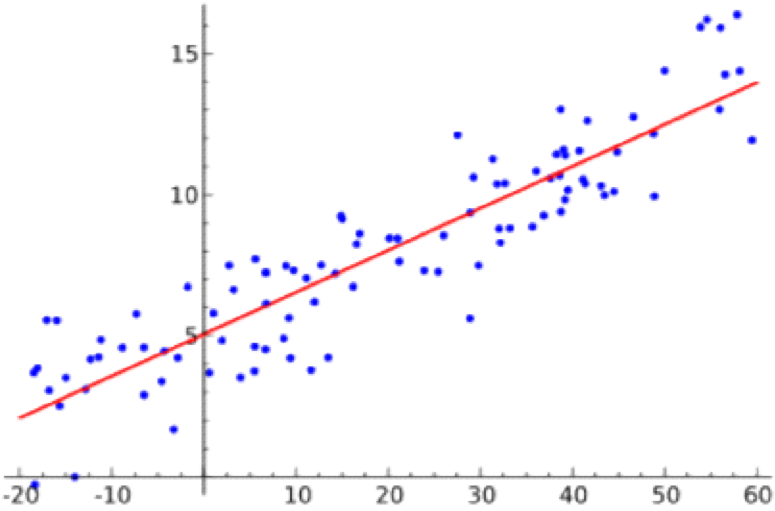
Ultimately, because your standard errors are too low, so are your p-values. Now you’re making Type I errors without realizing it.

That’s not good.

14) Ans.>>

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).
Simple and multiple linear regression[edit]



Example of simple linear regression, which has one independent variable
The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as multiple linear regression, also known as multivariable linear regression (not to be confused with multivariate linear regression [10]).
Multiple linear regression is a generalization of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable. The basic model for multiple linear regression is
{\displaystyle Y_{i}=\beta _{0}+\beta _{1}X_{i1}+\beta _{2}X_{i2}+\ldots +\beta _{p}X_{ip}+\epsilon _{i}}

for each observation i = 1, ... , n.
In the formula above we consider n observations of one dependent variable and p independent variables. Thus, Yi is the ith observation of the dependent variable, Xij is ith observation of the jth independent variable, j = 1, 2, ..., p. The values βj represent parameters to be estimated, and εi is the ith independent identically distributed normal error.
In the more general multivariate linear regression, there is one equation of the above form for each of m > 1 dependent variables that share the same set of explanatory variables and hence are estimated simultaneously with each other:

$${\displaystyle Y_{ij}=\beta _{0j}+\beta _{1j}X_{i1}+\beta _{2j}X_{i2}+\ldots +\beta _{pj}X_{ip}+\epsilon _{ij}}$$

for all observations indexed as i = 1, ... , n and for all dependent variables indexed as j = 1, ... , m.
Nearly all real-world regression models involve multiple predictors, and basic descriptions of linear regression are often phrased in terms of the multiple regression model. Note, however, that in these cases the response variable y is still a scalar. Another term, multivariate linear regression, refers to cases where y is a vector, i.e., the same as general linear regression.

15) Ans.>>

Data collection

Data collection is all about how the actual data is collected. For the most part

Descriptive Statistics

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You are most likely to be familiar with this branch of statistics, because many examples arise in everyday life. Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy. (Chapters 2 and 3 warn you of the common pitfalls of using descriptive methods.)

Inferential Statistics

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

INTERPRETATION When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer. (A full explanation of common inferential methods appears in Chapters 6 through 9.)

For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by surveying the speeds of cars (data collection) and then arrive at a conclusion as to whether the speed limit needs to be lowered

Discrete and continuous data

Data comes in two distinct types. Discrete data can take distinct values, which can be clearly identified and separated. An example of this is the score obtained by rolling a die, which can only take values of 1, 2, 3, 4, 5 or 6, with nothing in between, and all the scores can be distinguished. By contrast, continuous data can take any value. For example, when you measure the speed of a car, it could take any value, depending on how accurately you measure it – for example 31.2 or 48.28, or 48.281 – basically any value.