PART-I → Data & Programming Ideas.

Q1) What is the difference b/w long and wide data? What pandas function are used to transform one data type to the other?

Sol:- Long data is "long" - typically there will be 3 columns, id, variable, and value., which would look something like

ID	Variable	Value.
Barry Bonds	HR	762
Pete Rose	H	4256.
Barry Bonds.	RBI	1996

Wide format data will have a row for each ID, with variables as columns.

Pandas use pivot to go from long to wide, and melt to go from wide to long.

Q2) What is the difference between a pandas dataframe and a series?

Sol: Series is a type of list in pandas which can take integer values., string values, double values and more.

→ Series can only contain single list with index whereas dataframe can be made of more than one series or we can say that a dataframe is a collection of series that can be used to analyze the data.

~~Q3) What is the difference~~

Q3) What are the types of joins and how do they differ from one another?

Sol:- There are mainly 4 types of joins -

i) Inner Join
ii) Left Join
iii) Outer Join
iv) Right Join

~~Inner Join~~ Inner Join ~~gives~~ Returns a dataframe with only
~~Outer Join~~ these rows that have common characteristics
gives us rows with the intersection for index values that are
contained in both datasets.

Two dataframes A and B,

A:

index	a	b
0		
1		
2		
3		
4		

B:

index	c	d
0		
2		
3		
5		

Inner Join ⇒

index	a	b	c	d.
0	—	—	—	—
2	—	—	—	—
3	—	—	—	—

Left Join – Returns a dataframe containing all the rows of the left data frame. Contains values that are index on the left

A.join(B, how='left')

index	a	b	c	d.
0	—	—	—	—
1	—	—	NA	NA.
2	—	—	—	—
3	—	—	—	—
4	—	—	NA	NA.

Right Join – Similar to left join. The only difference is that all the rows of the right dataframe are taken as it is and only those of the left dataframe that are common to both

index	a	b	c	d.
0	—	—	—	—
2	—	—	—	—
3	—	—	—	—
5	NA	NA	—	—

Outer Join – Returns all these records which either have a match in the left or right dataframe.

index	a	b	c	d
0	—	—	—	—
1	—	—	NA	NA
2	—	—	—	—
3	—	—	—	—
4	—	—	NA	NA
5	NA	NA	—	—

Q4.). What is groupby and what does it do?

Sol: Groupby's find all the pairs of matching values in a set of columns, and then perform aggregations on them.

In other words, it is a function used to split the data into groups based on some criteria

eg:-

	Animal	Max Speed
0	Falcon	380
1	Falcon	370
2	Parrot	24
3	Parrot	26.

df.groupy(['Animal']).mean()

Animal	Max Speed
Falcon	375
Parrot	25

Q5.) When I do arithmetic operations on two series with partial overlapping indexes, what is the index of new series? What values are there for indexes in both the series, and ones that are only in one?

Sol: The index of the new series is a union of indices of both the series. The value for the indices in both the series is the resulting operation whereas for the ones that are only in one is NaN.

area = pd. Series ({'Alaska': 1723337, 'Texas': 695662, 'California': 423967}, name = 'area')

Population = pd. Series ({'California': 38332521, 'Texas': 26448193, 'New York': 19651127, name = 'pop'})

area | population

Alaska	NaN
California	0.011060
New York	NaN
Texas	0.026303

PART – II → PROBABILITY

Q1.) Define a sample space. What are the 3 axioms of probability?

Sol: The sample space of an experiment or a random trial is a set of all possible outcomes or results of that experiment. Probability is a function that maps events to $\mathbb{R}$ and follows 3 axioms –

i) $P(E) \in [0,1]$

ii) $P(S) = 1$

iii) If we have sequence of disjoint sets $E_1, E_2, \ldots, E_n$

$$E_i \cap E_j = \phi$$

$$P(E_1 \cup E_2 \ldots \cup E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$$

Q2.) What does it mean when 2 events are independent

Sol: Two events are independent if the occurence of one does not affect the probability of the other.

If E_1 and E_2 are 2 independent events then

$$P(E_1 \cap E_2) = P(E_1) P(E_2)$$

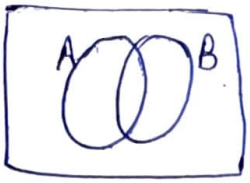
X, Y are independent iff

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

$$P(x,y) = P_x(x) P_y(y) \quad ; \quad f(x,y) = f_x(x) f_y(y)$$

Q3.) What is the probability of the union of 2 events (not necessarily disjoint)?

Let A and B be 2 events that are not disjoint.



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Q4.) What is Conditional Probability? What is Baye's Theorem?

The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occured.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B \cap A)}{P(A)}$$

Bayes Theorem is a way of finding a probability when we know certain other probabilities.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$\Rightarrow P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Q5.) Suppose I role 2 sin-sided dice. X_1 is the result of the first die roll, X_2 is the result of the second die roll and $X = X_1 + X_2$. Calculate the foll.:

$P(X=5)$

$P(X_1 = 2 | X = 6)$

$P(X = 6 \cup X_2 = 3)$

Sol: $P(X_1 + X_2 = 5) = P(\{(1,4) \cup (2,3) \cup (3,2) \cup (4,1)\})$

$$= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9}$$

$P(X=5) = 1/9$

$$P(X_1=2 \mid X=6) = \frac{P(X_1 \cap X)}{P(X)} = \frac{P(X_1=2 \cap X=6)}{P(X=6)}$$

$$P(X_1=2 \cap X=6) = P(\{(2,4)\}) = \frac{1}{36}$$

$$P(X=6) = P(\{(1,5),(2,4),(3,3),(4,2),(5,1)\})$$

$$= \frac{5}{36}$$

$$P(X_1=2 \mid X=6) = \frac{1/36}{5/36} = 1/5$$

$$P(X=6 \cup X_2=3) = P(X=6) + P(X_2=3) - P(X=6 \cap X_2=3)$$

$$= \frac{5}{36} + \frac{6}{36} - \frac{1}{36} = \frac{10}{36} \frac{5}{18} = \frac{5}{18}.$$

Q6) What is a random variable? What is the difference b/w discrete and continuous random variable?

Sol:- A random variable, usually written X is a variable whose possible values are numerical outcomes of a random phenomenon.

There are 2 types of random variables -

Discrete Random Variable - which takes finite values or countably infinite values.

Continuous Random Variables - takes every value on atleast one interval.

Q7.) Suppose a random variable X has the following density -

$$f(x) = \begin{cases} cx^2 & -1 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Calculate c so that this is a proper density and calculate the expectation and variance of X.

Sol: $\int_{-1}^{1} x^2 dx = \left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{2c}{3} = 1$

$\Rightarrow \boxed{c = \frac{3}{2}}$

$\therefore f(x) = \begin{cases} \frac{3}{2}x^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$

$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{-1}^{1} x \cdot \frac{3}{2} x^2 dx = \frac{3}{2}\left[\frac{x^4}{4}\right]_{-1}^{1}$

$E[X] = \frac{3}{2}\left(\frac{1}{4} - \frac{1}{4}\right) = 0$

$Var[X] = E[X^2] - (E[X])^2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$

$\Rightarrow Var[X] = \int_{-1}^{1} x^2 \cdot \frac{3}{2} x^2 dx = \frac{3}{2}\left[\frac{x^5}{5}\right]_{-1}^{1} = \frac{3}{5}$

Q8) Suppose a discrete random variable X has the following p.m.f.

$$p(x) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 1 \\ 0.25 & x = 2 \end{cases}$$

Calculate the expectation and variance of X.

Sol:- $E[X] = \sum_{x=0}^{2} x \cdot f(x) = 0(0.25) + 1(0.5) + 2(0.25)$

$= 1$

$Var[X] = E[X^2] - (E[X])^2$

$E[X^2] = 0^2(0.25) + 1^2(0.5) + 2^2(0.25)$

$= 0 + 0.5 + 1 = 1.5$

$\Rightarrow Var[X] = 1.5 - 1 = 0.5$

89) If X and Y have the joint density

$$f(x,y) = \begin{cases} x+y & x,y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Calculate the covariance and correlation of X and Y.

Sol:

$$f_x(x) = \int_0^1 (x+y)\,dy = x + \frac{1}{2}.$$

$$f_y(y) = \int_0^1 (x+y)\,dx = y + \frac{1}{2}$$

$$E[X] = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \left[\frac{x^3}{3} + \frac{x^2}{4}\right]_0^1 = \frac{7}{12}.$$

$$E[Y] = \int_0^1 y\left(y+\frac{1}{2}\right)dy = \left[\frac{y^3}{3} + \frac{y^2}{4}\right]_0^1 = \frac{7}{12}$$

$$E[X^2] = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \left[\frac{x^4}{4} + \frac{x^3}{6}\right]_0^1 = \frac{5}{12}.$$

$$Var[X] = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$\Rightarrow SD[X] = \sqrt{\frac{11}{144}}. \qquad Var[Y] = \frac{11}{144}. \qquad SD[Y] = \sqrt{\frac{11}{144}}$$

$$E[XY] = \int_0^1\int_0^1 xy*(x+y)\,dx\,dy = \int_0^1\int_0^1 (x^2y + xy^2)\,dx\,dy$$

$$= \int_0^1\int_0^1 x^2y\,dx\,dy + \int_0^1\int_0^1 xy^2\,dx\,dy.$$

$$= \int_0^1 \left[\frac{x^3}{3}\right]_0^1 y\,dy + \int_0^1 \left[\frac{x^2}{2}\right]_0^1 y^2\,dy$$

$$= \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{3} = \frac{2}{6} = \frac{1}{3}.$$

$$\therefore \text{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

$$= \frac{1}{3} - \frac{7}{12}\cdot\frac{7}{12}$$

$$= -\frac{1}{144}$$

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{SD(X)\,SD(Y)} = \frac{-1/144}{\frac{11}{144}} = \frac{-1}{11}$$

$$\text{Corr}(X,Y) = \frac{-1}{11}$$

PART-III : STATISTICS

Q1) Suppose we have a sample $X_1, \ldots, X_n$ of random variable with the following density:

$$f(x) = \begin{cases} ax * \exp\left(-\frac{ax^2}{2}\right), x>0 \\ 0 \quad , \text{otherwise} \end{cases}$$

Sol:

$$L(a|x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} a x_i * \exp\left(\frac{-a x_i^2}{2}\right)$$

$$l(a|x_1,x_2\ldots,x_n) = n\log(a) + \sum_{i=1}^{n} -\frac{a}{2} x_i^2 + \sum_{i=1}^{n} \log(x_i)$$

$$\frac{dl}{da} = 0 \Rightarrow \frac{n}{a} - \frac{1}{2}\sum_{i=1}^{n} x_i^2 + 0 = 0$$

$$\Rightarrow \boxed{\hat{a} = \frac{2n}{\sum_{i=1}^{n} x_i^2}}$$

Q2.) What is a confidence interval? What is the difference between a one sided lower/upper confidence interval and a two-sided confidence interval.

Sol:- Confidence Interval is a statistic you calculate from a sample that tries to quantify a reasonable range of values for parameter of interest

$(1-\alpha)100\%$ confidence interval means that for $(1-\alpha)100\%$ of the samples, the true value will be contained in the confidence interval.

Two sided confidence bounds -

When we use a 2 sided confidence interval, we are looking at a closed interval where a certain percentage of the population is likely to lie i.e. we determine the values, b/w which a specified percentage of the population lies.

One sided confidence bounds

One sided confidence bounds are esentially an open ended version of 2 sided bounds. A one-sided bound defines the point where a certain percentage of the population is either higher or lower than the defined point.

An upper one-sided confidence bound defines that a certain percentage of population is less than.

A lower one sided bound defines a point that a certain percentage of population is greater than.

Lower: $\left[\bar{x} - Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \infty\right]$

Upper: $\left(-\infty, \bar{x} + Z_{\alpha} \frac{\sigma}{\sqrt{n}}\right]$

Q3.) What is a hypothesis test? Understand the following terms

→ Critical region
→ level (of a test)
→ p-value
→ power.

Sol:- Hypothesis test is a test where you try to see whether you have sufficient evidence to reject a hypothesis about ~~what~~ a parameter you are interested in, in your data.

Critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level α.

Level of Test → (α) – probability of rejecting the null hypothesis when it is true.

p-value – probability of observing sample data at least as extreme as the actually obtained test statistic.

power – Power is the probability of not making a Type-II error.

Q4.) Suppose I have a normally distributed data $x_1, x_2, \ldots, x_{10}$ with $\bar{x} = 1$ with unknown variance σ^2, but an estimate of $S^2 = 0.8$ from the data. How do I get a $(1-\alpha)100\%$ confidence interval (do not use bootstrap) with $\alpha = 0.05$? What if I wanted to test

$H_0 : \mu = 0$ at $\alpha = 0.05$
$H_A : \mu \neq 0$

Sol: $\bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}} = 1 \pm t_{0.025, 9} \cdot \frac{\sqrt{0.8}}{\sqrt{10}}$

$= 1 \pm 2.262 \frac{\sqrt{0.8}}{\sqrt{10}}$ $(0.36, 1.64)$

We can reject H_0, as its not in the confidence interval.

Q5.) Describe the bootstrap procedure (write out an algorithm in pseudo code)

Sol:-

Step 1: Resample data with replacement $x_1, x_2 \ldots, x_n$

Step 2: Create m bootstrap samples for $x_1, x_2 \ldots, x_n$

Step 3: Calculate MLE $\hat{\theta}$ for each bootstrap sample

Step 4: Calculate the lower and upper bound by sampling percentile to calculate the confidence interval

lower_bound = np.percentile($\hat{\theta}$, $\alpha/2$)

upper_bound = np.percentile($\hat{\theta}$, $100-\alpha/2$)

PART-IV : Machine Learning and scikit learn.

Q1) How do ridge regression and LASSO differ from the traditional linear regression? What is the difference b/w LASSO & Ridge?

Sol:- Linear Regression – The linear regression gives an estimate which minimises the sum of square error.

$$\text{Min} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2$$

$$\underset{\beta}{\text{Min}} \sum_{i=1}^{N} \left[y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2} \ldots \ldots - \beta_d x_{i,d} \right]^2$$

$$= \underset{\beta}{\text{Min}} \sum_{i=1}^{N} (Y - X\beta)^T (Y - X\beta)$$

Ridge Regression – The ridge regression gives an estimate which minimises the sum of squared error. as well as satisfy the constraint $\sum_{j=1}^{d} \beta_j^2 <= S$

$$\underset{\beta}{\text{Min}} \sum_{i=1}^{N} (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

Lasso Regression – Lasso regression gives an estimate which minimizes the sum of the squared error as well as satisfy the constraint

$\sum_{j=1}^{d} |\beta_j| <= S$ $\quad \underset{\beta}{\text{Min}} (Y - X\beta)^T (Y - X\beta) + \lambda \| \beta^T \|$

Q2) What kinds of problems do we use logistic regression for? What is the relationship b/w the covariates, regression coefficients and the predicted probabilities in logistic regression?

Sol:- We use logistic regression for binary classification problems.

$$X = \begin{bmatrix} \vec{x_1} \\ \vdots \\ \vec{x_n} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad y_i = 0 \text{ or } 1$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \vec{x_i}^T \beta = \sum_{j=1}^{d} x_{ij} \beta_j$$

$p_i = P(y_i = 1 \mid \vec{x_i})$ → predictors for unit i

observed y for person i

$$\frac{p_i}{1-p_i} = \exp(\vec{x_i}^T \beta)$$

$$p_i = (1-p_i)\exp(\vec{x_i}\beta)$$

$$\Rightarrow p_i = \frac{\exp(\vec{x_i}^T \beta)}{1 + \exp(\vec{x_i}^T \beta)}$$

Q3) Describe the procedure of Cross Validation?

Sol:- Cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

Randomly partition my training set into 'k' parts of approx same size i.e. k-folds.

Fit the model for each 'k' in the grid, k-times.

Algorithm –

1.) Shuffle the dataset randomly

2.) Split dataset into k groups

3.) For each unique group –
   a) Take the group as a held out or test data set.
   b) Fit a model on training set and evaluate it on the test set
   c) Fit a model on the training and evaluate on the test set
   d) Retain the evaluation score and discard the model.

4.) Summarize the skill of model using the sample of model evaluation scores.

Q4.) Describe at high level, how a binary tree works?
i.e. given a datapoint, how do you end up with a prediction?
What are the parts of a tree? Nodes, leaves etc

Q5.) What are boosting and bagging?
Boosting and Bagging come under ensemble methods.

Sol:- The general principle of ENSEMBLE METHODS is to combine the predictions of several models. These are built with a given learning algorithm in order to remove the robustness over a single model.

2 types -

- Parallel Ensemble Methods - In these methods, the base learners are generated in parallel simultaneously
eg:- when deciding the movie you want to watch, you may ask multiple friends for suggestions and probably watch the movie which got the highest votes.
- Sequential Ensemble Methods - In this technique, different learners learn sequentially with early learners fitting simple model to the data.

Then the data is analyzed for errors. The goal is to solve for net error from the prior model. The overall performance can be boosted by weighing previously mislabeled samples with higher weight.

Bagging - (Boot Strap Aggregating)

Bagging, a parallel ensemble methods, is a way to decrease the variance of the prediction model by generating additional data in the training stage. This is produced by random sampling with replacement from the original set.

These multisets of data are used to train multiple models. As a result we end with ensemble of diff. methods. The average of predictions from all the different models is used.

Boosting -

Boosting is a sequential ensemble methods that in general decreases the bias error and builds strong predictive models. The term 'Boosting' refers to family of algorithms which convert a weak learner to a strong learner.

In each iteration, data points that are mispredicted are identified and their weights are increased so that the next learner pays extra attention to get them right.

During training, the algorithm allocates weight to each resulting model. A learner with good prediction results on the training data will be assigned a higher weight than a poor one.

Q6.) What is a random forest and how does it differ from a binary decision tree?