Machine Learning :

* Machine Learning refers to a set of tools for understanding data. (or)
Machine Learning is programming computers to optimize a performance criteria using example data (or) past experience. (or)
Machine learning is all about using right features to build the right models that acheives the right tasks. (or)
Machine learning is all about extracting knowledge from data.
* Machine learning is also known as statistical learning (or) Predict to analytics.

History of Statistical Learning :

* At the beginning of 19th century the programs called least squares method was developed.
* Least square method is used for the implementation of linear regression.
* linear regresion is used to predict the quantitative values.
* In 1936, the linear Discriman Analysis was developed to predict qualitative values.
* In 1940, Logistic Regression was developed.
  - It is used for classification.
* In 1970, that have generalised linear models is the term and developed to describe the entire class of statistical learning methods.
(* Linear modes are used to generalise the linear relation).
* In 1980, there have developed the algorithm called the classification and Regression Tree (CART).
* In same 1980, neural networks are developed.
* In 1990, support vector machines were developed.

Applications

* we
* Medic
* Astro
* Public
* Statist
* Ident
* class
* ram.
* Predi
the bas
* cust
* Iden
* Cla
classes
* Este
variabl
* clas
* Types
* Statis
understa
* these
1.) su
2.) Un
3.) Re
1.) Sup
Supe
for
more
* ov
vari
* ve

Machine learning :

* Machine Learning refers to a set of tools for understanding data. (or)
Machine learning is programming computers to optimize a performance criteria using example data (or) past experience. (or)
Machine learning is all about using right features to build the right models that acheives the right tasks. (or)
Machine learning is all about extracting knowledge from data.
* Machine learning is also known as statistical learning (or) Predict to analytics.

History of Statistical Learning :

* At the beginning of 19th century the programs called least squares method was developed.
* Least square method is used for the implementation of linear regression.
* linear regresion is used to predict the quantitative values.
* In 1936, the linear Discriman Analysis was developed to predict qualitative values.
* In 1940, Logistic Regression was developed.
- It is used for classification.
* In 1970, that have generalised linear models is the term and developed to describe the entire class of statistical learning methods.
* Linear models are used to generalise the linear relation).
* In 1980, there have developed the algorithm called the Classification and Regression Tree (CART).
* In same 1980, neural networks are developed.
* In 1990, support vector machines were developed.

Applications

* We can us
* Medicine
* Astro physi
* Public pol
* Statistical

* Identify
* classify
- ram.
* Predict w
the basis o
* customize
* Identify
* classify
classes ba
* Establish
variables i
* classify

* Types of

* Statistical
understandi
* These te
i) supervi
2) Unsuper
3) Reinforc

1.) Supervise

Supervise
for pred
more in
* outcom
variable
* vector

Applications of Machine learning:

* we can use machine learning in Business.
* Medicine field.
* Astro physics.
* Public policy. etc---.
* Statistical learning problems:
* Identify the risk factors for prostate cancer.
* classify a recorded phoneme based on a log-periodog ram.
* Predict whether someone will have a heart attack, are the basis of demographic, diet and clenical, measurments.
* customize an e-mail spam detection system.
* Identify the numbers in a hand written zip code.
* classify a tissue sample into one of several cancer classes based on a gene expression profile.
* Establish the relationship between salary and demographic variables in population survey data.
* classify the pixels in a satellite image.

* Types of learning:-

* Statistical learning refers to a vast set of tools for understanding the data.
* these tools is classified into 2 categories.

1) supervised Statistical Learning.
2) Unsupervised Statistical learning.
3) Reinforcement Learning.

1) Supervised learning:-

Supervised learning involves building a statistical model for predicting (or) estimating the output based on one (or) more inputs with the supervising outputs.

* outcome measurment (Y) is also called as dependent variable, response variable (or) target variable.
* vector of P predictors measurments 'x' also called as

input variables, features, independent variables, predictors, covariates, regressors.

* In regression problem, Y is quantitative.
* In the classification problem, Y is qualitative.
* Examples for classification.
* Hand-written digits classification,
* Unsupervised learning:

Unsupervised learning involves building a statistical model for predicting (or) estimating the outputs based on one (or) more inputs without supervising outputs.

* what is statistical learning?

Notation.

Here, sales is a response (or) target that we wish to predict. We generally refer to the response as Y.

TV is a feature, or input, or predictor, we name it X_1, likewise name Radio as X_2, and so on.

we can refer to the input vector collectively as,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}.$$

Now we write our model as

$$Y = f(X) + \epsilon.$$

where ϵ captures measurement errors and other discrepancies.

~~* with a good f~~

What is $f(X)$ good for?

* with a good f we can make predictions of what new points $X = x$.
* we can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y, and which are irrelevant, e.g., Seniority and years of Education have a big impact on income, but Marital status typically does not.
* Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.

The Regression Function $f(x)$

Is there an ideal $f(x)$? In particular, what is a good value for $f(x)$ at any selected value of x, say $x = 4$? There can be many Y values at $x = 4$. A good value is

$$f(4) = E(Y|x=4)$$

$E(Y|x=4)$ means expected value (average) of Y given $x = 4$.

* this ideal $f(x) = E(Y|x=x)$ is called the regression function.

* The regression function $f(x)$

* It also defined for vector X_i; e.g.:

$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$.

* Is the ideal or optimal predictor of Y with regard to mean-squared prediction error: $f(x) = E(Y|X=x)$ is the function that minimizes $E[Y - g(X)]^2 | X = x]$ over all functions g at all points $X = x$.

18/7.

* How to estimate 'f'?

* To estimate 'f' it was classified into two approaches

1) Linear Approaches
2) Non-linear Approaches.

1) Training data.

2) A statistical method on this data to learn the actual input and output.

* To estimate the function 'f' we may follow parametric methods and (or) non-parametric methods.

1) Parameteric method.

In this we follow 2 steps.

i) we have to assume, the functional form of 'f' has linear. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

ii) In this we can find the β values.

* there is a dis-advantage, i.e.) the actual data is not a linear data.

* Advantage is we need not to bother about the functional form.

2.) Non-parametric form.

(* we must).

* How to estimate 'f'.

* To estimate function 'f' we may use linear approach (or) non-linear approach.

* We assume that we have observed a set of a different data points. (Training data). Because, we will use these observations to train our method how to estimate f.

* Let x_{ij} represents the value of jth predictor for the input observation 'i'. where $i = 1, 2, \dots n$; $j = 1, 2, \dots p$.

* y_i represents the response variable for ith observation, Then, the training data consists of $\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$.

* $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$. i.e., representation of the training data.

* therefore, our goal is to apply a statistical learning method to training data inorder to estimate the unknown function 'f'.

$$y \simeq \hat{f}(x)$$

* Now the statistical method for the above task can be characterize as either parametric (or) non-parametric.

* Parametric method:

Parametric methods involve a two step model based approach.

i) we make an assumption about the functional form of 'f'.

That assumption is 'f' is linear.

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

$x_1, x_2, \dots$ are predictors.

Once we have assumed that 'f' is linear. Now, the problem is to simplify to estimating parameters.

i) * After the m
that uses the
* The most
least squares
* Dis-advanta
i) The mode
true unknown
to far from
estimate will
* Non-para
* Non-para
about the f
* It estimates
points,
Ex: thin plat
* Assessing th
* Inorder to
learning m
measure
MSE = 1/n
* If MSE
actual out
* If MSE
actual out
* MSE is
* Model me
(or) statistica
* Always
* Training
* The MS

ii) * After the model has been selected, we need the procedure that uses the training data to fit or) to train the model.
* The most common approach to train the model is least squares approach.

* Dis-advantages of the parametric Approach :-

i) the model we choose will usually not natural form to true unknown form of 'f'. If the choosen model is to far from the true functional form of 'f'. Then, our estimate will be very poor, which needs to overfitting.

* Non-parametric methods :-

* Non-parametric methods do not make any assumption about the functional form of 'f'.
* It estimates the 'f' which is very close to the data points.
Ex: thin plate spline.

* Assessing the model :-

* Inorder to evaluate the performance of a statistical learning method on a given data set, we need to measure mean squared error. $\hat{y}_i = \hat{f}(x_i)$.

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

* If MSE is small, the predictions are very close to the actual outputs.
* If MSE is large, predictions are far away from the actual outputs.
* MSE is measure for regression problem.
* Model means it is the output of machine learning method (or) statistical learning method.
* Always MSE must be small.

* Training MSE :-

* The MSE is computed using the training data that way

used to fit the model.

* we would like to select the model for which the test MSE is as small as possible.

* The Bias - variance TRade off :-

* the expected test MSE for a given value x_0 can always be decomposed into 3 fundamental quantities,

1.) the variance of $\hat{f}(x_0)$

2.) the squared bias of $\hat{f}(x_0)$.

3.) the variance of error (~~Eount~~ term ~~curve~~) ϵ.

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + (\text{Bias}(\hat{f}(x_0))^2 + \text{Var}(\epsilon)$$

* variance :

* Variance is refers to ~~about~~ amount by which $\hat{f}$ will change if we estimate it using a different training data set.

* Bias :-

Bias refers to the error i.e., introduced by approximation of a real life problem.

* If we want to minimize the expected test MSE we need to select the model. That simultaneously acheives low variance and low bias.

21/7/22

* simple linear regression using single predictor x :-

* we assume a model

$$Y = \beta_0 + \beta x + \epsilon$$

β_0, β_1 are 2 unknown constants that represent and slope also known as coefficients (or) parameters and ϵ_0 is error term (or) term.

* Given some estimates $\hat{\beta}_0$ & $\hat{\beta}_1$ for model coefficients we predict future sales when using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

* where $\hat{y}$ indicates a prediction of y is the basi

21/7/22

Supervised Learning:

1) **Linear Regression**: Linear Regression is a simple approach to supervised learning. It assumes that the dependencies of y on $x_1, x_2, \ldots, x_p$ is linear.

* True regression function never a linear.

simple linear regression using a single predictor x

* we assume a model

$$y = \beta_0 + \beta_1 x + \epsilon.$$

where, β_0, β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or the parameters and ϵ is the error term.

* Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of y on the basis of $x = x$. then that symbol denotes an estimated value.

Residual: The difference between actual output and the Predicted output caused Residual.

* **Estimation of the parameters by Least squares**:

* Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for y based on the ith value of x. then $e_i = y_i - \hat{y}_i$ represents the ith residual.

* we define the residual sum of squares (RSS) (or)

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2.$$

or equivalently as

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

* the least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

where, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

* **Assessing the accuracy of the coefficient estimates:**

* The standard error of an estimator reflets how it varies under repeated sampling we have,

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

where, $\sigma^2 = \text{var}(\epsilon)$

* These standard error can be used to compute confidence intervalse A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

that is, there is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like tree parent sample).

* For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$.

* **Hypothesis testing:**

* Standard Error can also be used to performed hypotheses tests on the coefficients. The most common hypothesis tests involves testing the null hypothesis of

H_0 : there is no relationship between X and Y, versus the alternative hypothesis.

H_A : There is some relationship between X and Y.

* Mathematically this corresponds to testing

$$H_0 : \beta_1 = 0.$$

versus

$$H_A : \beta_1 \neq 0.$$

Since, if $\beta_1 = 0$ then the model reduces to $y = \beta_0 + \epsilon$ and X is not associated with Y.

* To test

$$t = \frac{\beta_1 -}{SE}$$

* This will

assuming β_1

* Using stati

of observing

the p-value

* R- Code

library (M

library (I

Ad ← re

head (A

lm.fit ←

attach (

lm.fit <

lm.fit

Summar

names

coef (l

conf. int

Predict

plot (

abline

* To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}.$$

* This will have a t-distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$.

* Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ (or) larger we call this probability the p-value.

* R-Code for Linear Regression :-

```
library (MASS)
library (ISLR2)
Ad ← read ("Advertising.csv", header = TRUE)
head (Ad)
lm.fit ← lm (sales ~TV, date = Ad)
attach (Ad)
lm.fit ← lm (sales ~TV)
lm.fit
Summary (lm.fit)
names (lm.fit)
coef (lm.fit)
conf.int (lm.fit)
Predict (lm.fit, data.frame (TV=c(200, 150, 90)), interval "
                                                  Confidence")
plot (TV, sales)
abline (lm.fit, lwd = 3, col = "red")
```

* To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

* This will have a t-distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$.

* Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ (or) larger we call this probability the p-value.

* R-Code for Linear Regression :-

```
library (MASS)
library (ISLR2)
Ad ← read ("Advertising.csv", header = TRUE)
head (Ad)
lm.fit ← lm (sales ~TV, data = Ad)
attach (Ad)
lm.fit ← lm (sales ~TV)
lm.fit
Summary (lm.fit)
names (lm.fit)
coef (lm.fit)
confint (lm.fit)
Predict (lm.fit, data.frame (TV=c(200, 150, 90)), interval = "Confidence")
plot (TV, sales)
abline (lm.fit, lwd = 3, col = "red")
```

26/7/22

* Assessing the overall Accuracy of the Model :-

* we compute the Residual Standard Error.

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2},$$

where, the residual sum-of-squares is $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

* R-Squared or fraction of variance explained is

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where, $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.

* It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between x and y:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Quantity	Value
Residual standard Error	3.26
R^2	0.612
F-Statistic	312.1

* Multiple linear Regression :-

* It is like simple linear regression, but number of inputs will more than 1.

* Here our model is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

* Multiple linear Regression is supervised learning approach for predicting quantitative response based on more than one input.

* It can be model as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.

* we interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

sales $= \beta_0 + \beta_1 \times$ TV $+ \beta_2 \times$ radio $+ \beta_3 \times$ newspaper $+ \epsilon$

```
library (MASS)
library (ISLR2)
Ad <- read.csv ("Advertising.csv", header = TRUE)
head (Ad)
lm.fit <- lm (sales ~ TV + radio + newspaper, data = Ad)
attach (Ad)
lm.fit <- lm (sales ~ TV)
```

(lm. fit)
summary (lm
Output:
coefficients:
(Intercept)
TV
radio
newspaper
Correlations:
TV
radio
newspaper
Sales
* Is at le
* For
F = (TSS
RS
* F- stat
* If F- S
input and
Quantity
RS-E
R^2
f-statist

(lm. fit)

summary (lm.fit)

Output:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.938889	0.311908	9.422	< 2e-16	***
TV	0.045765	0.001395	32.809	< 2e-16	***
radio	0.188530	0.008611	21.893	< 2e-16	***
newspaper	-0.001037	0.005871	-0.177	0.86	

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

* Is atleast one predictor useful.
* For 1st one use F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p, n-p-1}$$

* F-statistic should be large.
* If F-statistic = 1, then their is no relation between input and output variable.

Quantity	value
RS-E	1.69
R^2	0.897
f-statistic	570

29/9/22 * Deciding on the important variables :-

* the most direct approach is called all subsets or best subsets regression : we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
* However we often can't examine all possible models, since they are 2^p of them ; for example, when $p = 40$ they are over a billion models!
* Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.
* Forward selection :-
* Begin with the null model - a model that contains an intercept but no predictors.
* Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
* Add to that model the variable that results in the lowest RSS amongest all two-variable models.
* Continue until some stopping rule is satisfied, for example, when all remaining variables have a p-value above some threshold.
* Backward selection :-
* Start with all variables in the model.
* Remove the variables with the largest p-value - that is, the variable that is the least statistically dependent significant
* the new (p-1) - variable model is fit, and the variable with the largest p-value is removed.
* Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

* Other consider
* Qualitative p
* Some predi
taking a d
* These are
variables.
(* See for ex
card.).
Example: In
males and fe
a new varia

x_i

Resulting mo
$y_i = \beta_0 +$
(Results for
* Qualitative
* with more
variables. For
two dummy

x_{i1}

and the

x_{ic}

* then both
regression eq
$y_i = \beta_0 + \beta_1 x_{i1}$

* Other considerations in the regression model:-

* Qualitative predictors:-

* Some predictors are not quantitative but are qualitative, taking a discrete set of values.

* These are also called categorical predictors or factor variables.

(* See for example the scatterplot matrix of the credit card.).

Example: Investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male.} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

(Results for gender model).

* Qualitative predictors with more than 2 variables:

* with more than two levels, we create additional dummy variables. For example, for the ethinicity variable we create two dummy variables. the first could be,

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is caucasian} \\ 0 & \text{if } i\text{th person is not caucasian.} \end{cases}$$

* then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

* there will always be one fewer dummy variable than the number of levels. The levels with no dummy variable - African American in this example - is known as the base line.

3/4/22

* Extensions of the linear model :-

* In linear models,
* the additive assumption implies that the effect of changes in a predictor x_j on the response Y is independent of the values of the other predictors.
* linear assumption implies that the change in the response 'Y', due to a '1' unit change in x_j is constant, regardless of the value of x_j.

* Interactions :-

* In our previous analysis of the Advertising data, we assumed that the effect on Sales of increasing one, the advertising medium is independent of the amount spent on the other media.
* For example, the linear model

$$Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper.$$

states that the average effect on Sales of a one-unit increases in TV is always β_1, regardless of the amount spent on radio.

* Modelling interactions

Model takes the form

$$Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon$$

$$= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon.$$

Results :-

	Coefficient	Std. Error	t-Statistic	P-value
Intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
radio	0.0289	0.009	3.24	0.0014
TV × radio	0.0014	0.000	20.73	<0.0001.

* Hierarchy principle :-

If we include an interaction in a model, we should do include the main effects, even if the p-values associated with their coefficients are not significant.

* Interaction between qualitative and quantitative variables :-

Consider the credit data set, and suppose that we wish to Predict balance using income (quantitative) and student (qualitative).

* without an interaction term, the model takes the form

$$\text{balance}_i \simeq \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i^{th} \text{ person is a student} \\ \beta_0 & \text{if } i^{th} \text{ person is not a student.} \end{cases}$$

with interactions, it takes the form

$$\text{balance}_i \simeq \beta_0 \times \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_2 \times \text{income}, & \text{if student} \\ 0 & \text{if not student.} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}, & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}, & \text{if not student.} \end{cases}$$

4/8/22

* Parametric Methods versus Non-Parametric Methods :-

* A non parametric method akin to linear regression is K-nearest neighbours regression which is closely related to the K-nearest neighbors classifier.

* Given a value for K and a prediction point x_0, K-nearest neighbors regression first identifies the K-observations that are closest to x_0, represented by N_0. $f(x_0)$ is then estimated using the average of N_0; like so

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i.$$

* A parametric approach will out perform a non-parametric approach if the parametric form is close to true form of $f(x)$.

* the choice of a parametric approach versus a non-parametric approach will depend largely on bias-variance trade-off & shape of function $f(x)$.

* when the true relationship is linear, it is difficult for a non-parametric approach to complete with the linear regression because non-parametric approach incurs a cost in variance that is not offset by a reduction in bias.

* Additionally, in higher dimensions, K-nearest neighbors regression often performs worse than linear regression.

* This is often due to combining too small an n with too large a p, resulting in a given observation having no nearby neighbors.

* This is often called the curse of dimensionality. In other words, the K-observations nearest to an observation may be far away from x_0 in a p-dimensional space when p is large, leading to a poor prediction of $f(x_0)$ and a poor K-nearest neighbors regression fit.

* As a general rule, parametric models will tend to outperform non-parametric models when there are only a small no. of observations per predictor.

* classification

Set C, such

eye color

email ∈ {

* Given a

taking values

build a fu

vector x an

* Often we

that x belo

* For exam

of the probe

than a cl

10/8/22 * Logis

* Logistic regr

particular cat

* Logistic re

Prediction

form

P(X

* This yields

* the logist

* $\frac{P(X)}{1-P(X)}$

between

* As an

$\frac{1}{4}$ since

* Taking

* Classification:- Qualitative variables takes values in an ordered set C, such as:
eye color ∈ {brown, blue, green}.
email ∈ {spam, ham}.

* Given a feature vector X and a qualitative response Y taking values in the set C, the classification task is to build a function C(X) that takes as input the feature vector X and predicts its value for Y : i.e., C(X) ∈ C.

* Often we are more interested in estimating the probabilities that X belongs to each category in C.

* For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

10/8/22

* Logistic Regression:-

* Logistic regression models the probability that y belongs to a particular category rather than modeling the response itself.

* Logistic regression uses the logistic function to ensure a prediction between 0 and 1. The logistic function takes the form

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

* This yields a probability greater than 0 and less than 1.

* The logistic function can be rebalanced to yield

$$\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X},$$

* $\frac{p(X)}{1-p(X)}$ is known as the odds and takes on a value between 0 and infinity.

* As an example, a probability of 1 in 5 yields odds of $\frac{1}{4}$ since $\frac{0.2}{1-0.2} = \frac{1}{4}$.

* Taking a logarithm of both sides of the logistic odds equation yields an equation for the log-odds or logit.

$$\log\left[\frac{P(X)}{1-P(X)}\right] = \beta_0 + \beta_1 X.$$

logistic regression has a logit that is linear in terms of X.

* Unlike linear regression where β_1 represents the average change in Y with a one-unit increase in X, for logistic regression, increasing X by one-unit yields a β_1 change in the log-odds which is equivalent to multiplying the odds by e^{β_1}.

* The relationship between P(X) and X is not linear and because of this β_1 does not correspond to the change in P(X) given one-unit increase in X. However, if β_1 is positive, increasing X will be associated with an increase in P(X) and, similarly, if β_1 is negative, an increase in X will be associated with a decrease in P(X). How much change will depend on the value of X.

Estimating Regression Coefficients:

* logistic regression uses a strategy called maximum likelihood to estimate regression coefficients.

* Maximum likelihood plays out like so: determine estimate for β_0 and β_1 such that the predicted probability of $\hat{P}(X_i)$ corresponds with the observed classes as closely as possible. Formally, this yield an equation called a likelihood function.

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} P(X_i) \times \prod_{j:y_j=0} (1 - P(X_j)) =$$

* Estimates for β_0 and β_1 are choosen so as to maximize this likelihood function.

* Linear regression's least squares approach is actually a special case of maximum likelihood.

* logistic regression measures the accuracy of coefficient estimates using a quantity called the z-statistic. The z-statistic is similar to the t-statistic. the z-statistic for β_1 is represented by

17/8/22

* Discriminant

	Outlo
0	Rainy
1	Rainy
2	over
3.	Sun
4.	Sun
5,	Su
6.	ov
7.	Ra
8.	R
9.	S
10.	R
11.	O
12.	O
13.	S

Yes - 9

Outlook:

Rainy - 5

overcast - 4

sunny - 5

P(Y=X) =

terms of X.
average
for logistic
β_1 change
lying the

linear and
ange in
if β_1 is
increase
increase
$P(X)$.
X.

m likelihood

ine estimate
of $\hat{P}(x_i)$
as possible.
lihood

to maximize

ctually a

coefficient
HISTIC. The
e z- statistic

z- statistic $(\beta_1) = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.

17/8/22

* Discriminant Analysis :-

	Outlook	temperature	Humidity	windy	play golf
0	Rainy	Hot	High	False	NO
1	Rainy	Hot	High	True	NO
2	overcast	Hot	High	False	yes
3.	sunny	Mild	High	False	yes
4.	sunny	cool	normal	False	yes
5.	sunny	cool	normal	True	no
6.	overcast	Cool	normal	True	yes
7.	Rainy	mild	high	False	no
8.	Rainy	cool	normal	False	yes
9.	Sunny	Mild	normal	False	yes
10.	Rainy	mild	normal	True	yes
11.	overcast	mild	high	True	yes
12.	overcast	Hot	normal	False	yes
13.	sunny	mild	High	True	~~yes~~ NO

Yes - 9 ; NO - 5

Outlook :-

Rainy - 5 — 2, 3

overcast - 4 — 4, 0

sunny - 5 — 3, 2

$$P(Y=x) = \frac{P(x=Y)\ P(Y)}{P(x)}$$

x = Sunny.

$$P(Y=Yes/X) = \frac{P(X/Yes.) * P(Yes)}{P(X)}$$

$$= \frac{\frac{3}{9} \times \frac{9}{14}}{\frac{5}{14}} = 0.6$$

$$P(Y=NO/X) = \frac{P(X/NO) * P(NO)}{P(X)} = \frac{\frac{2}{5} \times \frac{5}{14}}{\frac{5}{14}} = 0.4$$

Humidity

High - 7 - $\frac{7}{14}$ — 3 / 4

Normal - 7 - $\frac{7}{14}$ — 6 / 1

$P(High/Yes) = \left(\frac{3}{9} \times \frac{9}{14}\right) / \frac{7}{14} = \frac{3}{7} = 0.428$

$P(High/NO) = \left(\frac{4}{5} \times \frac{5}{14}\right) / \frac{7}{14} = \frac{4}{7} = 0.572$

* let us consider we have data of 13 observations (0-12).
* By using that data we need to predict the response of playgolf on features or predictors given. Yes - 9 ; NO - 4.

Outlook:

Rainy - 5 — 2 Y / 3 N

Overcast - 4 — 4 Y / 0 N

Sunny - 4 — 3 Y / 1 N

Humidity:

High - 6 — 3 Y / 3 N

Normal - 7 — 6 Y / 1 N

Temperature

hot - 4 — 2 Y / 2 N

Mild - 5 — 4 Y / 1 N

Cool - 4 — 3 Y / 1 N

windy

True - 5 — 3 Y / 2 N

mild - 6 — 4 / 2.

UNIT-III

* Resampling methods :- (to get additional information we are refitting).

Resampling methods involves repeatedly drawing samples from the given data, and refit the model on each sample to get additional info about the model.

* There are 2 resampling methods.

1.) cross validation.

2.) Boot strap.

1.) Cross validation :- It measures the test error rate associated with a statistical learning method to evaluate the model performance.

2.) Boot strap :- It measures the accuracy of parameter estimation (or) of given model.

* validation set :- Before going to know about cross validation we need to know about validation set.

* the data is divided into 2 groups.

1.) train error.

2.) testing error.

* (How to implement K-fold cross validation ?)

* Cross validation :-

Training Error versus Test error :-

* Recall the distinction between the test error and the training error.

* The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.

* In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.

* But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

* More on prediction - error estimates :-

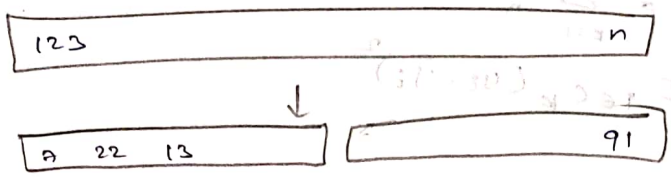
* Best solution : a large designated test set. often not available.

* Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate. These include the Cp statistic, AIC and BIC. (Ark of information criteria") - They are discussed elsewhere in this course.

* Here, we instead consider a class of methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

* Validation set Approach :

* Here, we randomly divide the available set of samples into two parts : a training set and a validation or hold-out set.
* The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
* The resulting validation-set provides an estimate of the test error. This is typically assessed using MSE in the case of quantitative response and misclassification rate in the case of qualitative (discrete) response.
* The validation process :



A random splitting into two halves : left part is training set, right part is validation set.

* Drawbacks of validation set approach :

* The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
* In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
* This suggests that the validation set error may tend to overestimate the test error for the model fit on the entire data set.

* K-fold Cross-Validation :

* widely used approach for estimating test error.
* Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.

* Idea is to randomly divide the data into K equal-sized parts. We leave out part K, fit the model to the other K-1 parts (combined) and then obtain predictions for the left-out Kth part.

* This is done in turn for each part $k = 1, 2, \ldots, K$ and then results are combined.

* The details :

* Let the K parts be $C_1, C_2, \ldots, C_K$, where C_k denotes the indices of the observations in part K. There are n_k observations in part k : if N is a multiple of K, then $n_K = n/K$.

* Compute :

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k,$$

where, $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2$