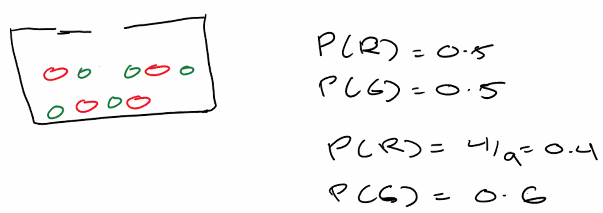
**Probability.**

****

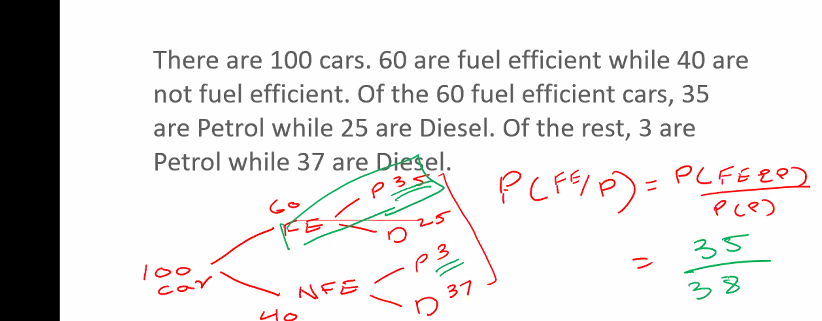
**Disjoint one time one color dependent since one used next one used will be different**

**Two cars can have same colors and one car color is green has no effect on the other car’s cars so independent**

**Contingency table**

**B given that C**

**Mutually exclusive in probability tree the last probability.**



* NFE|D=37/62

First of all we make probability tree

This HIV example is not mutally exclusive we need to convert it into the mutually exclusive.

**Machine learning Process**

ETP process is in machine learning

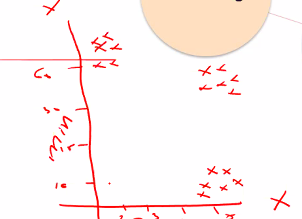
Machines need historical data,algorithms and testing.

Dimension reduction.

2000 data points is too much to consider so it shortlists top 20 variables.

Clustering

To find a set of users from overall base.



We need to pick the right option package for the right customer.

We divide overall base into a no of clusters in this case 3.We each unique cluster with different offer.

We find the performance of each particular cluster using performance metrics.

Association

A person who buys beer also buys pamper.Checking different demographics.

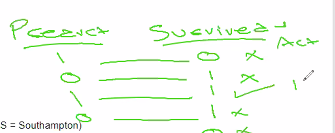
Regression.

We always predict numerical values through regression.

Classification.

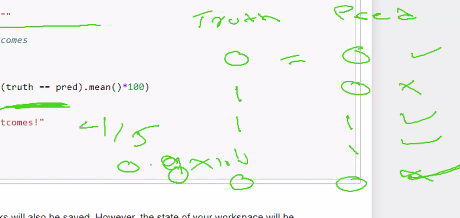
We use this to predict categorical values.

We provide the target variable in supervised learning.In unsupervised there is no label or target.



* We will save survive in another column

We will check the accuracy of the algorithm

2f means two decimal points

In ml x is always exploratory variable and y is target dependent variable

Y=label=dependent variable=target variable=response variable

X=independent variable=exploratory variable=features=predictors.

Test score cannot exceed training score Training score is always more.

Random state keeps the data consistent in testing and training.

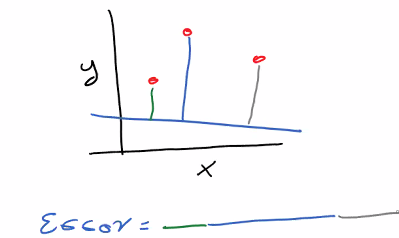
Random state must be same if data is shared.

The yellow line is helping us determine the y with respect to x

Best fit means distant between points and line is minimum best fit means accurate prediction.

Machine keeps fitting the line and optimization till best fit Is achieved and error is minimum.

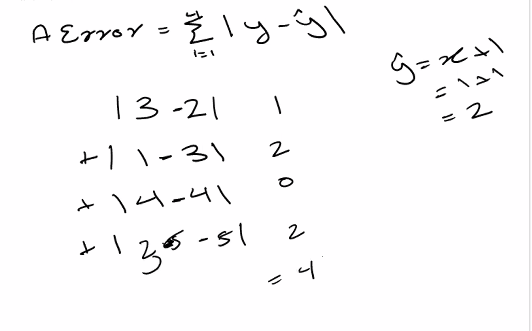
Gradient descent decides where error is minimum.



Local minimum=optimum point.

Absolute is to compensate for negative values.

We square because we need to penalize the error and compensate for negative values.



We use a constant such as 0.1 to control gradient descent in order to ensure that the line does not pass the optimum point.So order to ensure that the line does not move too quickly.

Y intercept is by default True.

False we want to normalize data ourselves not automatically.

N\_jobs is for parallel processing.

Gradient descent means line goes from maximum error to minimum error.

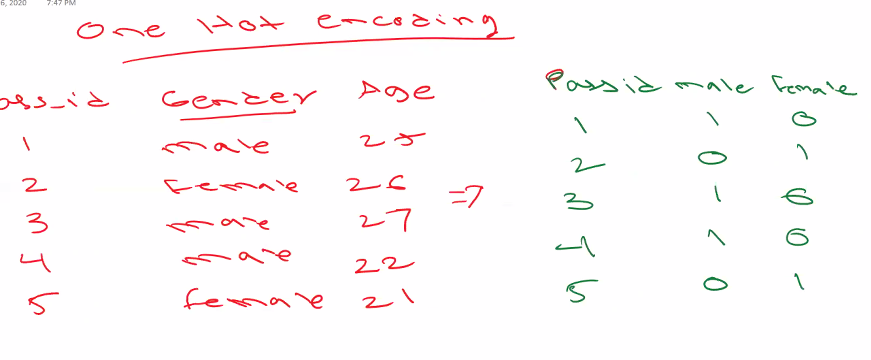
We scale the dependent variables. We do preprocessing.

Two types of scaling numerical and categorical.

1 will be the one with the max value(Feature scaling)

Min max scaler is used the most(we have library) we can use arguments to define min and max value.

We normalize numerical.we use one hot encoding on categorical.



Label encoding assigns alphabetically.



R2 score close to 0 poor.IF we take simple mean it gives same result so it is poor.

Model is not a linear regression model or is not properly fit these are the two problems.

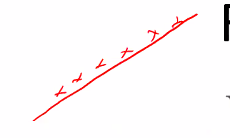
R2 score limitation is that no of features increases R2 score for no reason.

Another limitation is that it is sensitive to outliers.

R2 is for variable adjusted R2 is for which features have positive impact.

R2 score is used for variance.

Deterministic points align perfectly and can be predicted

Solistic cannot be predicted( usually the case not straight line of data points)

First case determined pattern covered only exceptional cases left.

Red dots are the errors and errors should not have a pattern.

Points should be near mean for good residual plot

Not all the features are as important .

We can remove features in arguments.arguments defines how many features

Linear regression is not scable only applicable to trained model.

<https://scikit-learn.org/0.16/datasets/index.html>

we change names to easier names to understand and represent.

Target is value we want to predict.

Skewness is very important in data science interview.

The type of transform returned is 2-d numpy array that is why we transform to dataframe.

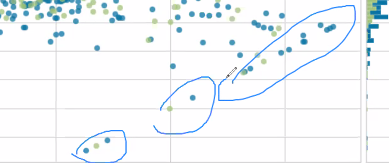
Scaling is not done on the response variable.

Trained data must have same no of rows and test data must have same no of rows.

Helper library given below.ways to program efficiently.

Pattern at the end of residual plot straight line right bottom.

In residual plot the straight line is y^ the regression line.



Np haystack converts 2-D numpy array to 1-D array.

True=selected false=not selected

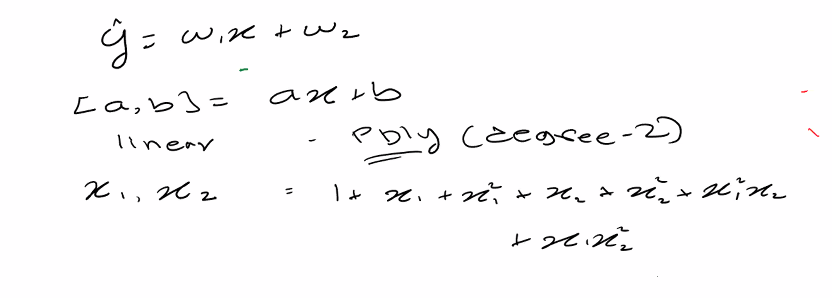
Delta of 2% between R2 and Adjusted R2 is acceptable.

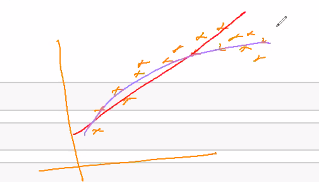
Every fold has different evaluation R2 scores should be consistent.

Default scorer is R2.

Step 1 means remove 1 feature at each iteration.

**Polynomial regression tasks 1-8 Do at home.**

****

****

We must consider all the possible combinations of degree 2.or higher degrees.

Model accuracy improved and data R2 is improved with polynomial in this case.

.values is multidimensional array.

Simple problem complex problem data linear but we are using polynomial degree(3) (overcomplicate)

Complex problem simple solution using linear regression on sinusoidal solution.(oversimplify)

Underfit=high bias

Overfit=high variance.

Ideal model=biasness variance tradeoff.=low variance low bias.

Regularization

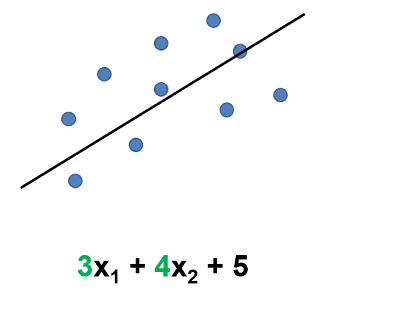
Used to improve model performance and to prevent overfitting.

We will consider both error and complexity.

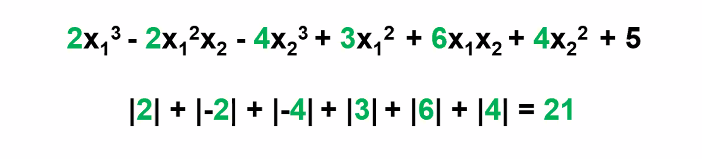
The green are the coefficients. We add coefficient to the error.

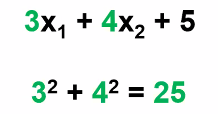
If we just look at error it will select wrong model. But in this case the generalized model is the linear line.

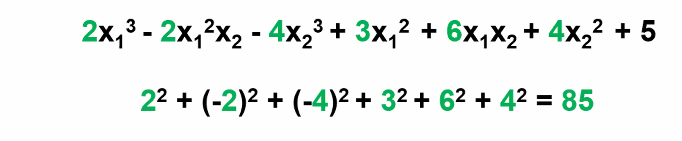
We penalize the complexity











We take absolute in lasso and square in ridge.

L1(lasso) is also used for feature selection.The feature that doesn’t make an impact it assigns it zero.

L1 is computatively expensive but L2 is not expensive.

L2 not sparse L2 is faster if sparse L1 is faster.

We use lambda to control the complexity punishment.



Small lambda

Complex model is selected

If model complexity is increased error due to variance is more. Error due to bias is less.

If model complexity is less error due to bias is more.(yellow one).

In large lambda

Model complexity is penalized. Simple model is selected.

Alpha parameter decides model complexity.

**Put three projects in resume.**

When we communication we use R2 score for testing.

Instead of giving it value one by value we give it range.