SML

U-2

① Simple Linear Regression

→ one dependent var y predicted from an indepnt var X.

→ r^2: proportion of variation in dependent var X predictable from Y

→ Represented as: $Y = a + bX$

→ slope gives the change in Y for a unit change in X

→ used in: economics, finance,

Multiple Linear Regression

→ one dependent var y predicted from a set of independent var $(X_1, X_2, \ldots, X_k)$

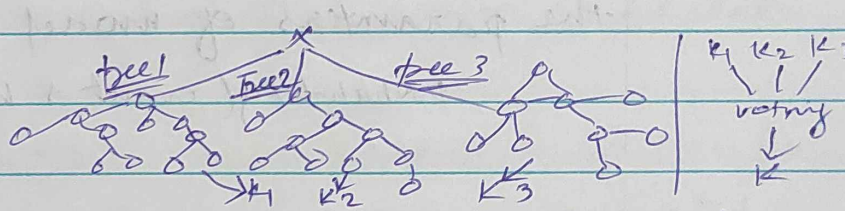
→ R^2: proportion of variation in dependt var y predictable from set of indepnt var. (x's)

→ Represented as: $Y = a + b_1X_1 + b_2X_2 + \cdots b_nX$

→ slope gives the change in Y for a unit change in each var.

→ used in marketing, health care

② Random forest



→ Powerful technique used in data science field for solving various problem.

→ used for both classification & regression

→ Ensemble Method: It is an ensemble method that combines multiple decision trees to create more accurate model

→ Decision Tree: Each tree in random forest is decision tree, which makes sequence of decision to arrive at a result.

→ It uses feature randomness to construct tree. This means only random subset of features is selected

→ Bagging involes creating multiple random samples of dataset on each sample. final prediction is made by avg. of all predict. trees.

③ Logistic regression & its advantages

→ It's a statistical method used to model the relationship b/w binary outcome (0,1, true, false) and predictor var.
→ It uses logistic function to transform the output of linear regression model into probability values b/w 0 and 1
→ Used in ML and data science for classification task like predicting whether email is spam or not.
→ It is trained using maximum likelihood estimation
→ Max. Likelihood estimation is the process of estimating the parameters of model.

probability(event + non-event) = 1

Adv

1. easily implemted & interpreted
2. can easily extend to multiple classes.
3. very fast at classification
4. Good Accuracy
5. less over-fitting

Information Value (IV)
Akaike info. criteria (AIC)
Receiver operating chas. (ROC)
Rank ordering
C-statistics
K-S Statistics.

④ Random forest using Germen credit data

- popular ML algo, used for classification & regression
- uses decision tree to improve accuracy.

For German credit data, the dataset contains:

(1) status of existing acc
(2) credit history
(3) purpose
(4) saving acc.
(5) employment duration
(6) property
(7) installment_plans
(8) housing
(9) Job
(10) Telephone

(1) Lets import some packages for the German credit data

```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
credit_data = pd.read_csv("credit_data.csv")
```

(2) Create the dummy var.

```
>> dummy_stat = pd.get_dummies(credit_data['Status of existing data'],
   prefix = 'stat_of_exist_acc')
>> dummy_ch = pd.get_dummy(credit_dat["Credit_history"],
   prefix = "Cred_hist")
```

— Similarly for all the datasets —

(3) Remove extra dummy var: of all categories

credit_data_new = pd.concat ([dummy_stat, dummy_ch, dummy_purpose, ...])

(4) Split the data into 70-30

x_train, x_test, y_train, y_test = train_test_split (credit_data_new.drop(['class'], axis=1), train_size=0.7, random_state=42)

(5) Assume hyperparameter values:

no. of trees is 1000
Max. decision tree can grow is 100
Min obs. req is 3
Min no. of obs in node is 2

and Apply these into the models.

(6) Result:

Random-forest Train CM

Predicted	0	1
Actual		
0	501	0
1	18	181

Accuracy 0.974

Random forest Test CM

Predicted	0	1
Actual		
0	270	1
1	43	44

Accuracy 0.855

(5) Compare ridge and Lasso regression models

Ridge Regression Model	Lasso Regression Model
→ Uses penalty regression in L2 norm	→ uses penalty regression in L1 norm
→ shrinks the coeffcent towards zero, but rarely exact zero	→ shrinks the coeffcent exactly to zero
→ It increases the bias of the model	→ It decreases the bias of the model.
→ uses closed form algo to solve the problem	→ uses iterative optimization algo to solve the problem.
→ If 2 or more predictors are highly corelated, Ridge Reg. will shrink their coeffcents	→ while Lasso can choose one of them & set remaing to zero
→ Handling dataset is difficult	→ Handling dataset is easy

(6) Grid search on Random forest

→ While finding soln with Random forest, we get less accuracy. To increase the accuracy we will use Grid Search

→ It is performed by changing hyperparameters for the model.

→ These Hyperparmiter models are:

No. of Trees (1000, 2000, 3000)

Max depth (100, 200, 300)

Min sample split (2, 3)

Min sample in leaf node (1, 2)

→ It is used to get the best performe of the model.

→ How to perform

(1) Define grid hyperparmeters

(2) Train Random forest models.

(3) Evaluate performe of each model

(4) Select the combi. of best hyperparmeters on performe

(5) Test the final model.

(7)

x_i	y_i
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

To apply linear regression, plot the data to find the relationship b/w x and y.

So, $y = mx + b$

m is slope of line

b is y intercept.

$$m = \left(n\Sigma x.y - \Sigma x.\Sigma y\right) / \left(n\Sigma x^2 - (\Sigma x)^2\right)$$

$$b = (\Sigma y - m\Sigma x)/n$$

n is no. of data pnts.

Σxy sum of product of x and y

Σx sum of x values

Σy sum of y values

Σx^2 sum of squared x values.

$n = 5$

$\Sigma x = 15$

$\Sigma y = 12.6$

$\Sigma xy = 44.4$

$\Sigma x^2 = 55$

$m = (n\Sigma xy - \Sigma x \Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$

$= (5 \times 44.4 - 15 \times 12.6) / (5 \times 55 - 225)$

$= 33 / 50 = 0.66$

$b = (\Sigma y - m \Sigma x) / n$

$= (12.6 - 0.66 \times 15) / 5$

$= 2.7 / 5 = 0.54$

So, eqn becomes $y = mx + b$

$y = 0.66x + 0.54$

for 7th week sales, $x = 7$

$y = 0.66 \times 7 + 0.54$

$= 5.16$

for 9th week sales, $x = 9$

$y = 0.66 \times 9 + 0.54$

$= 6.48$

$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$ = [redacted]

(8)

Sno	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	"	Cool	Normal	Weak	Yes
6	"	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

Overall prob. of yes

$P(Y) = 9/14$

Overall prob. of no

$P(N) = 5/14$

Outlook:	Y	N	P(Y)	P(N)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
Rainy	3	2	3/9	2/5

Temperature:	Y	N	P(Y)	P(N)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5

Humidity:	Y	N	P(Y)	P(N)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5

Windy:	Y	N	P(Y)	P(N)
weak	6	2	6/9	2/5
strong	3	3	3/9	3/5

$P(X/play = yes) \cdot P(play = yes)$

$\Rightarrow 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14$

$\Rightarrow \frac{486}{91854} = 0.00529$

$P(x/play = no) \cdot P(play = no)$

$\Rightarrow 3/5 \times 1/5 \times 3/5 \times 4/5 \times 5/14$

$\Rightarrow 180/8750 = 0.0205$

$$P(x) = P(yes) + P(no)$$
$$= P(x/play=yes) = P(x/play=no)$$
$$= 0.00529 + 0.0205$$
$$= 0.0253$$

Now, ~~$P(play=yes/x)$~~

$$P(play=yes/x) = \frac{P(x/play=yes) \cdot P(play=yes)}{P(x)}$$
$$= \frac{0.0053}{0.0253} = 0.209$$

$$P(play=No/x) = \frac{P(x/play=no) \cdot P(play=NO)}{P(x)}$$
$$= \frac{0.0205}{0.0253} = 0.810$$

P(NO) is high

∴ The player cannot play.