

## The Medium App

An app designed for readers.

[OPEN IN APP](#)

~~selected neurons are ignored during~~

training.

## Handling Underfitting:

- Get more training data.
- Reduce the capacity of the network.
- Add weight regularization.
- Add dropout.

With these techniques, you should be able to improve your models and correct any overfitting or underfitting issues.

## Connect with me on:



# ← Question



Håkon Hapnes Strand, Data Scientist



Answered Mar 11, 2018

The idea behind cross-validation is the same as with a single holdout validation set, to estimate the model's predictive performance on unseen data. Cross-validation simply does this more robustly, by repeating the experiment multiple times, using all the different parts of the training set as validation sets. This gives a more accurate indication of how well the model generalizes to unseen data.

In other words, cross-validation does not prevent overfitting in itself, but it may help in identifying a case of overfitting.

6.5k views · View Upvoters



Upvote · 37



Share



...



Comment...

Recommended All

Sponsored by BITS Pilani

...

Welcome a promising career in

• • •

12:22 AM

File Edit Search Source Run Debug Consoles Projects Tools View Help

Editor - C:\Users\Imran\Desktop\python\Python\Class2\fittransform.py

```
19 imputer = Imputer(missing_values='NaN', strategy='mean', axis=0)
20
21 imputer.fit(df[['Age', 'Salary']])
22 X = imputer.transform(df[['Age', 'Salary']])
23
24 imputer.fit_transform(df[['Age', 'Salary']])
25
26 #####
27
28 encode = LabelEncoder()
29
30 encode.fit(df['Country'])
31 encode.transform(df['Country'])
32
33 encode.fit_transform(df['Country'])
34 #####
35
36 import numpy as np
37 from sklearn.preprocessing import StandardScaler
38
39 x1 = np.array([[1,2,3],
40                 [4,5,6],
41                 [7,8,9]])
42
43 standscaler = StandardScaler()
44
45 x_scaler = standscaler.fit_transform(x1)
46 print(x_scaler)
47
48 """ (Xi - Xmean) / (standard Deviation of that feature) """
49
50 standscaler.fit(x1)
51 standscaler.transform(x1)
52
53
54
55
56
```

Variable explorer

Name	Type	Size	Value
X	float64	(20, 2)	array([[ 3.4000000e+01, 7.2000000e+04], [ 4.2000000e+01, ...])
df	DataFrame	(20, 8)	Column names: Country, Age, Gender, Occupation, Employment Status, Emp ...
x1	int32	(3, 3)	array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
x_scaler	float64	(3, 3)	array([-1.22474487, -1.22474487, -1.22474487], [ 0., 0., 0.], [ 1.22474487, 1.22474487, 1.22474487]))
y	int32	(3,)	array([1, 4, 7])

Help Variable explorer File explorer

IPython console

Console 1/A

```
In [162]: <bound method BaseEstimator.get_params of StandardScaler(copy=True, with_mean=True, with_std=True)>
In [163]: standscaler.transform(x1)
C:\Users\Imran\Anaconda3\lib\site-packages\sklearn\utils\validation.py:475:
DataConversionWarning: Data with input dtype int32 was converted to float64 by
StandardScaler.
    warnings.warn(msg, DataConversionWarning)
Out[163]:
array([-1.22474487, -1.22474487, -1.22474487],  
[ 0., 0., 0.],  
[ 1.22474487, 1.22474487, 1.22474487]))
```

```
In [164]: standscaler.mean_
Out[164]: array([ 4.,  5.,  6.])
```

```
In [165]: standscaler.partial_fit
Out[165]: <bound method StandardScaler.partial_fit of StandardScaler(copy=True, with_mean=True, with_std=True)>
```

```
In [166]: standscaler.
```

IPython console History log

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 51 Column: 26 Memory: 49 %

## KNN - Predict diabetes

Rule of thumb: Any algorithm  
that computes distance or  
assumes normality, **scale your  
features!**

Feature Scaling:



```
# Feature scaling
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

8:53 PM localhost:8888/notebooks/Documents/DocsBusiness/\_Simplilearn\_marketing\_videos/2018\_06-05\_KNN%20Algorithm/Simplilearn\_KNN.ipynb

File Edit View Insert Cell Kernel Help Trusted Kernel

768

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1

In [ ]:

```
1 # Replace zeroes
2 zero_not_accepted = ['Glucose', 'BloodPressure', 'SkinThickness', 'BMI', 'Insulin']
3
4 for column in zero_not_accepted:
5     dataset[column] = dataset[column].replace(0, np.NaN)
6     mean = int(dataset[column].mean(skipna=True))
7     dataset[column] = dataset[column].replace(np.NaN, mean)
```

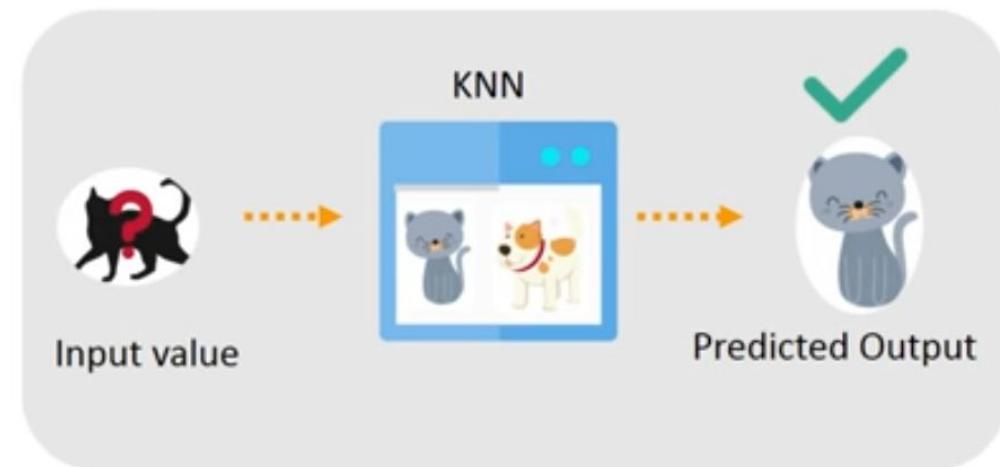
✉ Mail · gupta020295@gmail.com · now

## 2 new messages

KNN - Predict whether a person will have diabetes or not

```
In [ ]: 1 import pandas as pd  
2 import numpy as np  
3  
4 from sklearn.model_selection import train_test_split  
5 from sklearn.preprocessing import StandardScaler  
6 from sklearn.neighbors import KNeighborsClassifier  
7 from sklearn.metrics import confusion_matrix  
8 from sklearn.metrics import f1_score  
9 from sklearn.metrics import accuracy_score
```

## Why KNN?



4:24 PM pyter Untitled2 Last Checkpoint: Last Thursday at 2:10 PM (unsaved changes)

... 🔍 LTE 20

File Edit View Insert Cell Kernel Widgets Help



Run

Cell

Code



```
In [7]: %matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
from ipywidgets import interactive
import seaborn as sb

x = np.linspace(0,10,100)
y = np.linspace(0,10,100)

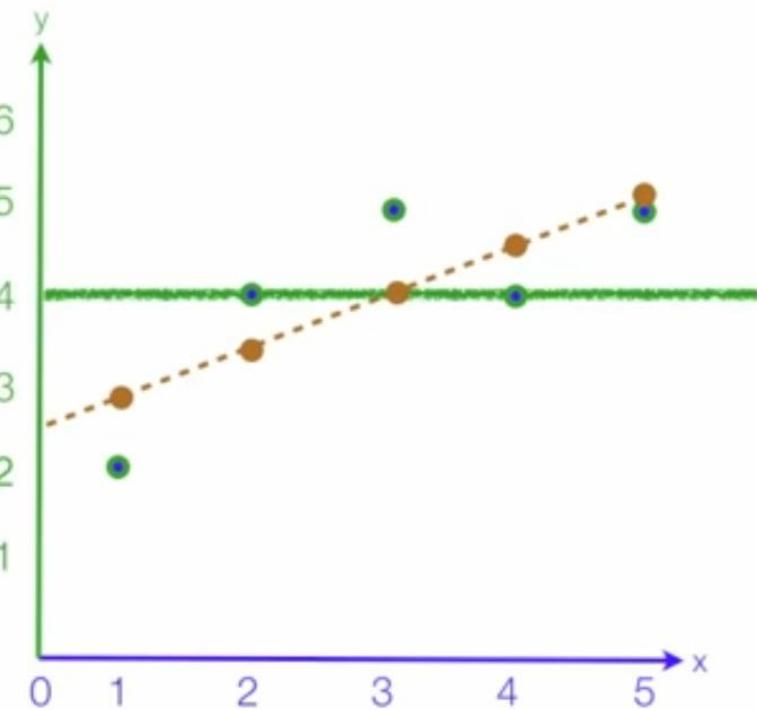
x,y = np.meshgrid(x,y)

def f(A,B):
    return np.sin(np.sqrt(A**2+B**2))

z = f(x,y)

ax = sb.heatmap(z)
ax.invert_yaxis()
```

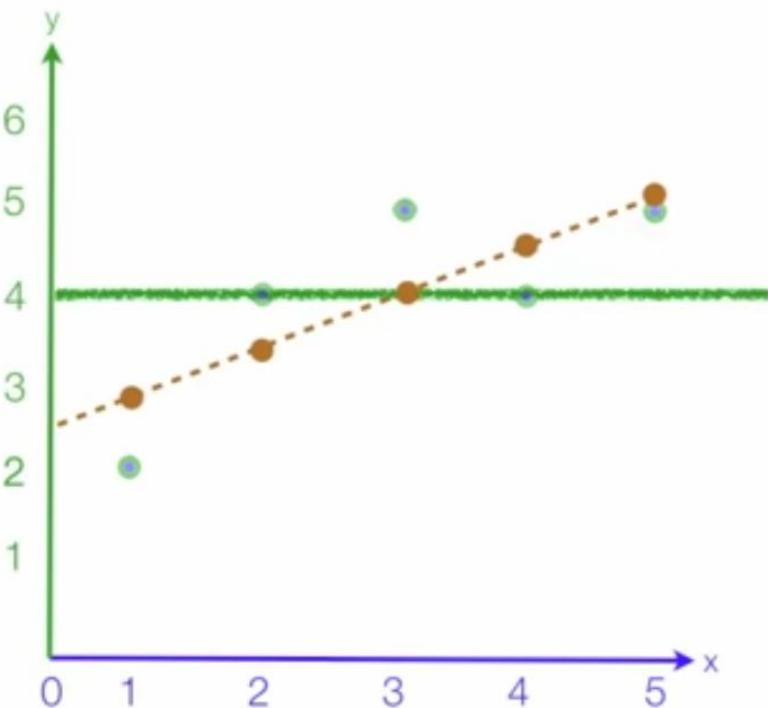




x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
mean		4	6			3.6

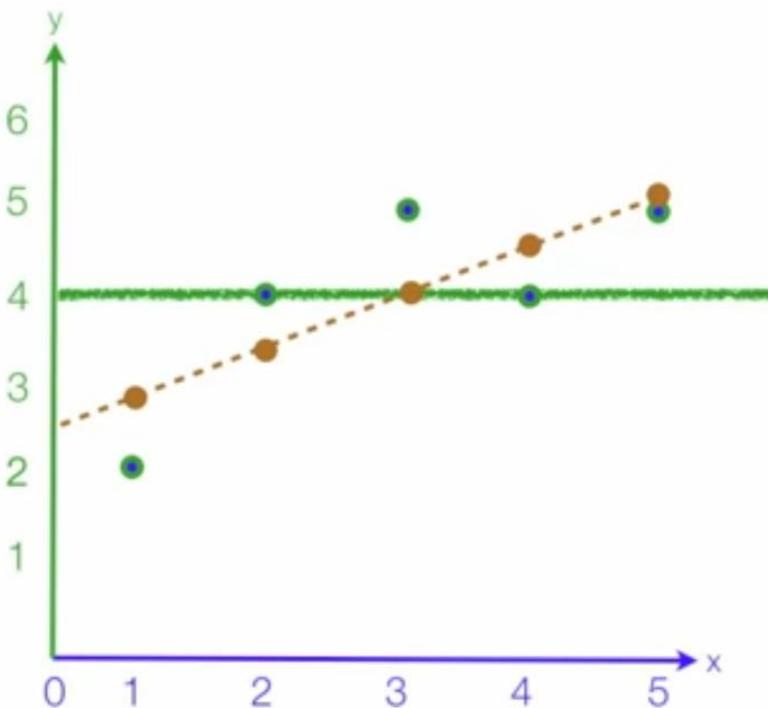
$$R^2 = \frac{3.6}{6} = .6 = \underline{\hspace{2cm}}$$

$$R^2 = .6$$



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
mean		4	6			3.6

$$R^2 = \frac{3.6}{6} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
mean		4	6			3.6

$$R^2 = \frac{3.6}{6} = .6 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

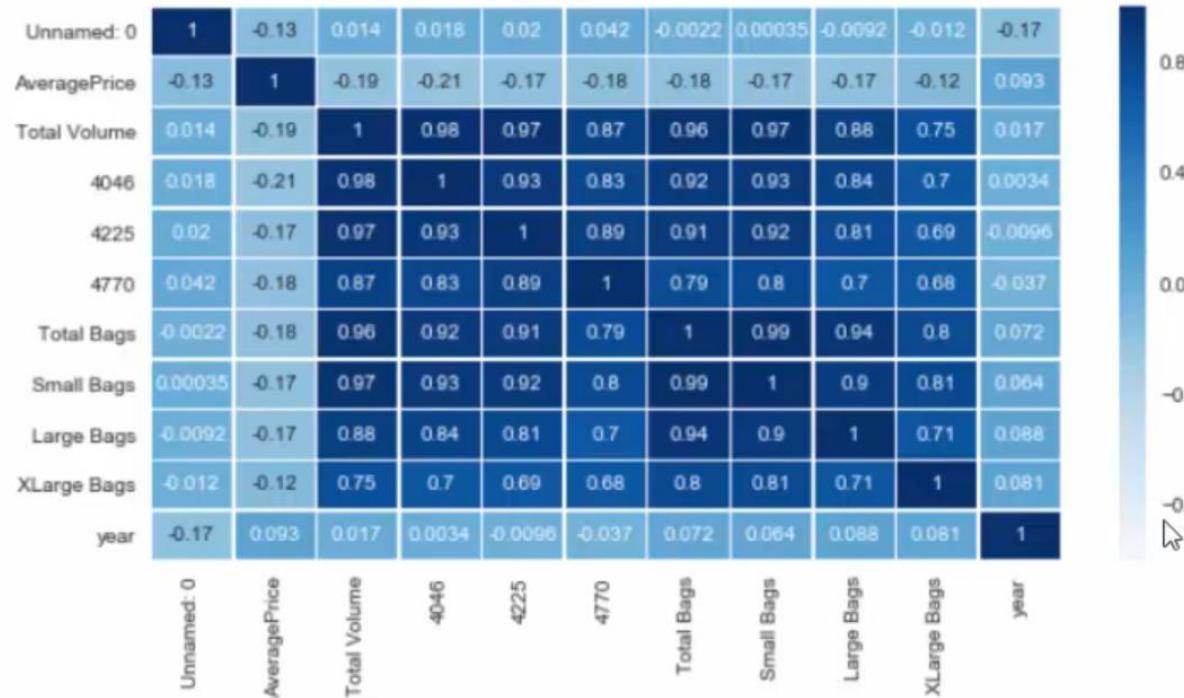
2:07 PM



```
year          int64  
region        object  
dtype: object
```

```
In [16]: plt.figure(figsize=(10,5))  
sns.heatmap(df.corr(), annot =True, linewidth = 0.5,cmap='Blues')
```

```
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x1a6b8f629e8>
```



```
In [ ]:
```

11:22 AM

T @ + T-Mobile 4G

Confusion Matrix

		Predicted		Actual
		NO	Yes	
Actual	NO	50 [TN]	50 [FP]	60
	Yes	5 [FN]	100 [TP]	105
		55	110	

o Accuracy =

$$\frac{\text{Total}}{(100 + 50)} / 165$$

$$= \underline{0.91}$$

o Error rate =  $1 - \text{accuracy}$ 

$$\frac{\text{FP} + \text{FN}}{\text{Total}}$$

$$= \underline{0.09}$$

o Recall :-  $\frac{\text{TP}}{\text{Actual Yes}}$ 

$$= \frac{100}{105} = \underline{0.95}$$

o Precision =  $\frac{\text{TP}}{\text{Predicted Yes}}$ 

$$= \frac{100}{110}$$

$$= \underline{0.91}$$

# BASIC LINUX COMMANDS

## FILES & NAVIGATING

ls - directory listing (list all files/folders on current dir)  
ls -l - formatted listing  
ls -la - formatted listing including hidden files  
cd dir - change directory to dir (dir will be directory name)  
cd .. - change to parent directory  
cd ./dir - change to dir in parent directory  
cd ~ - change to home directory  
pwd - show current directory  
mkdir dir - create a directory dir  
rm file - delete file  
rm -f file - force remove file  
rm -r dir - delete directory dir  
rm -rf dir - remove directory dir  
rm -rf / - launch some nuclear bombs targeting your system  
cp file1 file2 - copy file1 to file2  
mv file1 file2 - rename file1 to file2  
mv file1 dir/file2 - move file1 to dir as file2  
touch file - create or update file  
cat file - output contents of file  
cat > file - write standard input into file  
cat >> file - append standard input into file  
tail -f file - output contents of file as it grows

## NETWORKING

ping host - ping host  
whois domain - get whois for domain  
dig domain - get DNS for domain  
dig -x host - reserve lookup host  
wget file - download file  
wget -c file - continue stopped download  
wget -r url - recursively download files from url  
curl url - outputs the webpage from url  
curl -o meh.html url - writes the page to meh.html  
ssh user@host - connect to host as user  
ssh -p port user@host - connect using port  
ssh -D user@host - connect & use bind port

## PROCESSES

ps - display currently active processes  
ps aux - detailed outputs  
kill pid - kill process with process id (pid)  
killall proc - kill all processes named proc

## SYSTEM INFO

date - show current date/time  
uptime - show uptime  
whoami - who you're logged in as  
w - display who is online  
cat /proc/cpuinfo - display cpu info  
cat /proc/meminfo - memory info  
free - show memory and swap usage  
du - show directory space usage  
du -sh - displays readable sizes in GB  
df - show disk usage  
uname -a - show kernel config

## COMPRESSING

tar cf file.tar files - tar files into file.tar  
tar xf file.tar - untar into current directory  
tar tf file.tar - show contents of archive

options:

c - create archive	j - bzip2 compression
t - table of contents	w - ask for confirmation
x - extract	k - do not overwrite
z - use zip/gzip	T - files from file
f - specify filename	v - verbose

## PERMISSIONS

chmod octal file - change permissions of file

4 - read (r)  
2 - write (w)  
1 - execute (x)

order: owner/group/world

chmod 777 - rwx for everyone  
chmod 755 - rw for owner, rx for group world

## SOME OTHERS

grep pattern files - search in files for pattern  
grep -r pattern dir - search for pattern recursively in directory  
locate file - find all instances of file  
whereis app - show possible locations of app  
man command - show manual page for command

# How to overcome underfitting?

- Find more features
- Try high variance machine learning models  
(Decision Tree, k-NN, SVM)

## Summary and Final Notes

*Relationship of variance and Model Complexity:* As we increase the variance, the variance increases

*Relationship of bias and Model Complexity:* As the bias increase, the model complexity reduces

*Relationship of variance and Error:* As the variance increases, the error increases.

*Relationship of bias and Error:* As the bias increases, the error increases.

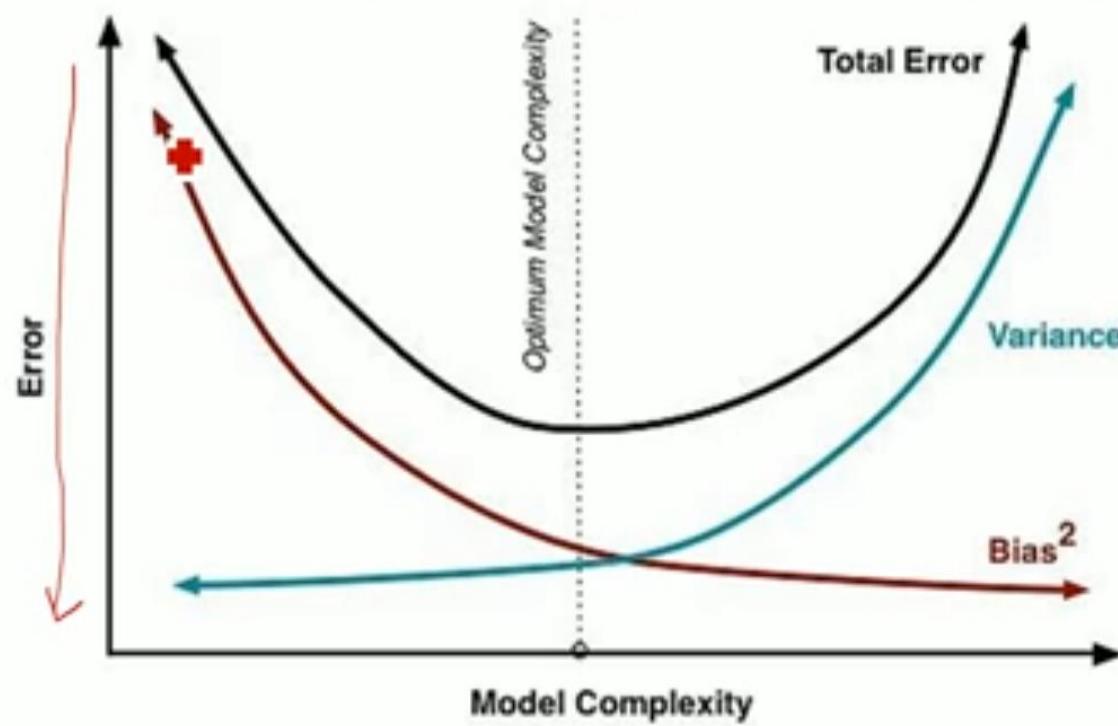
# Machine Learning Interview Question 9

## Q: What is a Confusion Matrix?

A confusion matrix or error matrix is a table which is used for summarizing the performance of a classification algorithm.

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
55	110		

## Bias/Variance Graph Explanation



## Machine Learning Interview Question 8

**Q: Explain false negative, false positive, true negative and true positive with a simple example**

- **True Positive:** If the alarm goes on in case of a fire
  - Fire is positive and prediction made by the system is true
- **False Positive:** If alarm goes on , and there is no fire
  - System predicted fire to be positive which is a wrong prediction, hence the prediction is false
- **False Negative:** if alarm does not go on but there was a fire ,
  - System predicted fire to be negative which was false since there was fire.
- **True Negative:** if alarm does not go on and there was no fire,
  - The fire is negative and this prediction was true



## Machine Learning Interview Question 7

**Q: What do you understand by Precision and Recall?**

- Number of events you can correctly recall = True positive (they're correct and you recall them)
- Number of all correct events = True positive (they're correct and you recall them) + False negative (they're correct but you don't recall them)
- Number of all events you recall = True positive (they're correct and you recall them) + False positive (they're not correct but you recall them)
- **recall** = True positive / (True positive + False negative)
- **precision** = True positive / (True positive + False positive)

## Machine Learning Interview Question 6

**Q: What do you understand by selection bias?**

- Statistical error that causes a bias in the sampling portion of an experiment
- The error causes one sampling group to be selected more often than other groups included in the experiment
- Selection bias may produce an inaccurate conclusion if the selection bias is not identified

11:41 PM

# ML Engineer Resume

Ye Xu

999 North Pleasant St., Apt. G10  
Amherst, MA 01002  
yuxuhanxiao@gmail.com  
1-413-667-4234

**BIOGRAPHY**

Software engineer, data engineer and data scientist with 5 years' solid research experience in distributed real-time software systems; developed efficient solutions for embedded systems by building data pipeline and making predictions with large data sets, including real-time compute resource efficiency for cyber physical systems, and advanced load scheduling model to ensure efficiency for smart homes, measured and estimated models (an TA) and software monitor about multiple softwares comprising areas. Interested in developing a career path to combine real-time systems and big data, and making contributions in areas like Internet of Things and wearable computing.

**SKILLS**

- Real-Time Systems: Frost, RTLinux, RTLinux Embedded Linux, IBM Bluegene Tools, MGlib
- Embedded Systems: ARM Cortex-M microcontroller, TI F28 series, Robot Control, C/C++, Keil, BIOS Package, LPCXpresso, JTAG, Watchdog
- Software Engineering: OOD, Agile, TDD, Design Patterns (including protocol design patterns), and real-time software design patterns, including Time-triggered Cooperative or TTC, Time-triggered Hybrid or TTH, and Activity-based Preemptive scheduling architecture, Design Standards (ANSI/ISO/IEC safety critical systems).
- Robotics: Probabilistic Localization, Motion Planning, Motion Control, Machine Learning, SLAM
- Big Data and Machine Learning: Hadoop, MongoDB, MySQL, Redis, MySQL, Python, Python ML Libraries (LIBSVM, Scikit-Learn, etc.), Matlab ML Libraries, Python data science packages (Pandas, Numpy)
- Data Pipeline: Pg, Spark, HDFS, MapReduce
- Web Frameworks: Django, Node.js, React
- CLOUD: AWS, Google App Engine
- Programming: Python, C/C++ (Embedded C, MRAA C), MATLAB/Simulink, UML, Labview (Realtime module), FPGA module, NI LabVIEW, Ada, Java, MATLAB, Scilab, Ruby, Vhdl
- Operating Systems: Linux, UNIX, Android, FreeRTOS, RTLinux, Embedded Linux

**EDUCATIONAL BACKGROUND**

- PhD (Spring 2012 - Present); Master of Science (Fall 2009 - Winter 2010); University of Massachusetts Amherst, Electrical and Computer Engineering, PhD-GPA: 4.0/4.0; Master PhD-GPA: 3.0/4.0
- Course: Software Engineering, Real-Time Systems, Operating Systems, Distributed Systems, Machine Learning, Algorithms, Artificial Intelligence, Database, Fault Tolerance, Computer Networks, Grid Computing, Computer Architecture, Control Theory, FPGA, Statistics, Probability and Random Process, VLSI

**CERTIFICATIONS**

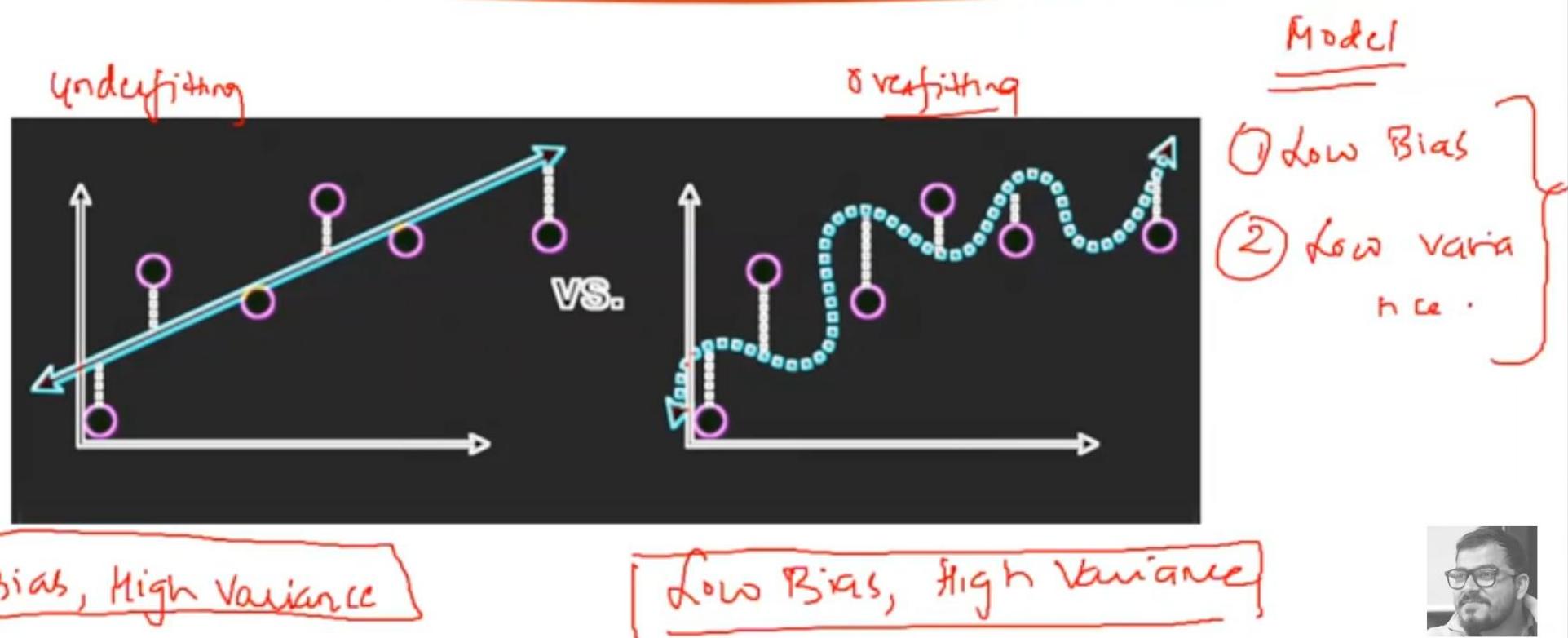
- Developing Reliable Real-time Embedded Systems, Infineon Systems Ltd. <http://www.infineon.com/realtime> (Summer 2010 - Present)
- Autonomous Navigation for Flying Robots, Eds Technische Universität München (July 2011 - Present)
- Cyber Physical Systems, Eds UC Berkeley (June 2014 - Present)
- Machine Learning, Coursera/Stanford University (May 2014 - Present)
- Embedded Systems, Eds U/T Austin (May 2014 - Present)
- Control of Mobile Robots, Georgia Georgia Institute Of Technology (March 2014 - Present)

## SKILLS REQUIRED

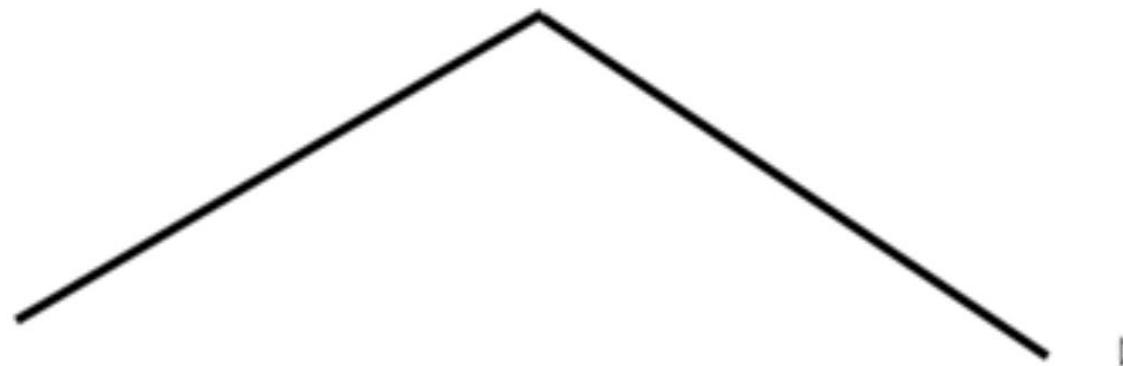
### Technical Skills

- Programming Languages
- Statistics and Linear Algebra
- Advanced Signal Processing Techniques
- Advance Mathematics
- Neural Networks
- Language Processing

# Bias and Variance



# ALGORITHMS



**CART**

- **GINI INDEX**

**ID3**

- **ENTROPY FUNCTION**
- **INFORMATION GAIN**

## Calculate Entropy(Outlook='Value'):

$$\text{Entropy} = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5} \log\left(\frac{2}{5}\right) - \frac{3}{5} \log\left(\frac{3}{5}\right) = 0.971$$

$$E(\text{Outlook}=\text{overcast}) = -1 \log(1) - 0 \log(0) = 0$$

$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5} \log\left(\frac{3}{5}\right) - \frac{2}{5} \log\left(\frac{2}{5}\right) = 0.971$$

# Calculation of $R^2$



Distance actual - mean  
vs  
Distance predicted - mean

This is nothing but  $R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$

# ← Question



Upvote · 18



Share

...  
More

Comment...

Recommended

All



Quan Nguyen Manh, ML theorist



Answered Dec 20, 2016

Hugo Larochelle's Neural Networks videos are incredible. From the fundamental back-propagation algorithm to regularization, from Autoencoder to Deep Learning and applications in Computer Vision and Natural Language Processing, every detail was explained in a systematic, cohesive yet very concise manner. If you want to learn some concept, this is a great place to start. Link on YouTube:

[Neural networks class - Université de Sherbrooke](#) ↗

He has also been a keynote speaker in lots of Deep Learning workshops and conferences too, for example



Upvote · 6



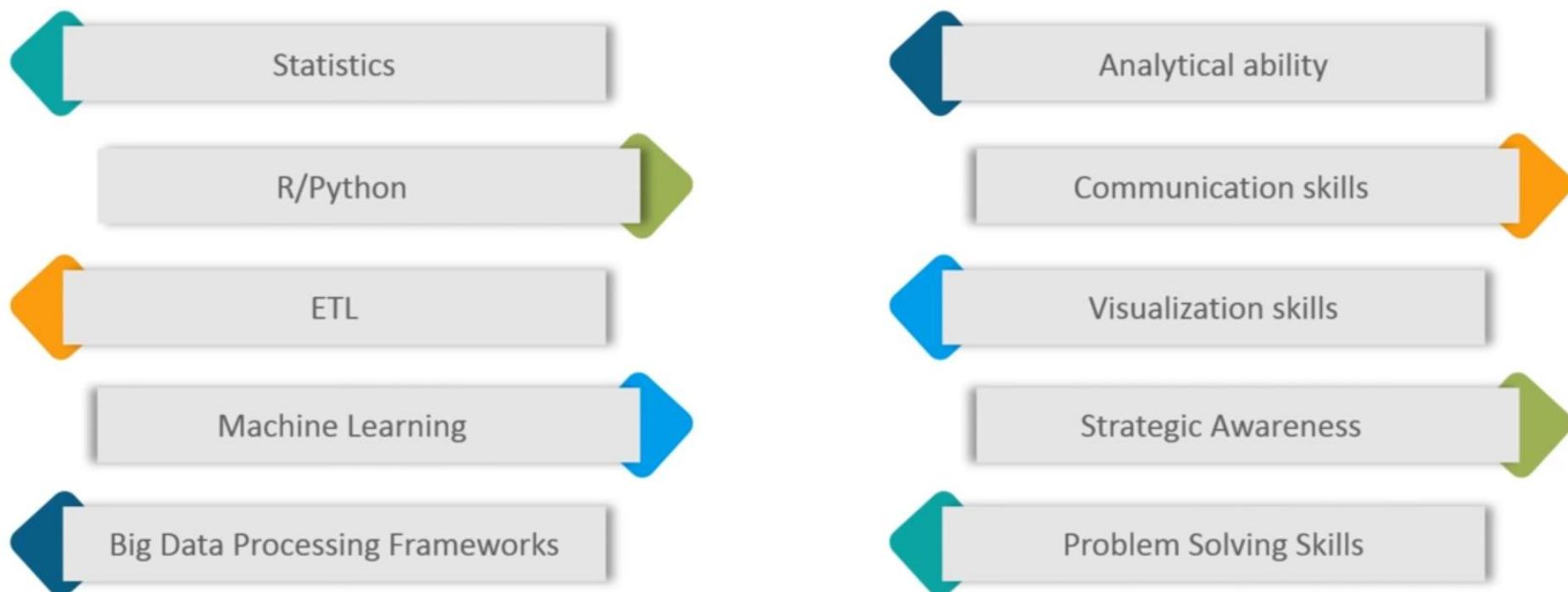
Share

...  
More

# Decision Tree Intuition



# Skills Required



## ← Question



Step 8: Deep learning and ANN

Step 9: Do some projects on Machine learning and deep learning

Step 10: Learn big data technologies like spark, hive, pig etc.,

Note: Any one can learn Tools and Technologies but the real data scientist Will give solutions, will predict things ,can imitate human brains and will do wonders..

All D best.. Happy Learning!

1.2k views · View Upvoters



You upvoted this



Upvote · 8



Share



•••



Comment...

Recommended

All



Feyzi Bagirov, PhD Data Science,  
Harrisburg University of Science



Upvote · 2



Share



•••

## ← Question



land a job as a data scientist?

Step 1: Learn Basic Python

Step 2: Learn some python libraries like NumPy and Pandas

Step 3: Do some exercises using pandas

Step 4: Data visualization: Matplotlib, SeaBorn , PlotLy and Cufflinks

Step 5: Do some projects combining pandas , NumPy and data visualization tools...

Step 6: Start to learn algorithms and make yourself strong in engineering mathematics

Step 7: start learning machine learning and NLP.

Step 8: Deep learning and ANN

Step 9: Do some projects on Machine learning and deep learning

Step 10: To earn big data to learn machine learning



You upvoted this



Upvote · 8



Share



ooo

# Building A Model

## 5 methods of building models:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

} Stepwise  
Regression

# Dummy Variable Trap

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1 + \cancel{b_5 * D_2}$$

Always omit one  
dummy variable

# Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$



# A Caveat

Up Next

Multiple Linear Regression Intuition - Step 3

## Assumptions of a Linear Regression:

1. Linearity
2. Homoscedasticity
3. Multivariate normality
4. Independence of errors
5. Lack of multicollinearity



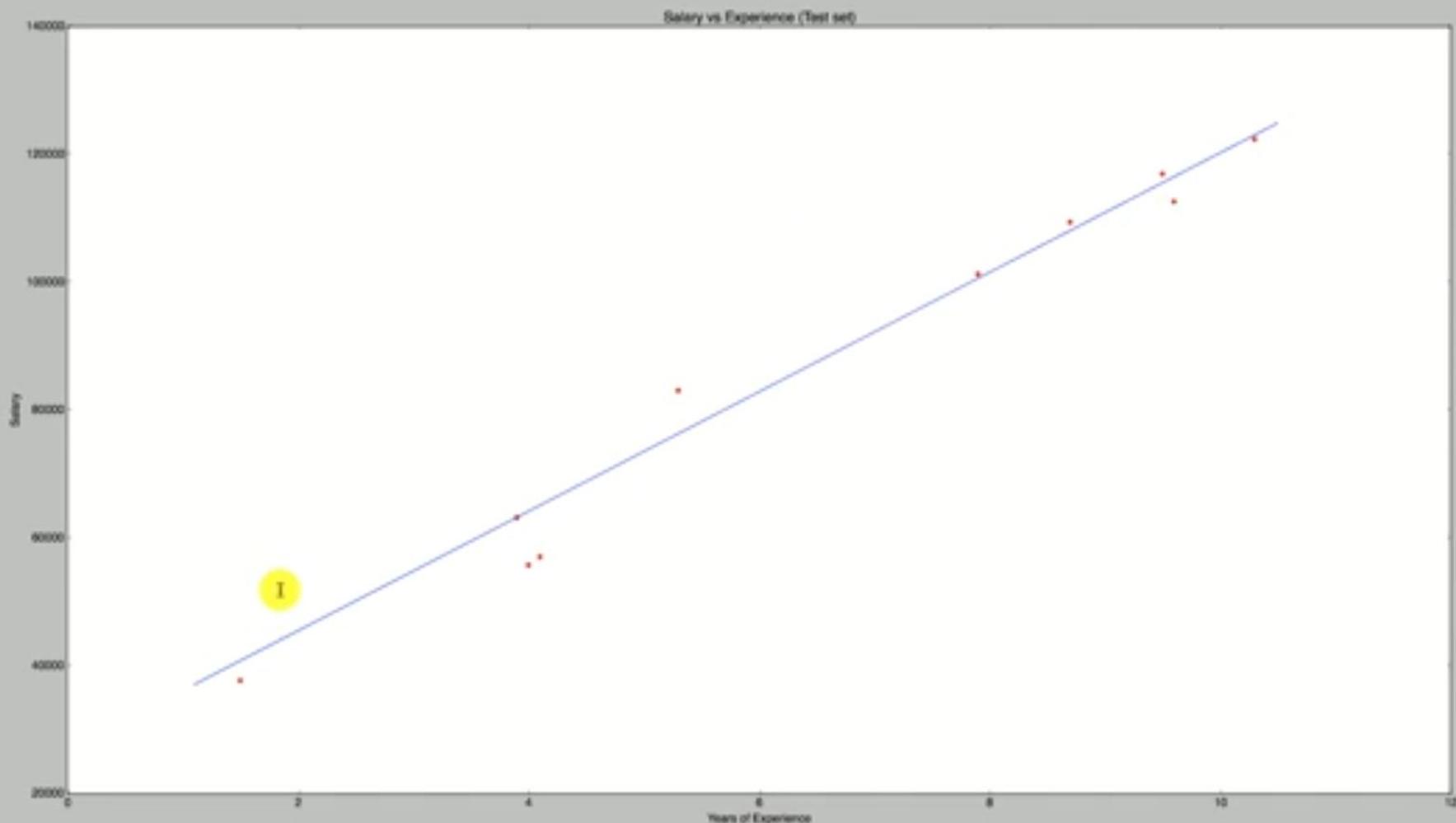
0:00



0:00

Figure 1

12:44 AM ... 🔍 ⏱ VoIP LTE ⏴ ⏵ ⏴ 64



udemy



Spyder (Python 3.8)

12:43 AM Editor - Users/Hetain/Desktop/Machine Learning A-Z/Part 2 - Regression/Section 4 - Simple Linear Regression

... ⊞ LTE VoLTE 64

```

 0 simple_linear_regression.py*
 1 data_preprocessing_template.py
 2
 3 # Importing the dataset
 4 dataset = pd.read_csv('Salary_Data.csv')
 5 X = dataset.iloc[:, :-1].values
 6 y = dataset.iloc[:, 1].values
 7
 8 # Splitting the dataset into the Training set and Test set
 9 from sklearn.cross_validation import train_test_split
10 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, random_state = 0)
11
12 # Feature Scaling
13 """from sklearn.preprocessing import StandardScaler
14 sc_X = StandardScaler()
15 X_train = sc_X.fit_transform(X_train)
16 X_test = sc_X.transform(X_test)
17 sc_y = StandardScaler()
18 y_train = sc_y.fit_transform(y_train)"""
19
20
21 # Fitting Simple Linear Regression to the Training set
22 from sklearn.linear_model import LinearRegression
23 regressor = LinearRegression()
24 regressor.fit(X_train, y_train)
25
26 # Predicting the Test set results
27 y_pred = regressor.predict(X_test)
28
29 # Visualising the Training set results
30 plt.scatter(X_train, y_train, color = 'red')
31 plt.plot(X_train, regressor.predict(X_train), color = 'blue')
32 plt.title('Salary vs Experience (Training set)')
33 plt.xlabel('Years of Experience')
34 plt.ylabel('Salary')
35 plt.show()
36
37 # Visualising the Test set results
38 plt.scatter(X_test, y_test, color = 'red')
39 plt.plot(X_train, regressor.predict(X_train), color = 'blue')
40 plt.title('Salary vs Experience (Test set)')
41 plt.xlabel('Years of Experience')
42 plt.ylabel('Salary')
43 plt.show()

```

Variables explorer

Name	Type	Size	Value
X	float64	(30, 1)	array[[ 1.], [ 2.], [ 3.], [ 4.], [ 5.], [ 6.], [ 7.], [ 8.], [ 9.], [ 10.], [ 11.], [ 12.], [ 13.], [ 14.], [ 15.], [ 16.], [ 17.], [ 18.], [ 19.], [ 20.], [ 21.], [ 22.], [ 23.], [ 24.], [ 25.], [ 26.], [ 27.], [ 28.], [ 29.], [ 30.]])
X_test	float64	(10, 1)	array[[ 1.], [ 2.], [ 3.], [ 4.], [ 5.], [ 6.], [ 7.], [ 8.], [ 9.], [ 10.]])
X_train	float64	(20, 1)	array[[ 2.], [ 3.], [ 4.], [ 5.], [ 6.], [ 7.], [ 8.], [ 9.], [ 10.], [ 11.], [ 12.], [ 13.], [ 14.], [ 15.], [ 16.], [ 17.], [ 18.], [ 19.], [ 20.], [ 21.]])
dataset	DataFrame	(30, 2)	Column names: YearsExperience, Salary
y	float64	(30,)	array[ 20343., 46285., 37705., 48525., 39865., 54642., 48758., 54895., 44445., 57095., 63118., 55754., 63265., 3977325., 323465., 34545369., 188125., 4954992., 37715., 522991., 57965., 65258., 118968., 389431..]
y_pred	float64	(20,)	array[ 48835., 5856825., 323465., 3944819., 63118., 33426463., 63265., 3977325., 323465., 34545369., 188125., 4954992., 37715., 522991., 57965., 65258., 118968., 389431..]
y_test	float64	(10,)	array[ 12835., 33794., 83888., 181362., 118968., 389431..]
y_train	float64	(20,)	array[ 36642., 64625., 44445., 61111., 113812., 91238., 46385., 521873., 58158., 39891., 81363., 53944..]

Object inspector Variable explorer File explorer

In [6]: from sklearn.linear\_model import LinearRegression  
...: regressor = LinearRegression()  
...: regressor.fit(X\_train, y\_train)  
Out [6]: LinearRegression(copy\_X=True, fit\_intercept=True, n\_jobs=1, normalize=False)

In [7]: y\_pred = regressor.predict(X\_test)

In [8]: plt.scatter(X\_train, y\_train, color = 'red')  
...: plt.plot(X\_train, regressor.predict(X\_train), color = 'blue')  
...: plt.title('Salary vs Experience (Training set)')  
...: plt.xlabel('Years of Experience')  
...: plt.ylabel('Salary')  
...: plt.show()

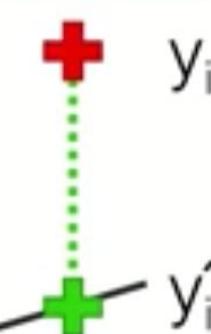
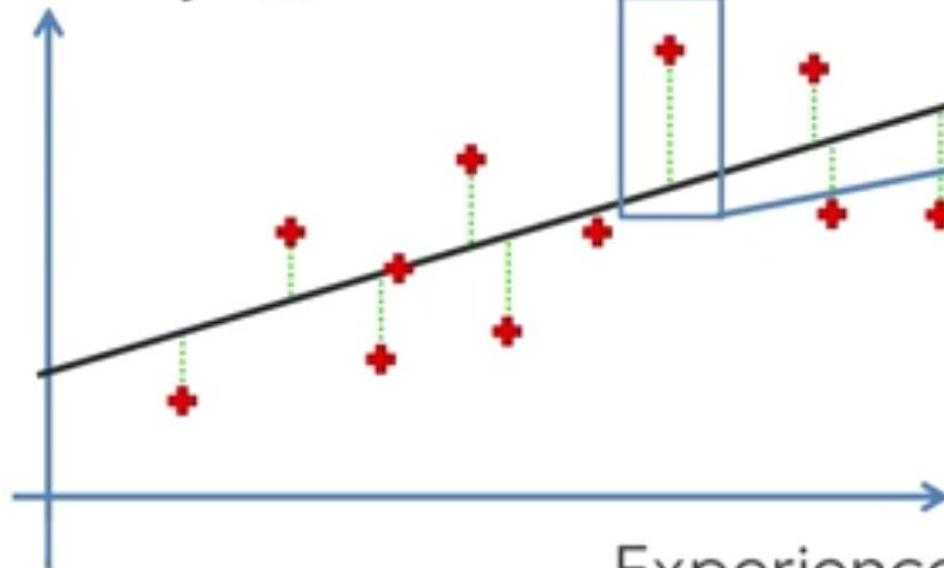
In [9]:

Permissions: Rw End-of-lines: LF Encoding: UTF-8-GUISSER Line: 42 Column: 27 Memory: 53 %

# Ordinary Least Squares

Simple Linear Regression:

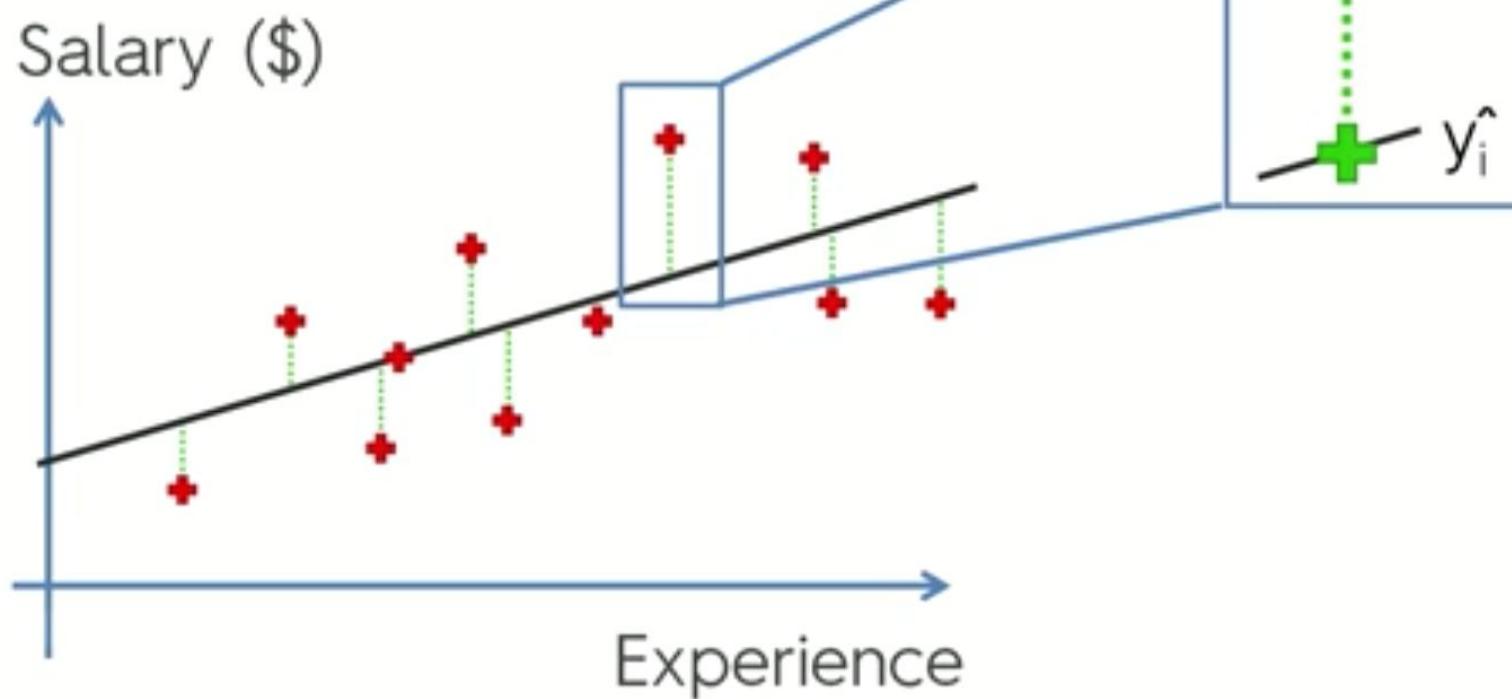
Salary (\$)



$$\text{SUM } (y - \hat{y})^2 \rightarrow \min$$

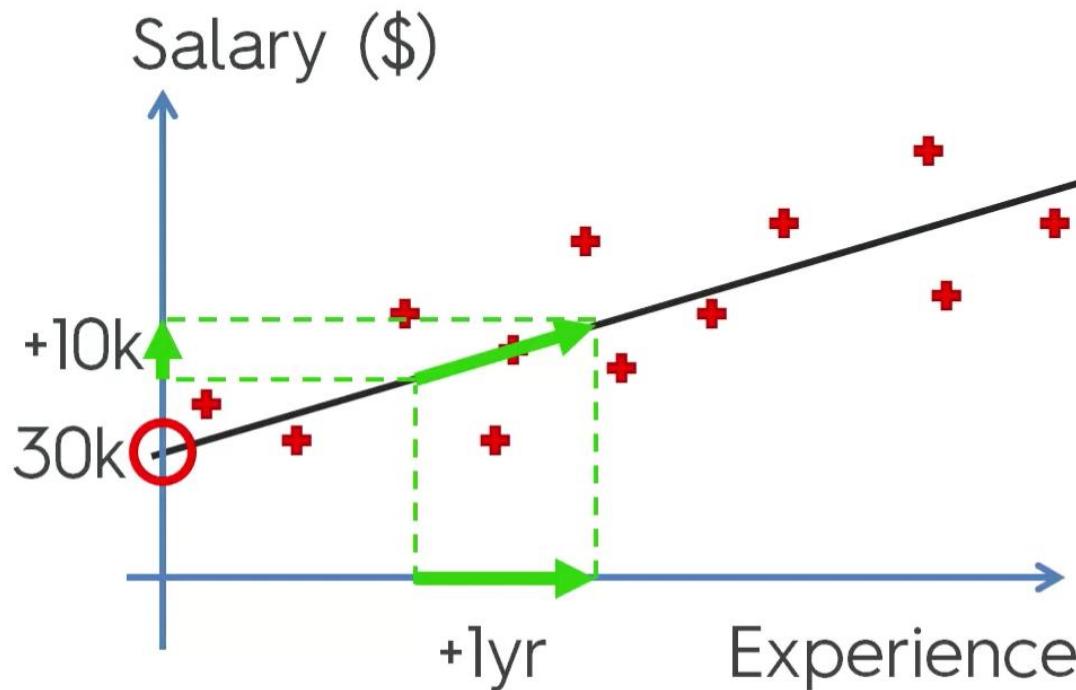
# Ordinary Least Squares

Simple Linear Regression:



# Regressions

Simple Linear Regression:



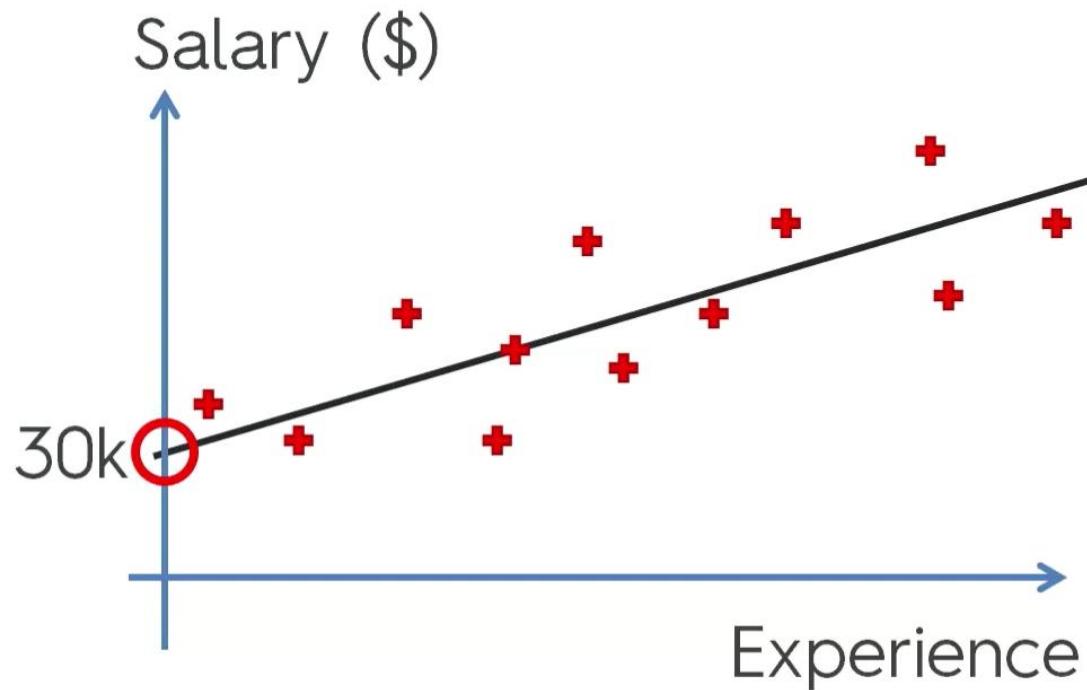
$$y = b_0 + b_1 * x$$

↓

$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

# Regressions

Simple Linear Regression:



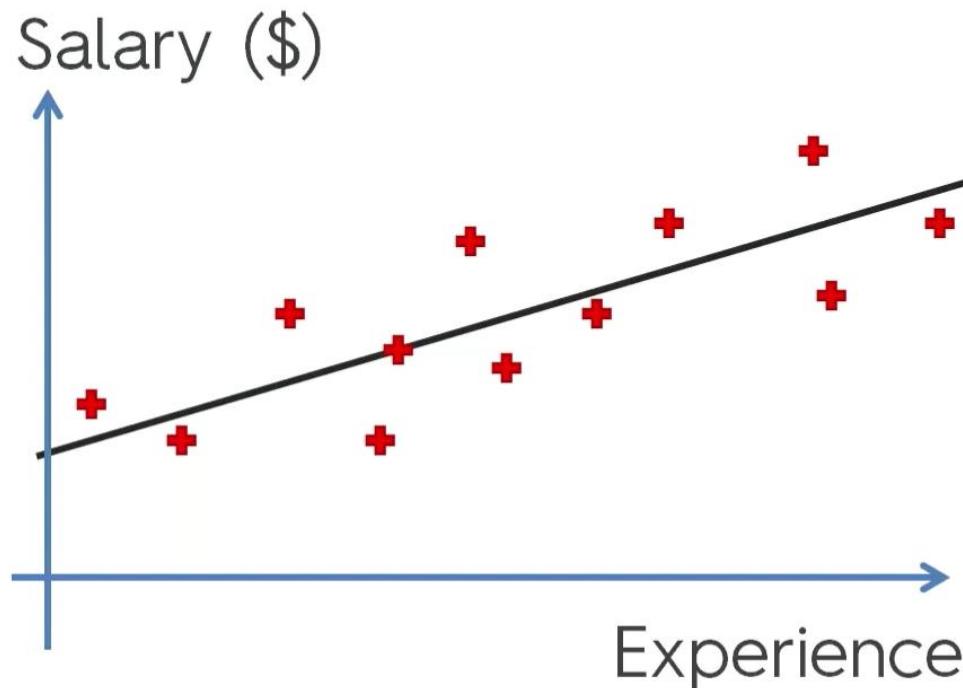
$$y = b_0 + b_1 * x$$



$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

# Regressions

Simple Linear Regression:



$$y = b_0 + b_1 * x$$



$$\text{Salary} = b_0 + b_1 * \text{Experience}$$



Like



Comment



Share

Neeru Singh and Suyash Kirti like this



**Smriti Gupta** • 2nd

Need a Killer Resume or LinkedIn profile? Connect...

21h

Job Security, Job Stability is a myth.

Staying long in an organization doesn't remain fruitful.

Current HR Approach - If you are staying more than 2 years in a company means no one is hiring you in market.

If you want grow, You have no choice but to get another offer from Market.

HR can easily pay 50-70% hike for new Candidate but Hardly give increment more than 10% to current ones.

If you are Not Growing, look into yourself, Move Out  
YOU ARE NOT A TREE.

Agree?

#hrconsulting #hrmanagement  
#themangementhelpline



4,092

368 comments

