Machine Learning

Getting Started
Regression
Classification

Dom Huh

Overview

Motivation to learn machine learning

What is Machine learning?

Introductions

- Data preprocessing

For both regression and classification:

- Conceptual understanding
- Group activities

What is Machine Learning?

Libraries we using today

Pandas - Dataframes and many associated functions

NumPy - Arrays and many associated functions

SciKit Learn - Premade algorithms, and other useful general functions

Pickle - Save and restore trained models

Matplotlib - Visualization of data

Quandl / Pandas-datareader/ Kaggle - A source to get stock data

Other libraries to look into

Tensorflow - General

Theanos- General

OpenCV- Computer Vision

NLTK - Language processing

Seaborn - Heatmaps

libOPT - Optimizations

SciPy - Useful math functions

Pillow - Image processing

Keras - General

Statsmodel - Use of data models

Neuralpy- Neural nets

gzip- Data preprocessing

… and many more (to find more, one resource is to look on Github Collection Machine Learning):
https://github.com/collections/machine-learning

Look at the source code/documentations to learn more about how to incorporate the library.

Review of how to download libraries

Terminal:

```
pip install library_name
```

Libraries used today:

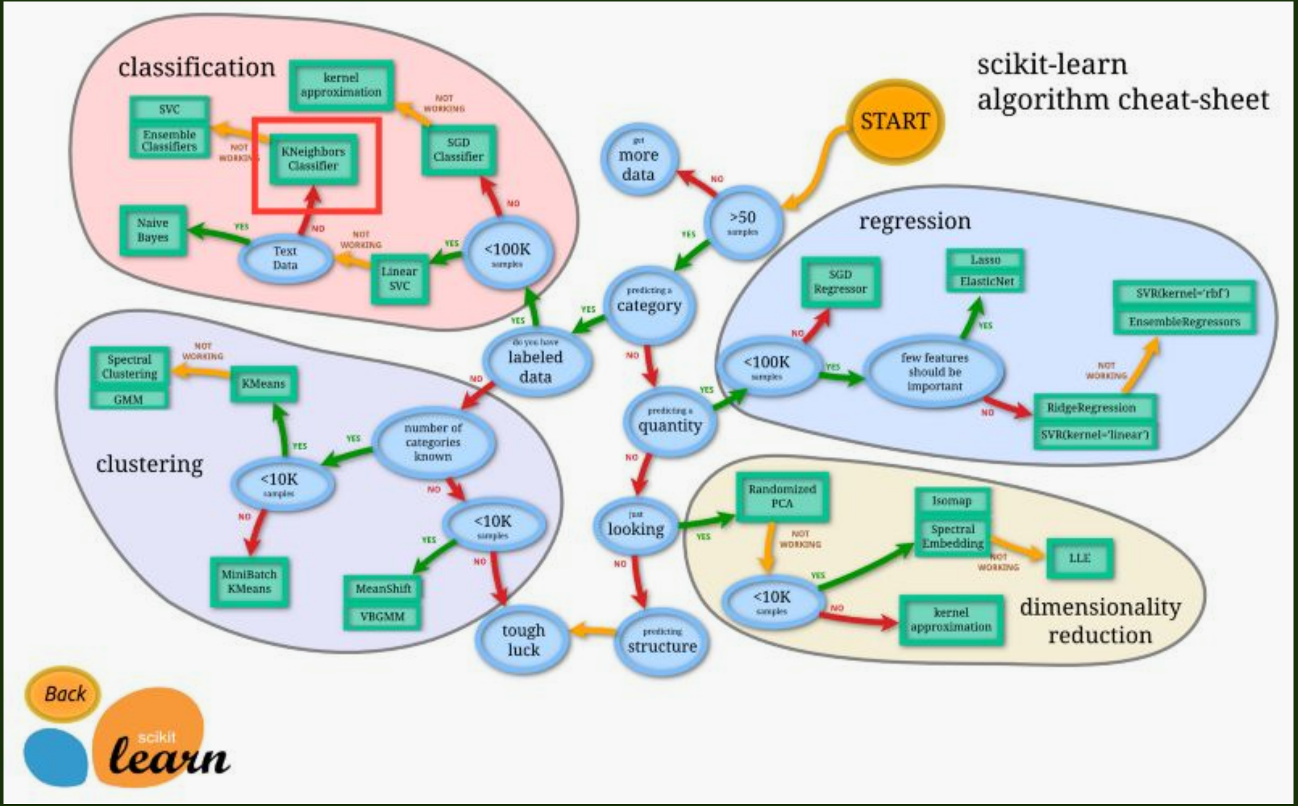
scikit-learn, pandas, numpy, pickle, matplotlib, quandl

What we need to do

Problem

Data

Method



Data

Data preprocessing

Get and look at data (visualize)

Figure out what is important based on what you trying to get

Data ‘formatting’

Data [Cleaning, Integration, Transformation, Reduction, Discretization]

More information on feature engineering:

https://www.youtube.com/watch?v=leTyvBPhYzw

Machine Learning

Regression

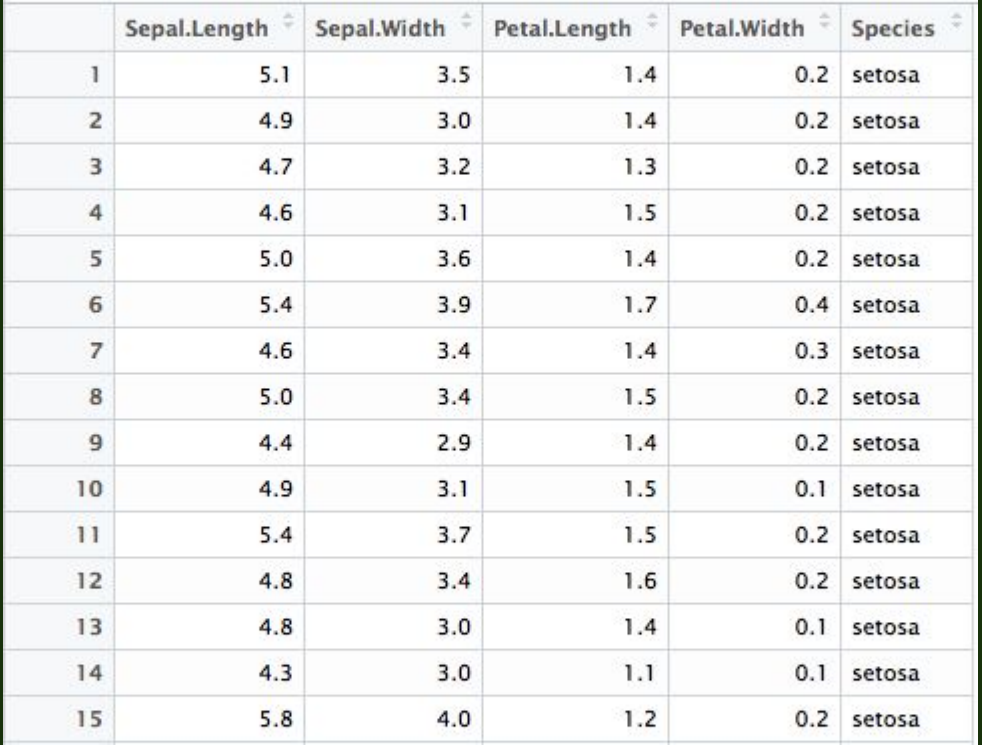
What is regression?

What is the house cost if income is 150 and house rooms is 10?

House cost	Income (K)	House Rooms
1,000,000	100	5
1,500,000	200	9
900,000	80	3

What is regression?

What species is it when sepal length is 6.0, sepal width is 3.9, petal length is 1.3, petal width is 0.1?



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa

Regression

Linear

~~Logistic~~

Polynomial

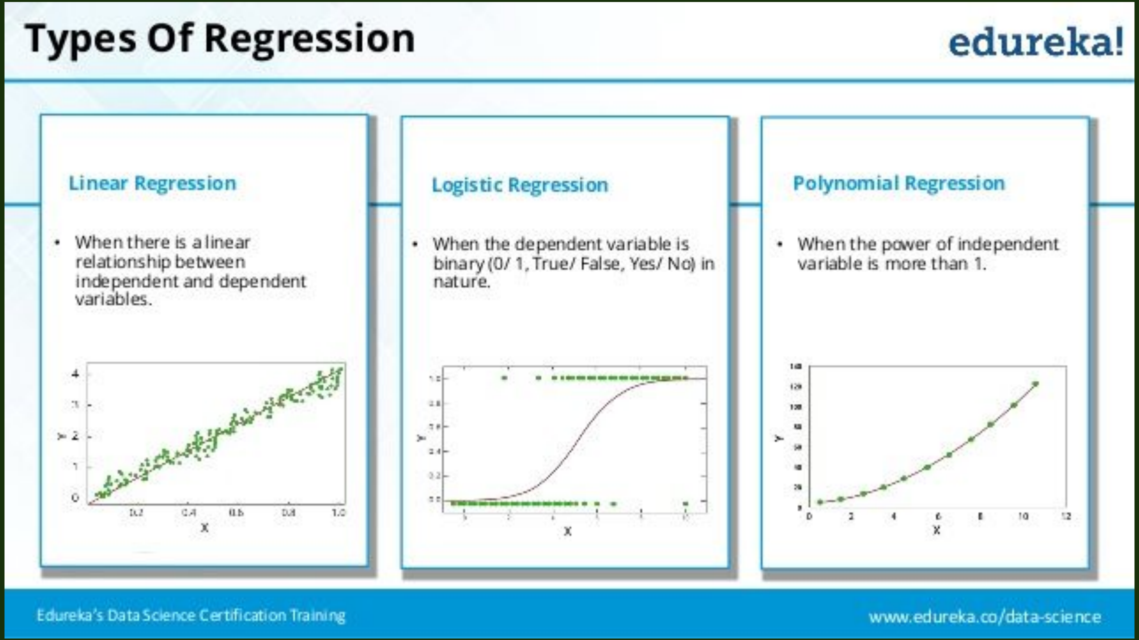
Ridge

Lasso

Elastic Net

SVR

many more...



Types Of Regression

edureka!

Linear Regression

- When there is a linear relationship between independent and dependent variables.

Logistic Regression

- When the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

Polynomial Regression

- When the power of independent variable is more than 1.

Edureka's Data Science Certification Training

www.edureka.co/data-science

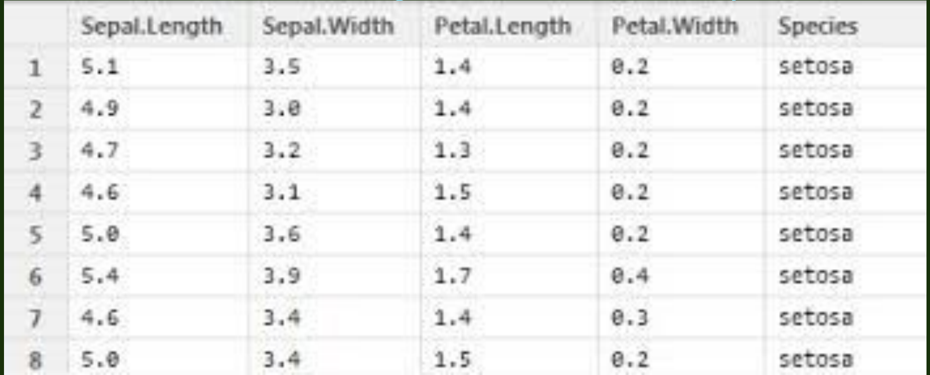
Linear Regression

y= mx+b

Fit data into this best fit line, thus the program is trying to find this line

Can use it to interpolate/extrapolate, and find the accuracy of this line

Visualization: https://towardsdatascience.com/linear-regression-the-easier-way-6f941aa471ea



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Linear Regression Implementation

Inside LinearRegression()

y= mx+b

How do we find m? How do we find b?

Define function for creating the best fit line.

Return m, b

Create regression var set for all x in label list

Tips in making your own Linear Regression Class

Def fit():
Def predict():
Def accuracy():

$$y=wx+b$$

Equation for m:
[mean(x) +mean(y) - mean(x*y)] / [mean(x)^2)-mean(x^2)]

Equation for b:
Use formula above

Equation for R^2 (accuracy):
1- [SquaredError(Y original, Y line) / SquaredError(Y original - Y mean line)]

Equation for Squared Error:
Sum((Array1 - Array2)^2)

Ridge, Lasso, Polynomial, Elastic Net Regression

Implement the following

Ridge, Lasso, Polynomial, Elastic Net Regression

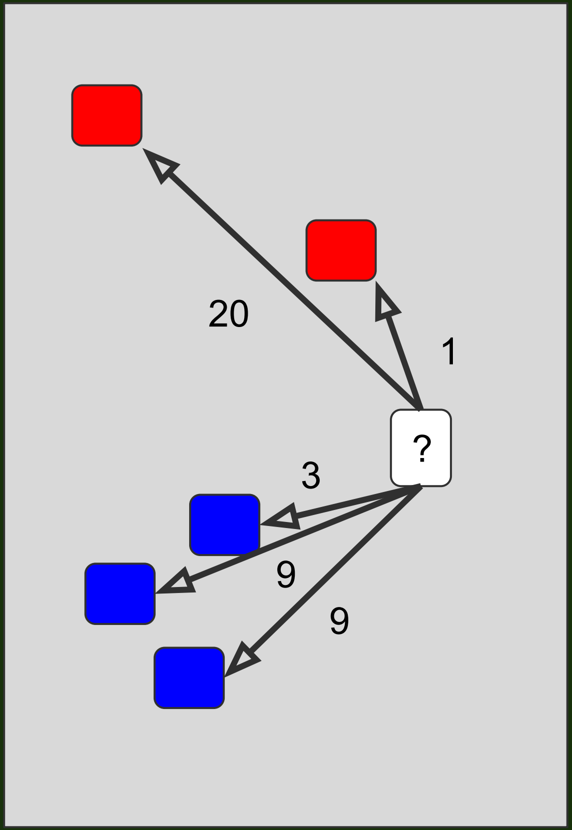
What are they doing?

How are they doing it?

Machine Learning

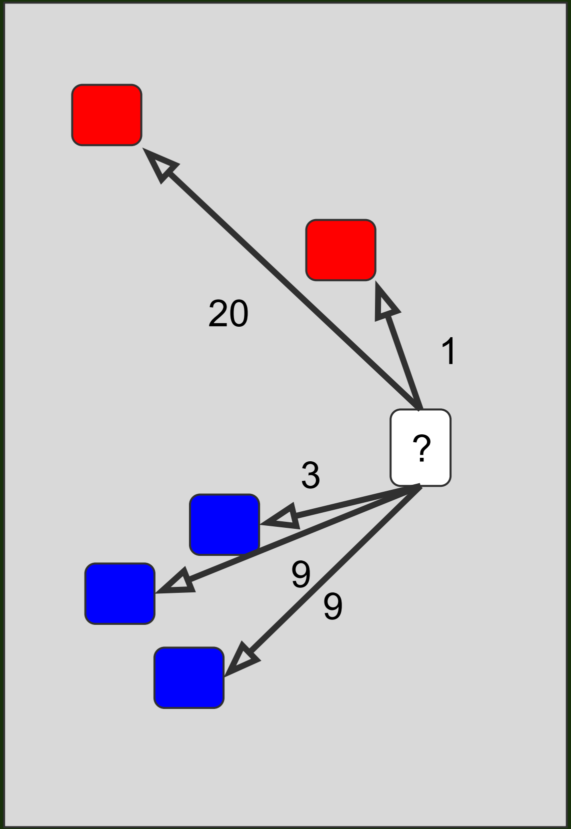
Classification

Classification



What is this unknown's classification?
Red or blue?

Classification



What is this unknown's classification?
Red or blue?

Depends on the algorithm

With K-Nearest Neighbors:
What colors are 'k' number of closest neighbors?
If k=1?
If k=2?
What property of k should be considered?

Neighbors Classifiers

Distance metrics:

Euclidean, Chi-square, Manhattan, Minkowski, Hamming

What does SciKit Learn use?

Brute, KD trees, Ball trees

KNN implementation

Inside KNearestNeighbor()

Find Euclidean distance of each data point to our point

Find the k closest data points to our point

Find the most common data point

Find confidence of that classification

Return that vote with confidence

Tips on making your KNN function

x^2+y^2=d^2

Use np.linalg.norm(data- prediction) for distance

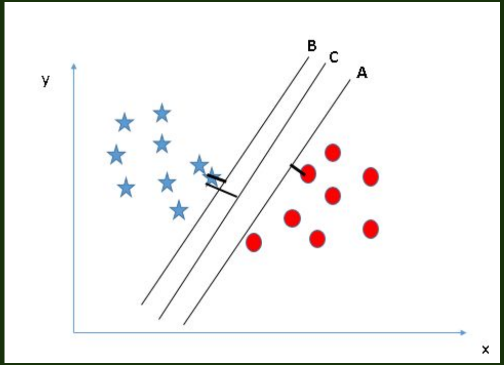
To count how many most common objects in array, use collections.Counter(array).mostcommon(1)

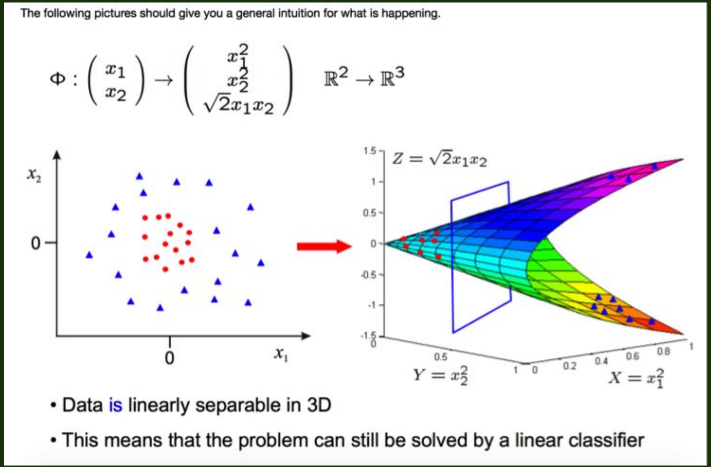
Next time...

Basic Intuition

SVM: Classification by separation using hyperplanes (n-dim) adjusted by maximized width possible between known data points

Clustering: Classification by centroids using k-means or hierarchical mean shift to based on position and bandwidth.



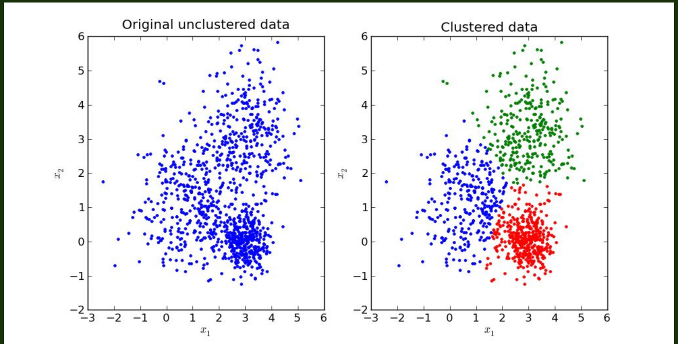


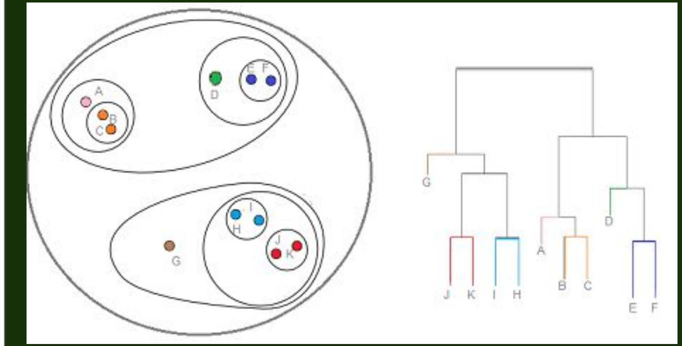
The following pictures should give you a general intuition for what is happening.

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \qquad \mathbb{R}^2 \to \mathbb{R}^3$$

x_2
0
0
x_1
$Z = \sqrt{2}x_1x_2$
1.5
1
0.5
0
-0.5
-1
-1.5
0
0.5
1
$Y = x_2^2$
0
0.2
0.4
0.6
0.8
1
$X = x_1^2$

- Data is linearly separable in 3D
- This means that the problem can still be solved by a linear classifier


Original unclustered data
Clustered data
x_2
x_1



A
B
C
D
E
F
G
H
I
J
K
G
J
K
I
H
A
B
C
D
E
F