Introduction to Machine Learning

By:

Shyamal Vaderia

Agenda

- Introduction
- Basics
- Regression
- Classification
- Clustering
- Decision Trees

In other words…

"Machine learning is a set of techniques, which help in dealing with vast data in the most intelligent fashion (by developing algorithms or set of logical rules) to derive actionable insights (delivering search for users in this case)."

How ML is different from Simple if-else rules?

Use-Cases

- Spam Email Detection
- Machine Translation (Language Translation)
- Image Search (Similarity)
- Amazon Recommendations : Clustering
- Google News : Classification

Terminology

- Samples
- Features
- Feature vector

Data about Movies

Name	Actors	Director	Budget	Type	Rating
Lagan	Aamir khan	T.R ramana	25 Cr	Drama	7.2
Wanted	Salman khan	Ali Zafar	50 Cr	Action	6.2
3 idiots	Aamir khan	Rajiv irani	35 Cr	Drama	8.4
Queen	Kangana	Abhi chopra	15 Cr	Drama	7.9
Neerja	Sonam kapur	Niraj pandey	20 Cr	Thriller	7.6
Dangal	Aamir khan	Nitesh Tiwari	70 Cr	Sport	8.9
Drishyam	Ajay Devgn	Rahul Prasad	30 Cr	Thriller	8.5

- Features

Name	Actors	Director	Budget	Type	Rating

- Feature Vector

  [178, 5, 2, 25, 1, 7.2]

- Matrix Representation of Data

Let’s dig deep into it…

What do you mean by

Apple

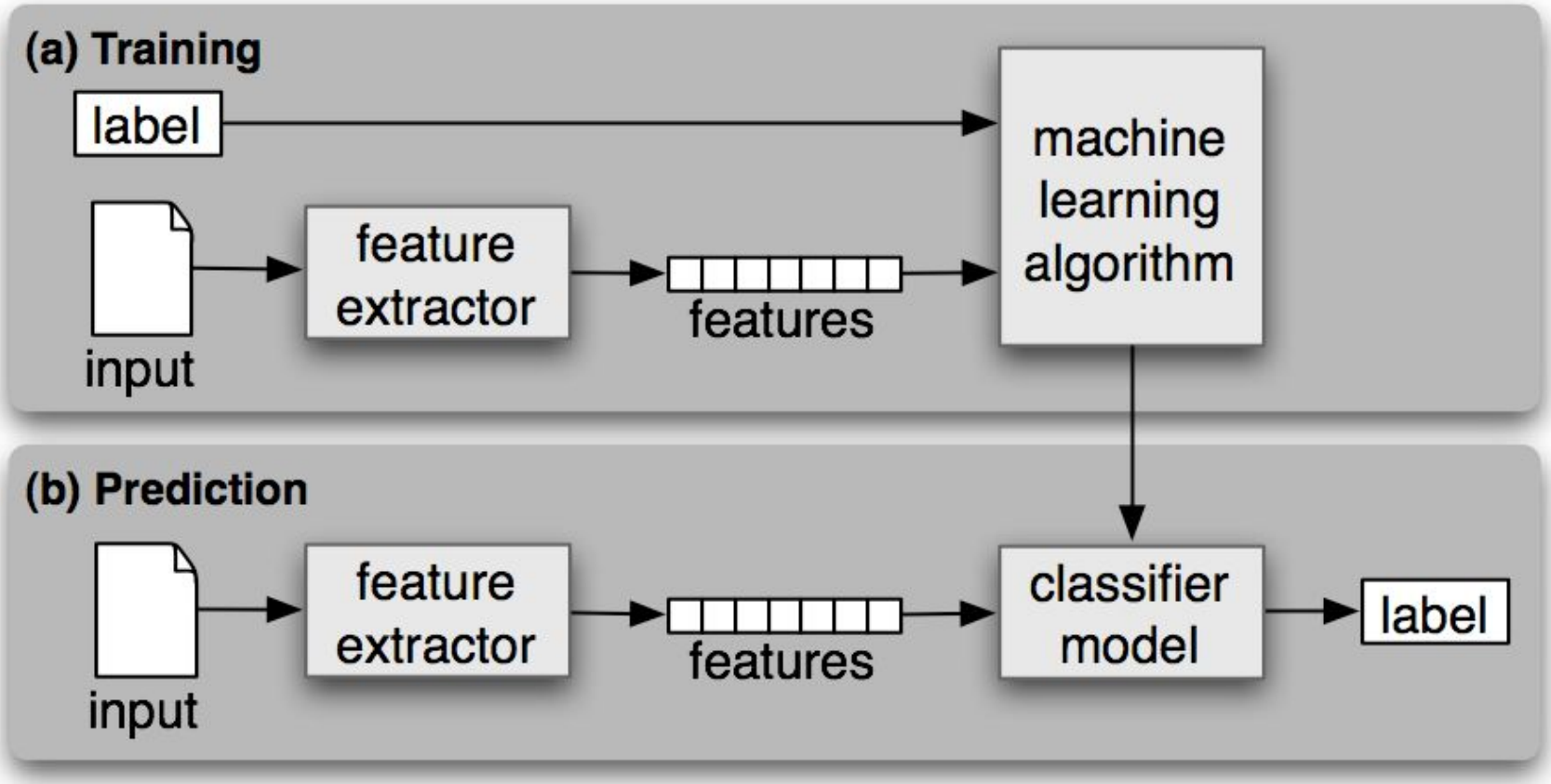
Learning

Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc…

Features:
1. Sky Blue
2. **Logo**
3. Shape
etc…

Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc…

Workflow


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

Categories

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised Learning

- Predictive Models.
- The correct classes of the training data are known.
- Algorithms: Nearest Neighbor, Naïve Bayes, Decision Trees, Regression.

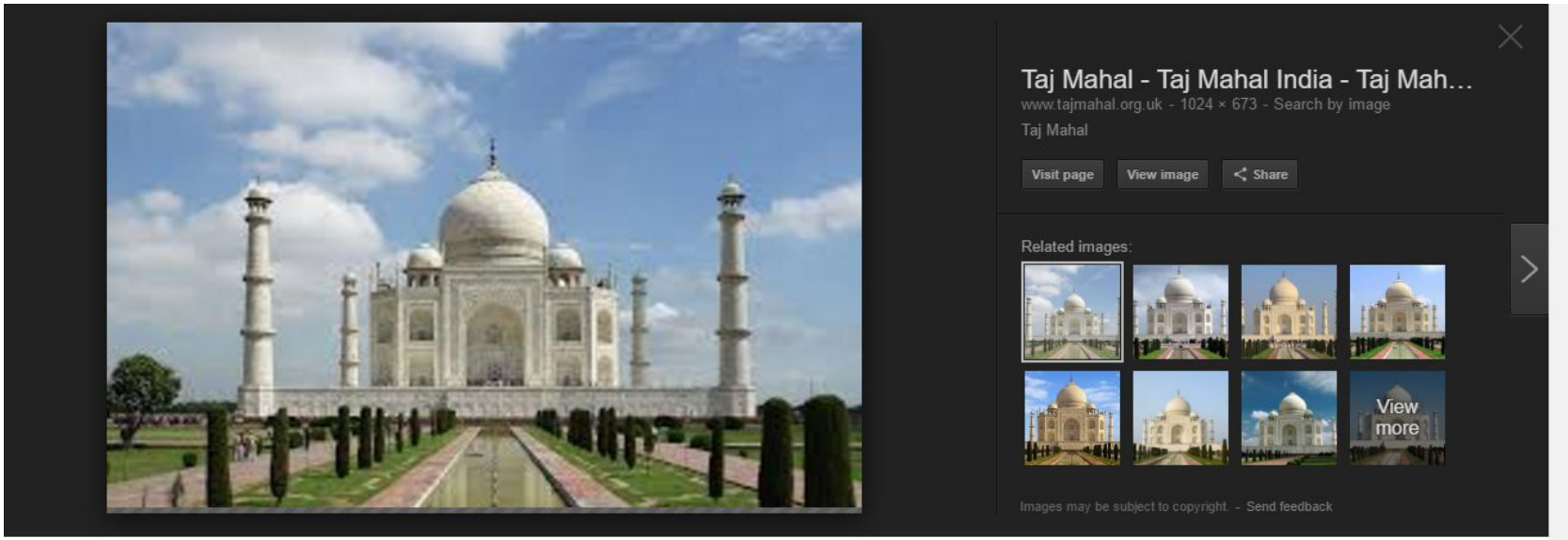
Labels

Name	Actors	Director	Budget	Type	Rating
Lagan	Aamir khan	T.R ramana	25 Cr	Drama	7.2
Wanted	Salman khan	Ali Zafar	50 Cr	Action	6.2
3 idiots	Aamir khan	Rajiv irani	35 Cr	Drama	8.4
Queen	Kangana	Abhi chopra	15 Cr	Drama	7.9
Neerja	Sonam kapur	Niraj pandey	20 Cr	Thriller	7.6
Dangal	Aamir khan	Nitesh Tiwari	70 Cr	Sport	8.9
Drishyam	Ajay Devgn	Rahul Prasad	30 Cr	Thriller	8.5

Unsupervised Learning

- The correct classes of the training data are not known.
- You are trying to find out the structure or relationship between different inputs.
- Algorithms: K-means clustering

Example


Taj Mahal - Taj Mahal India - Taj Mah…
www.tajmahal.org.uk - 1024 × 673 - Search by image
Taj Mahal
Visit page
View image
Share
Related images:
View more
Images may be subject to copyright. - Send feedback

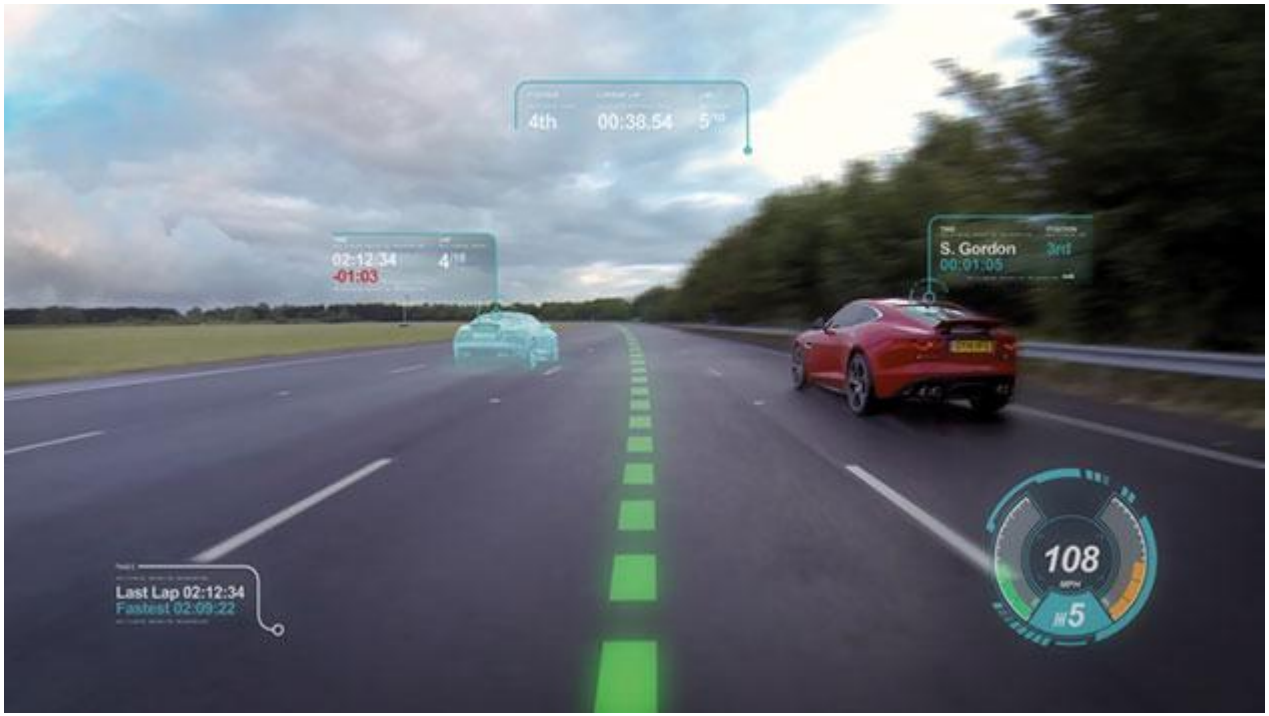
Example

Name	Actors	Director	Budget	Type
Lagan	Aamir khan	T.R ramana	25 Cr	Drama
Wanted	Salman khan	Ali Zafar	50 Cr	Action
3 idiots	Aamir khan	Rajiv irani	35 Cr	Drama
Queen	Kangana	Abhi chopra	15 Cr	Drama
Neerja	Sonam kapur	Niraj pandey	20 Cr	Thriller
Dangal	Aamir khan	Nitesh Tiwari	70 Cr	Sport
Drishyam	Ajay Devgn	Rahul Prasad	30 Cr	Thriller

Reinforcement Learning

- Learning from feedback from the environment.
- Good example of this is Self driving car,
  - It decide continuously from it's environment about which route to take? What speed to drive on?
- Algorithm: Markov Decision Process

Example of RL



Machine Learning Techniques

Techniques

- Classification: predict class from observations
- Regression (prediction): predict value from observations
- Clustering: group observations into "meaningful" groups

Classification

- Classify a document into a predefined category.
- Documents can be text, images etc.

Spam detection

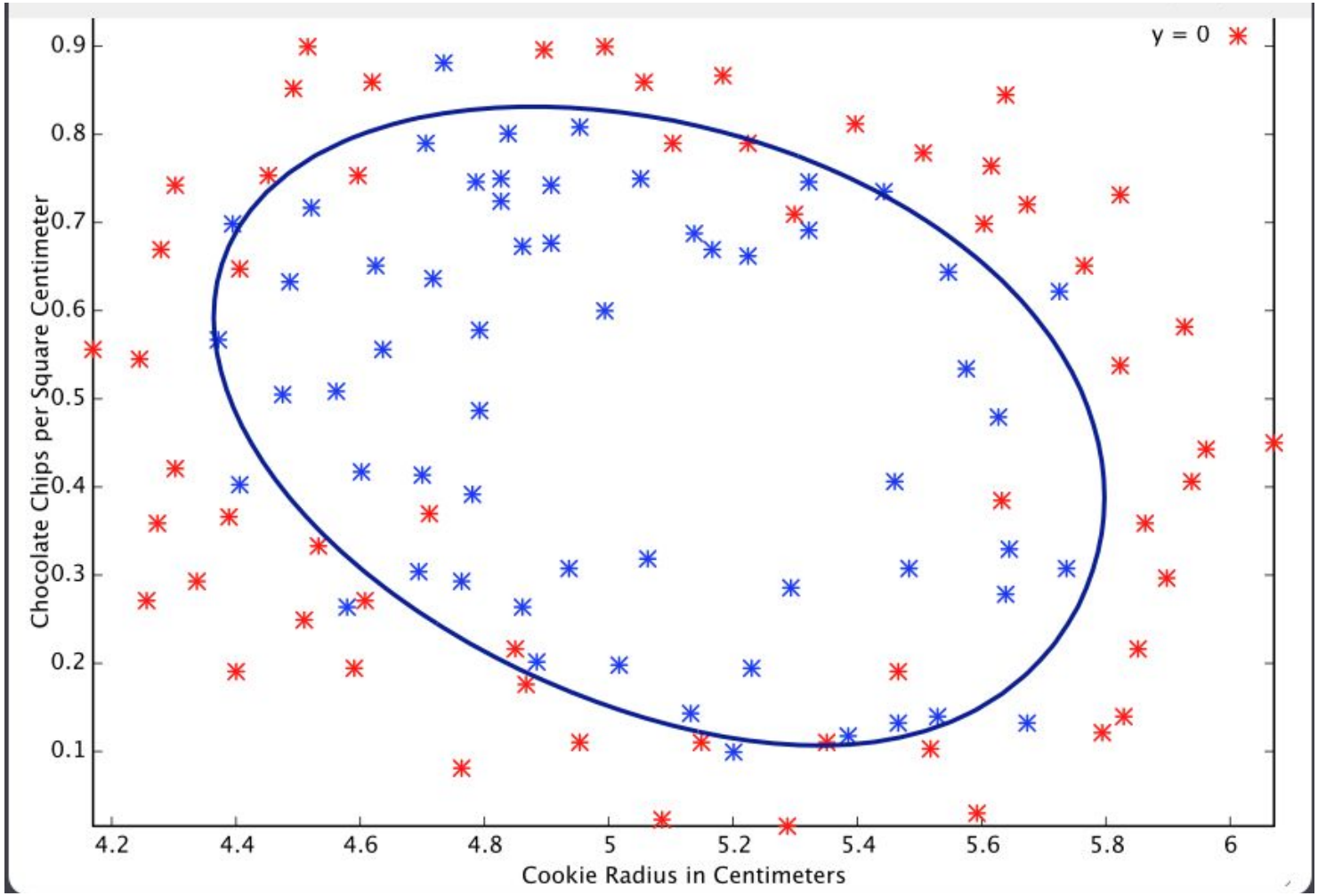
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab -- If you do not want to receive any more newsletters, please click here	9:40 pm
iEntry	**Welcome iEntry Member** - Ultimate Guide To Assessing	9:23 pm
New-Zealand-Jobs.067L	**Come to New Zealand to find a great job and settle here (Search for all Jobs from diffe...** - Search for all Jobs from different kinds of industries Find a Job in Enchanting	8:18 pm
CarSizzler	**Assured Free Luxurious Ride worth Rs.300 with Uber Cabs** - Home Home Buy New Car Buy New Car Sell Car Sell Car Tech Tics Tip & Tale Facebook 41727 others	6:05 pm
Supermarket Promotion	**Enjoy Rs.1700 voucher valid at any supermarket!** - If you are unable to view this mailer Click here HOW TO CONTACT US? BY EMAIL: support@savethedeals.in	4:51 pm
Entireweb Newsletter	**Hire an SEO the Right Way – 6 Tips You Must Remember for Life** - Unsubscribe me View web version Become a fan on Facebook Follow us on Twitter September 5th, 2(	1:24 pm
Max Bupa	**A policy that understands your family's medical need** - open in fresh tab -- If you do not want to receive any more newsletters, pleas	11:08 am
Scoop.it	**Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure | SME a...** - Scoop.it Facebook Twitter G+ H re the top	9:30 am
standard charterer Bank	**Instant approval on your Credit Card`**	7:27 am
CAR TRADE	**Sell your car at no cost at all** - If you are having trouble viewing this email,view web version | View this message in your mobile |	Sep 4
Uday	**VPS Web Hosting Services Provider** - Dear Sir, I am Uday Sharma, Business development executive. We are providing quality VPS hosting for	Sep 4
Mark Regan, SPN	**How to Find Your Most Valuable Keywords [Free Guide]** - This is a SiteProNews/ExactSeek Webmaster Exclusive Mailing! To drop your subscription, use the link	Sep 4
HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab You have received this mailer from Shop@Best on behalf of HDFC Bank because you	Sep 4
CAR TRADE	**Sell your car at no cost at all** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
ICICI Bank	**Home Loan Interest Rate starting from 10.15%*. Get Instant Approval!** - open in fresh tab -- If you do not want to receive any more newsletters, please Click Here	Sep 4
calculateyourwealth	**It's good when your bank helps you manage your wealth and fulfill your ambitions** - Calculate Now Dreams you wish to realize in your lifetime require enough wealth. (	Sep 4
Angel Broking	**Get Low Brokerage - Free Demat & Trading Account** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
Bankbazaar	**7 Minute Instant Online Approval for your PESONAL LOAN** - Now get instant online Personal Loan approval in 7 minutes by BankBazaar.com from leading Banks in	Sep 4
Jayde	**Welcome To The Jayde Newsletter!** - Welcome To WebProNews Welcome To The Jayde Newsletter! Before we begin, make sure to add	Sep 4
ineedhits noreply	**[ineedhits] Your ineedhits Account and Password** - ACCOUNT CREATION Account ID : A1588368 Dear Rah, Welcome to ineedhits. Yo	Sep 4
Rekha	**Mobility Apps for Your Business** - While we look at the span of last 20 years, we could broadly look at two distinct eras, - Life in	Sep 4
SlideShare Newsletter	**Top Tips From the World Champions of PowerPoint** - View online version Remember to display images Meet the PowerPoint World Champs Top Tips From the	Sep 4
Dilshad Pathan	**Feeling Hesitate to Discuss personal Health Queries** - My Life Care Follow Us on facebook twitter linkedin Google+ Feeling Hesitate to Discuss personal	Sep 4
Vaishu	**TAKE YOUR PICK. Register in SimplyMarry** - TAKE YOUR PICK. Register in SimplyMarry -- Regards Vaishu	Sep 4

Spam

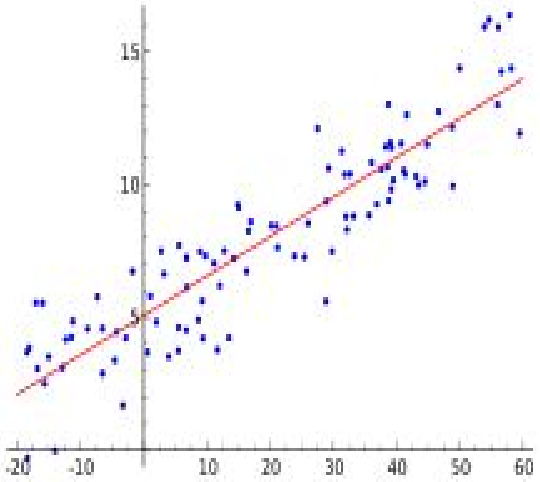
Spam

Classification Example

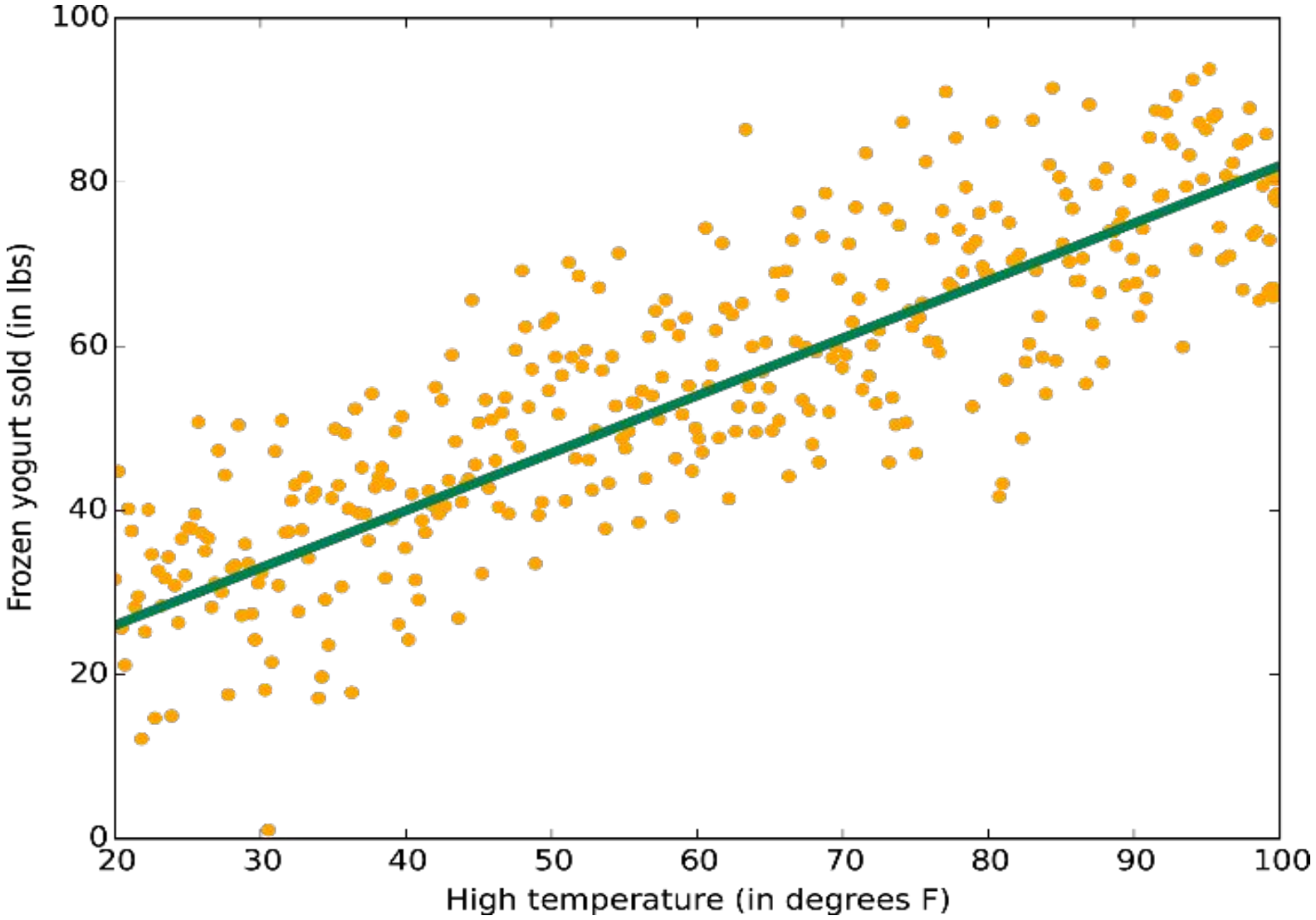


Regression

- **Best-Fit curve**
- **Regression analysis** is a statistical process for estimating the relationships among variables.
- Regression means to **predict** the output value using training data.
- Popular one is Logistic regression (binary regression)



Regression Example



Classification vs Regression

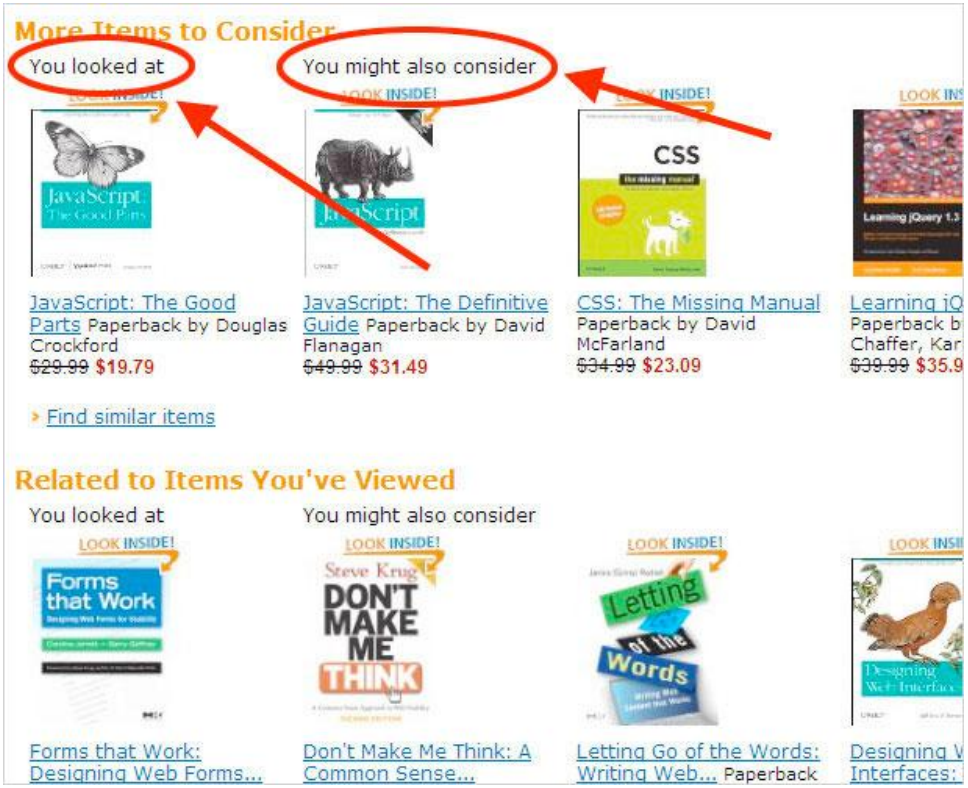
- Classification means to group the output into a class.
- discrete/categorical

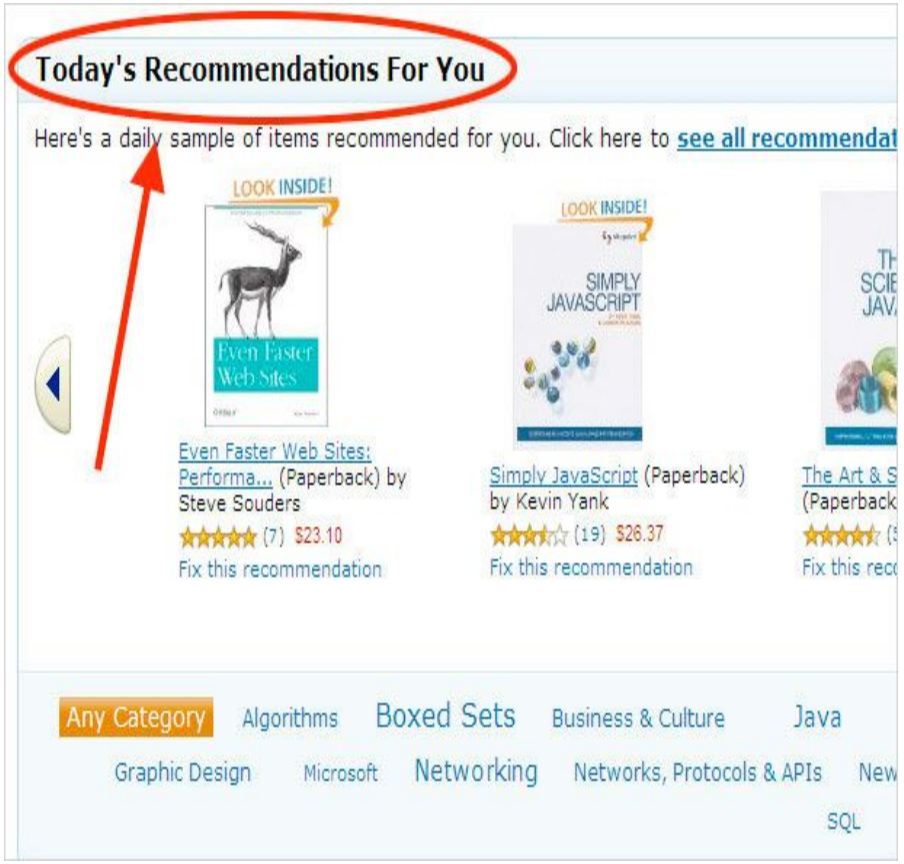
- Regression means to predict the output value using training data.
- real num/continuous

Clustering

- **Clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar to each other
- Objects are not predefined
- For e.g. these keywords
  - "man's shoe"
  - "women's shoe"
  - "women's t-shirt"
  - "man's t-shirt"
  - can be cluster into 2 categories "shoe" and "t-shirt" or "man" and "women"
- Popular ones are **K-means clustering** and **Hierarchical clustering**

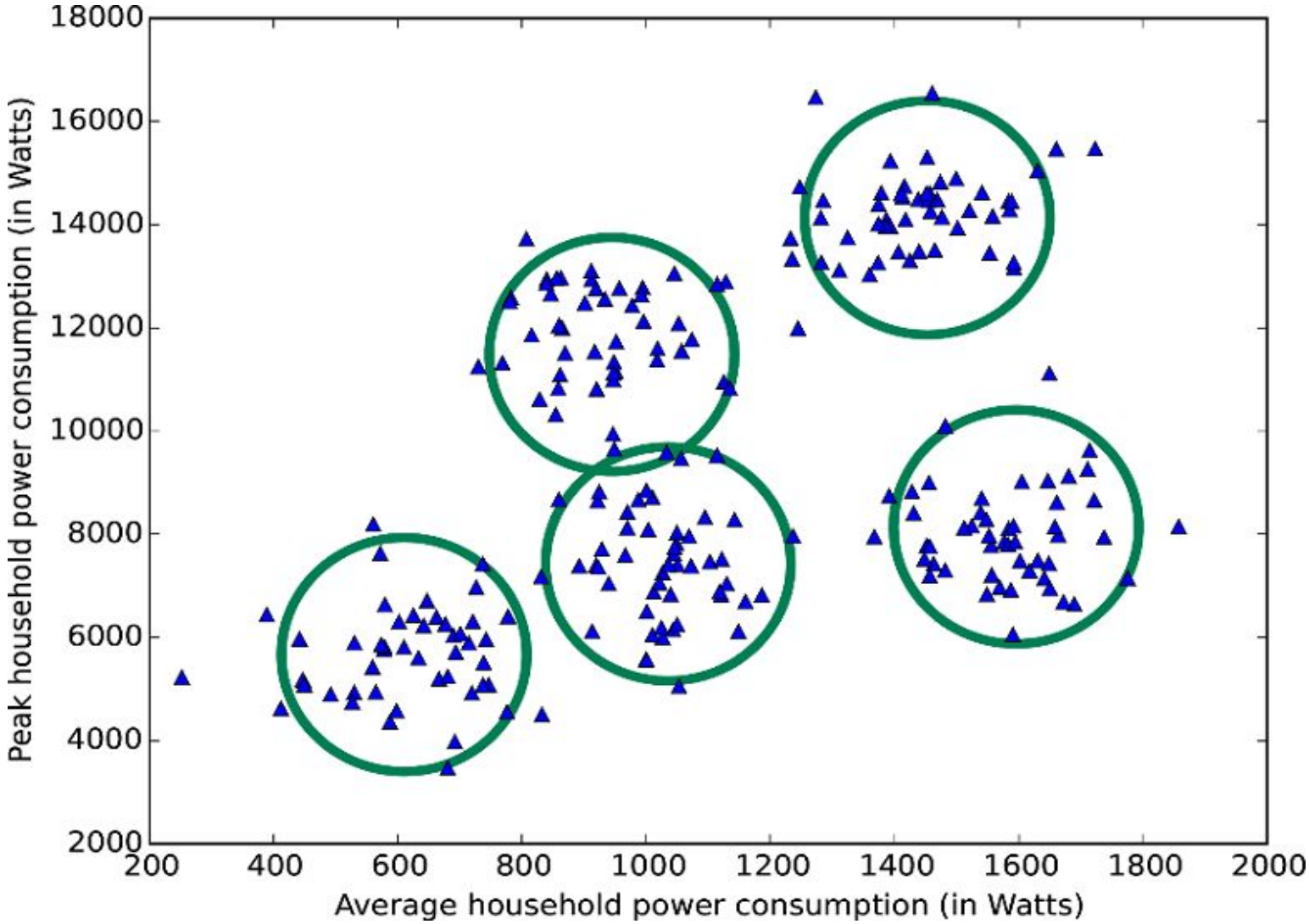
Recommendations





http://www.webdesignerdepot.com/2009/10/an-analysis-of-the-amazon-shopping-experience/

Clustering Example



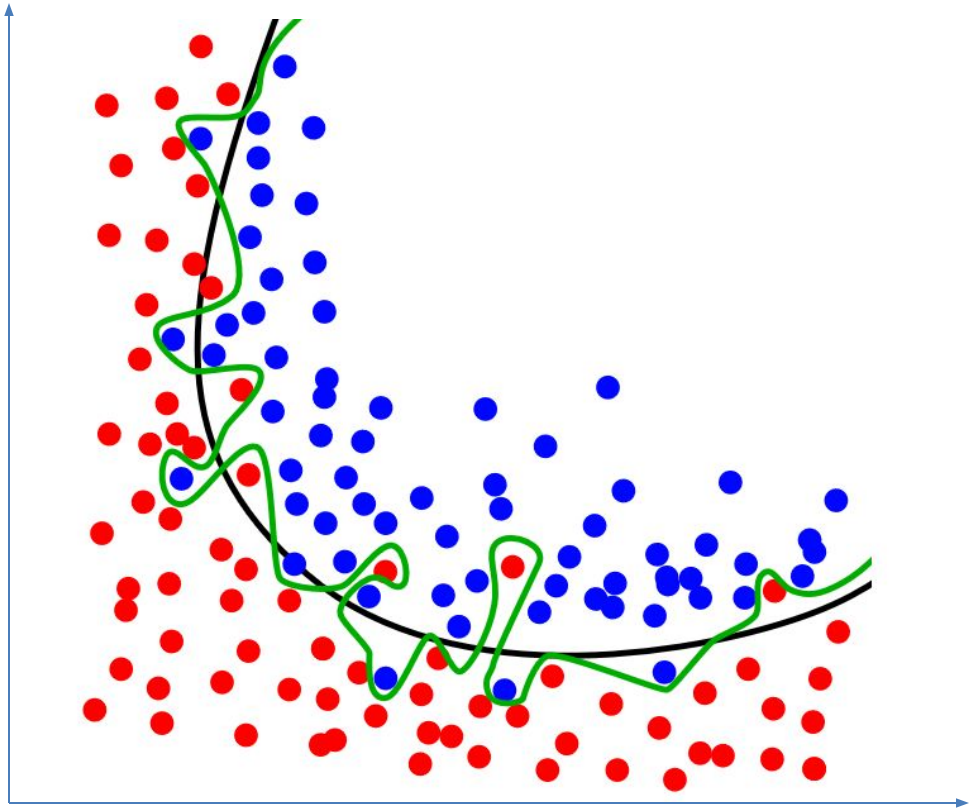
Pre-processing of data

Missing Values

- Reasons for missing values
  - Information is not collected
  - Attributes may not be applicable to all cases

- Handling missing values
  1. Eliminate Data Objects
  2. Estimate Missing Values
  3. Ignore the Missing Value During Analysis
  4. Replace with possible values

Over fitting



Evaluation Of Model

Evaluation of Model

- Partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3)
  - used for data set with large number of samples
- Cross-validation
  - divide the data set into k subsamples
  - use k-1 subsamples as training data and one sub-sample as test data --- k-fold cross-validation
  - for data set with moderate size

Metrics for Performance Evaluation

- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

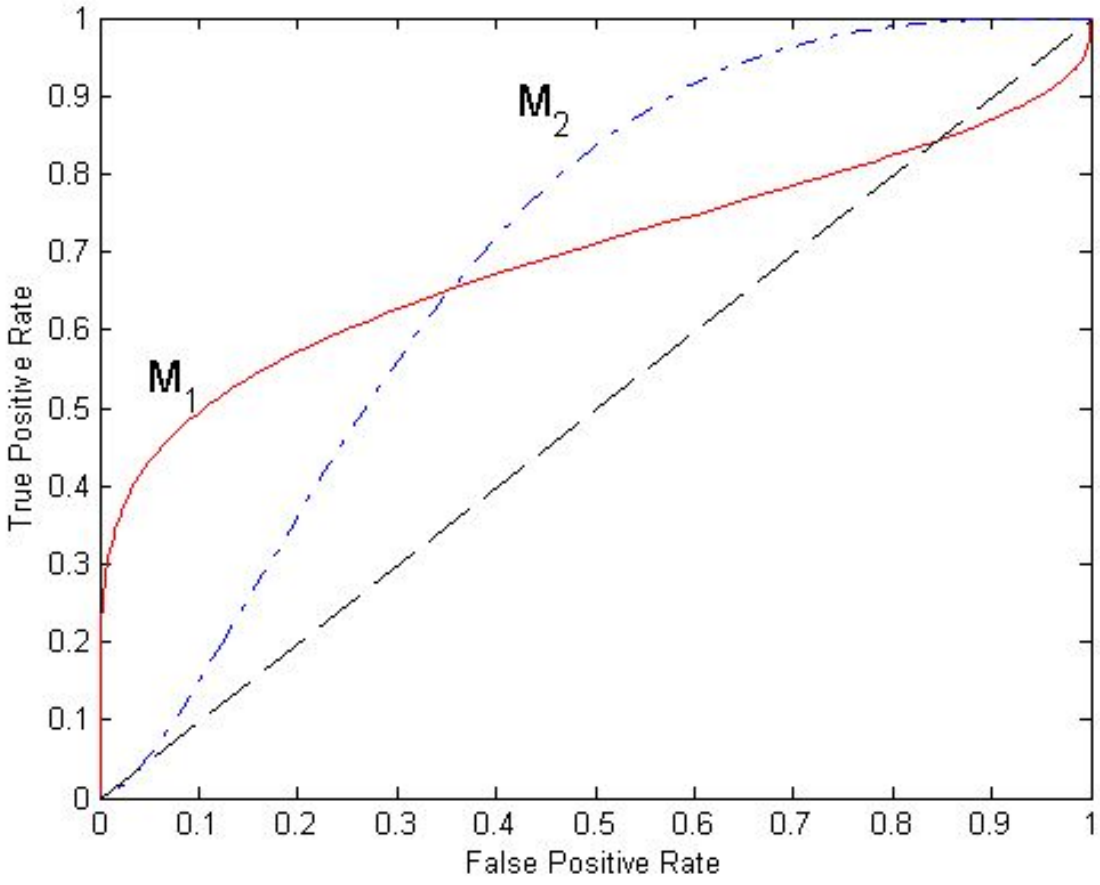
- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

ROC Curve

- ROC curve plots TP (on the y-axis) against FP (on the x-axis)

- Area under ROC curve is used to evaluate Any Model
- M_1 is better for small FPR
- M_2 is better for large FPR



Important Libraries

- Numpy
- Pandas : load data
- Sklearn
- Xgboost

Deep Learning Libraries

- Tensorflow
- Theano
- Keras
- Torch
- Nolearn
- Lassagne