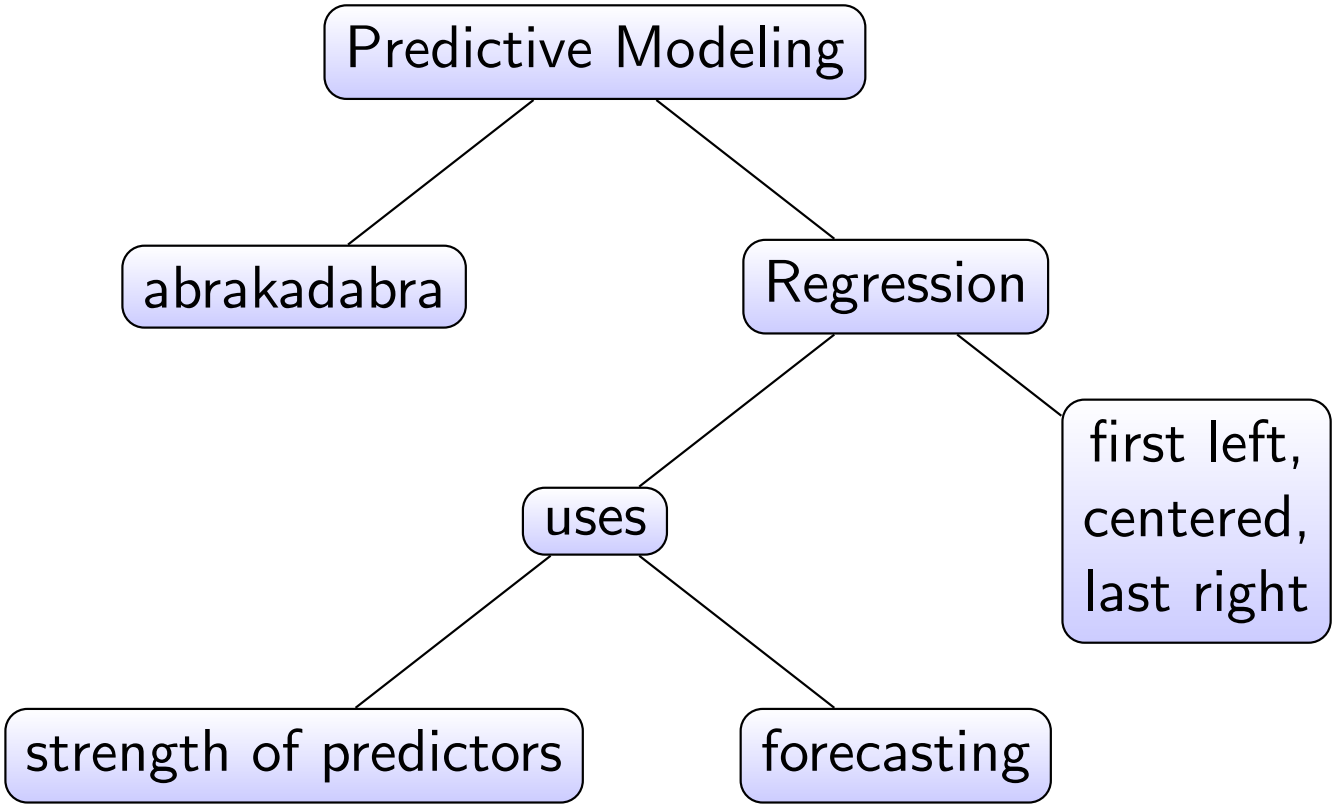
Stats for Data Science

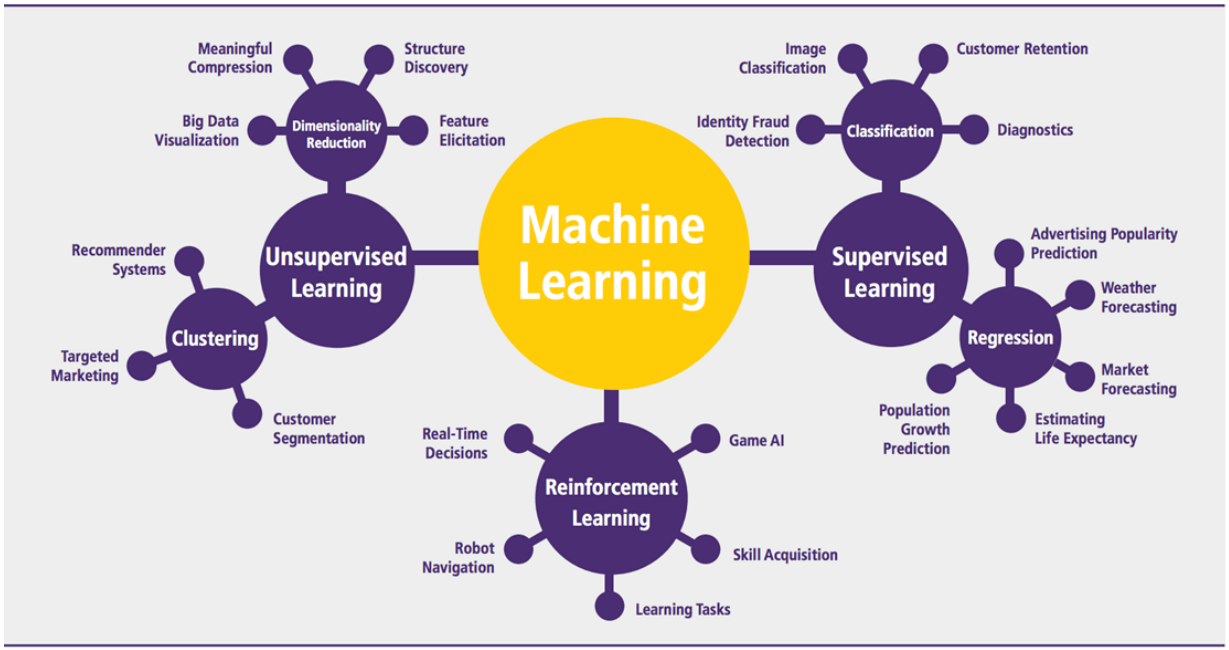
Saumya Bhatnagar

February 26, 2020

Table of contents I

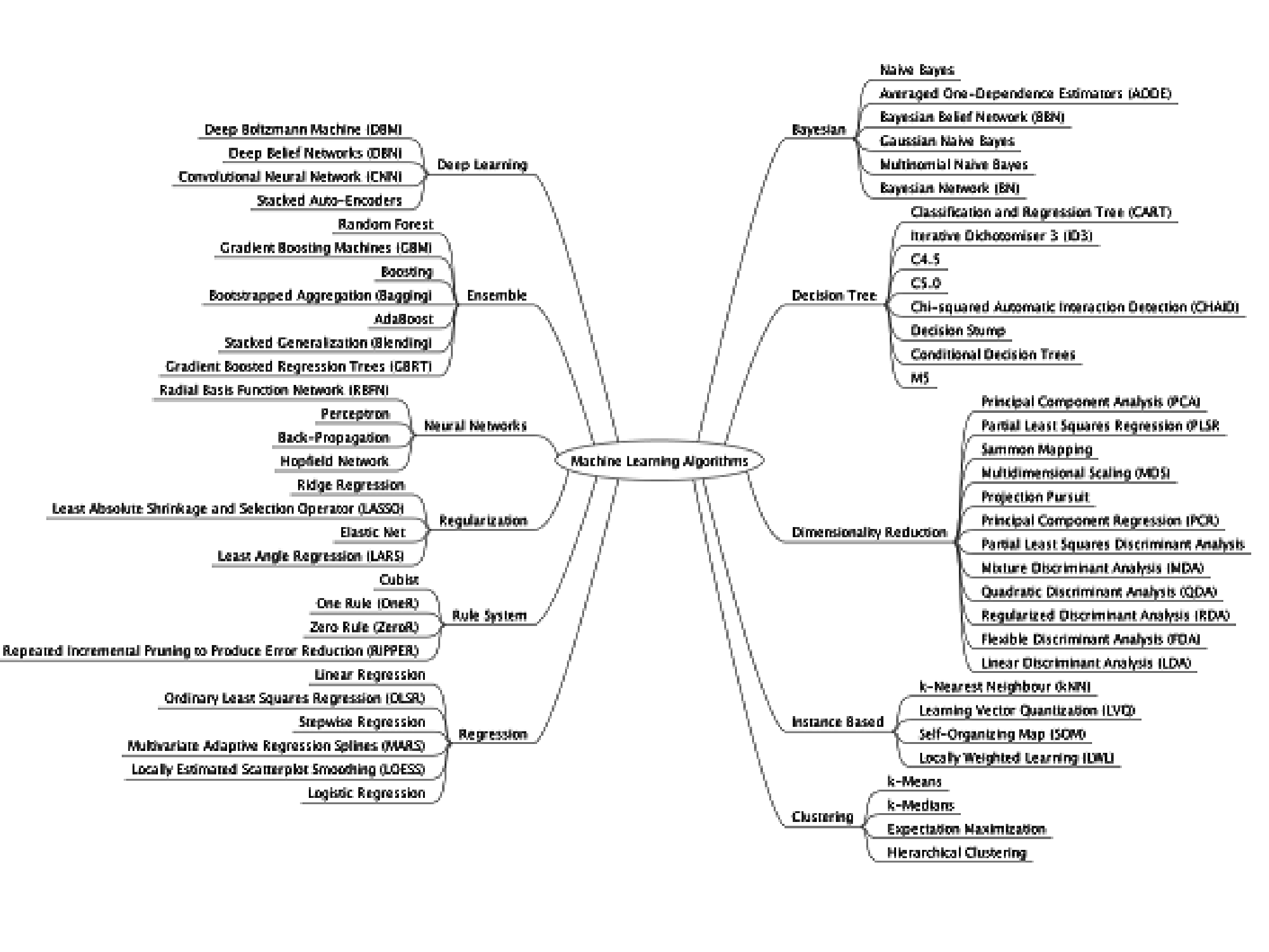
Table of contents II


Predictive Modeling
abrakadabra
Regression
uses
first left, centered, last right
strength of predictors
forecasting



	Categorial values	Continuous values
Supervised learning	**Two-class/Binary classification** • Decision Forest • Decision Tree • Naive Bayes/Bayes Classifier • (Deep) Neural Networks • Support Vector Machines **Multi-class classification** • Decision Forest • Logistic Regression • (Deep) Neural Networks	**Regression** • Bayesian Linear Regression • Decision Forest Regression • Decision Tree Regression • Linear Regression • Neural Network Regression • Ordinary Least Squares Regressior
Unsupervised learning	**Association Rule Learning** • Apriori algorithm • Frequent Pattern Growth **Classification** • Autoencoders • Restricted Boltzmann Machines	**Clustering** • Density-based Clustering • Hierarchical Clustering • Partitional Clustering (incl. K-means) **Dimensionality reduction** • Principal Component Analysis (PCA)

	Categorial values
Supervised learning	**Two-class/Binary cla** • Decision Forest • Decision Tree • Naive Bayes/Bayes • (Deep) Neural Netv • Support Vector Ma **Multi-class classificat** • Decision Forest • Logistic Regressior • (Deep) Neural Netv
Unsupervised learning	**Association Rule Lea** • Apriori algorithm • Frequent Pattern Gr **Classification** • Autoencoders • Restricted Boltzma


Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR)
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

content...

content...

content...

Univariate and Multivariate Analysis

Dimensionality Reduction

Under and Over Sampling

Class Imbalance

what is regression?

Regression analysis is a statistical technique to assess the relationship between an predictor variable and one or more response factors.
http://www.statisticshowto.com/probability-and-statistics/regression-analysis/ (go to definitions)

Linear vs Logistic

Basis	Linear Regression	Logistic Regression
Core Concept	The data is modelled using a straight line	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Used with	Continuous Variable	Categorical Variable
Output/Prediction	Value of the variable	Probability of occurrence of event
Accuracy and Goodness of fit	measured by loss, R squared, Adjusted R squared etc.	Accuracy, Precision, Recall, F1 score, ROC curve, Confusion Matrix, etc

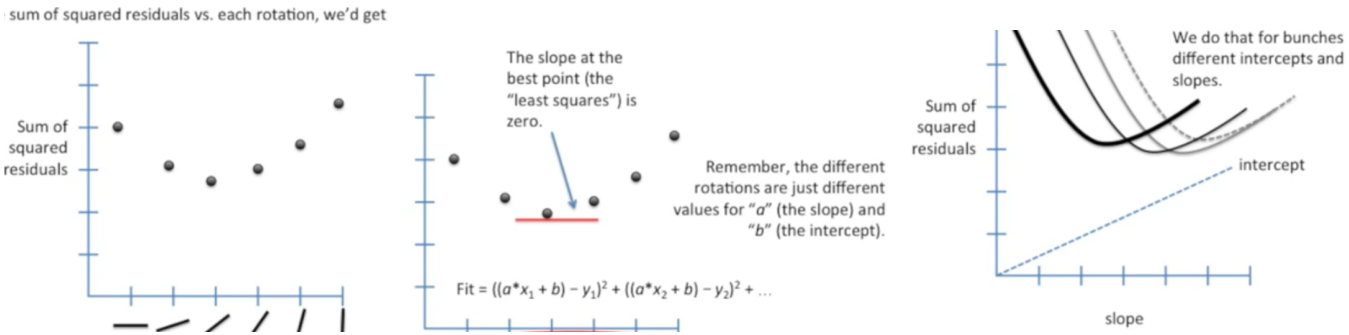
Data is modelled using a straight line vs using a sigmoid function maps continuous x to cont y, vs maps cont x to binary y

Outcome Variable	**GLM Family**	**Link**	**Mean to Variance**
Continuous, unbounded	Normal or Standard Gaussian	Identity	
Continuous, non-negative	Gamma or inverse Gamma		
Discrete/ counts/ rate	Poisson Quassi-poisson or negative binomial	Log	Identity If not Identity
Count	Gamma		Over dispersion
Counts with multiple zero	Zero inflated poisson may be checked for fitting		
Binary	Binomial or Logistic regression		
Nominal	Multinomial regression		

Regression Model Selection Criteria

Linear Regression using Least Squares

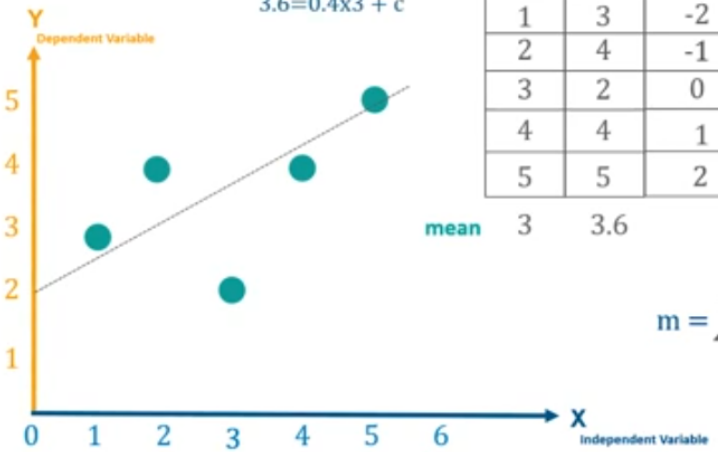
1. **fitting a line to the data**: as shown below
2. Find best fit using **Least Squares**
   sum of squared residuals $= \sum(y - \hat{y})^2$
3. Find goodness of fit using **R squared** method
   R squared (aka coefficient of determination) is a statistical measure of how well the data fits line
   low R squared doesn't always mean bad



Linear regression

$y = mx+c$

$3.6=0.4x3 + c$



x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean	3	3.6		$\Sigma = 10$	$\Sigma = 4$

$$m = \sum \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2}$$

x	y	$y-y$	$(y-y)^2$	y_p	(y_p-y)	$(y_p-y)^2$
1	3	-0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		Σ 5.2			Σ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\Sigma\left(y_p-\bar{y}\right)^2}{\Sigma\,(y-\bar{y})^2}$$

characteristics of Linear Regression

outliers have big bad impact
Computational complexity O(n)
comprehensible and transparent

Linear vs Multiple Regression

Linear

fit line

Calculate R^2

$R^2 = \frac{SS(mean_y) - SS(fit)}{SS(mean_y)}$

cal F-score and p-val

$F = \frac{\frac{SS(mean) - SS(fit)}{p_{fit} - p_{mean}}}{\frac{SS(fit)}{n - p_{fit}}}$

Multiple

fit plane or higher dimensional

Cal R^2

Adjust R^2 to compensate for additional parameters

Multiple Regression

Fit plane of higher dimensional obj to the data

Logistic Regression

content...

Three methods to classifier

1. model a classification rule - knn, decision tree, perceptron, svm
2. model the probability of class membership given input data - perceptron with cross-entropy cost
3. make a probabilistic model of data within each class - naive bayes 1 & 2 are discriminative classifications 3 is generative classification 2 & 3 probabilistic classification

Decision Tree

content...

Random Forest

content...

"Help me understand our customers better so that we can market our products to them in a better manner!

Monothetic: Cluster members have some common property Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

Polythetic: Cluster members are similar to each other. Distance between elements define relationship

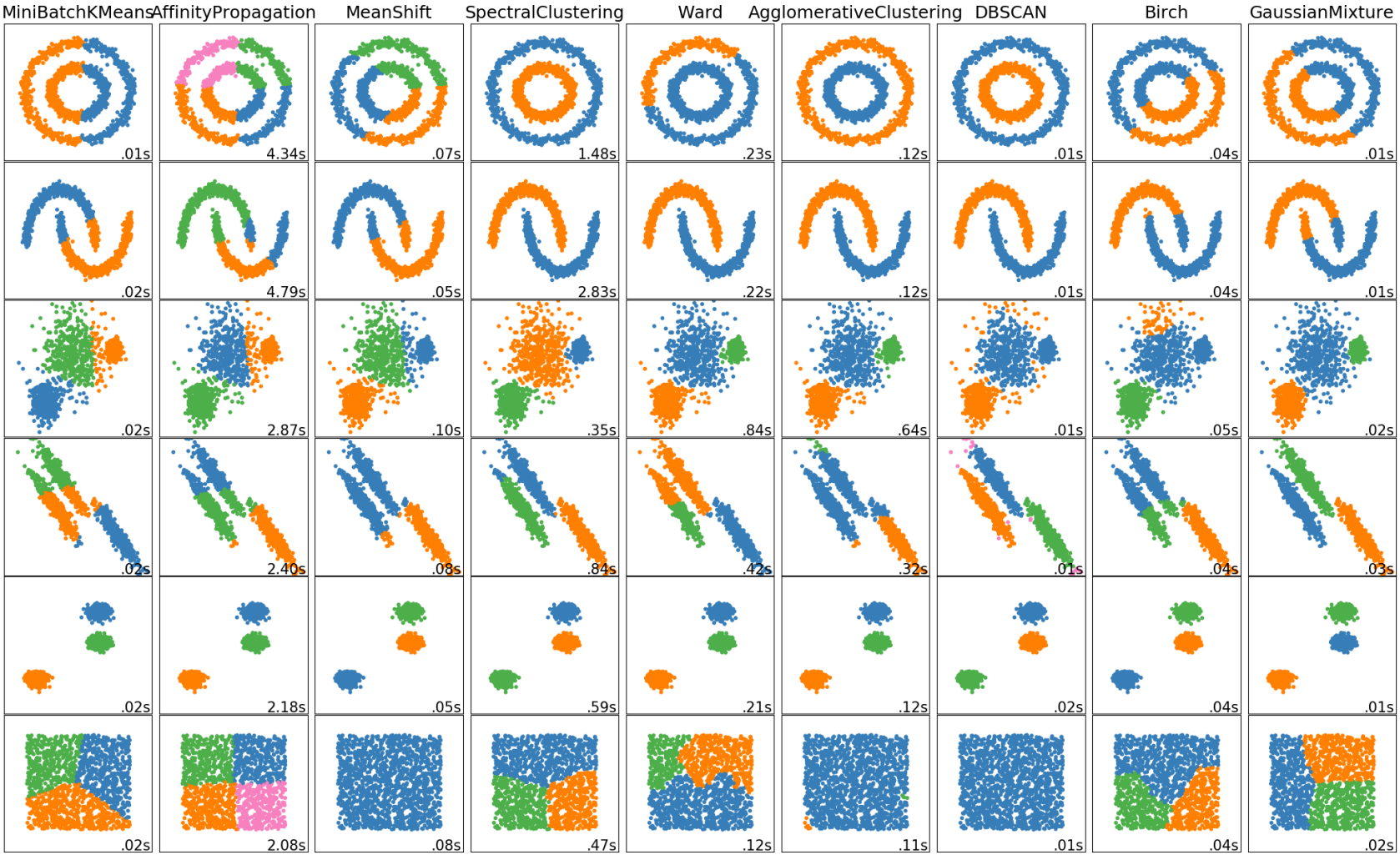
Hard Clustering: each data point either belongs to a cluster completely or not

Soft Clustering: a probability or likelihood of that data point to be in those clusters is assigned.

Clustering Models

Connectivity models	Distribution models	Centroid models	Density models
data points closer in data space exhibit more similarity to each other than the data points lying farther away	how probable is it that all data points in the cluster belong to the same distribution (e.g: Normal, Gaussian)	iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters	isolates various different density regions and assign the data points within these regions in the same cluster
hierarchical clustering	Expectation-maximization	K-Means, k-median	mean-shift, DBSCAN and OPTICS
Approaches: 1) Top-bottom, 2) bottom-up	EM uses multivariate normal distributions	DZA	DBSCAN uses radius ϵ and Center c
lacks scalability for handling big datasets, Time complexity: $O(n^2)$	These models often suffer from over-fitting. Prior knowledge to define num clusters	important to have prior knowledge of the dataset. results change in every trial	DBSCAN doesn't perform as well when the clusters are of varying density
Results are reproducible	more flexibility in terms of cluster covariance due to $\mu and \sigma$ (additional σ)	can handle big data , Time complexity: O(n)	DBSCAN identifies outliers as noises
chk1	elliptical shape (since we have a standard deviation in both the x and y directions	work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D)	DBSCAN: can find arbitrarily sized and arbitrarily shaped clusters
Angola	GMMs support mixed membership since is probability based	AGO	DBSCAN: drawback in high-dimensional data since the distance threshold ϵ becomes challenging to estimate

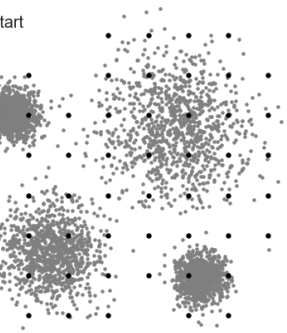
Clustering Types

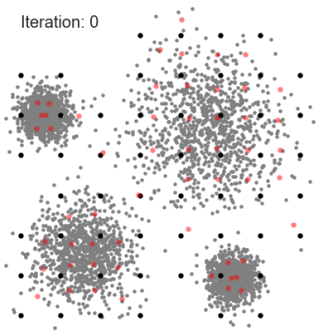


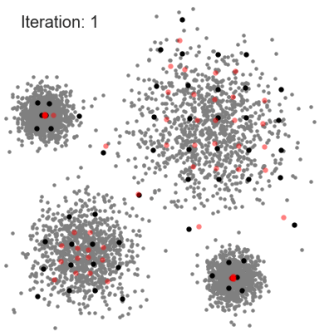
mean-shift clustering

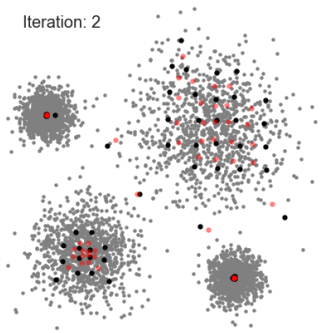
consider a set of points in two-dimensional space
a circular sliding window C centered and radius r as the kernel
hill-climbing algorithm that involves shifting this kernel iteratively to a higher density ($\propto$ number of points) region until convergence
At every iteration,
- shift the center point to the mean of the points within the window (hence the name)
-gradually move towards areas of higher point density
- until no longer increase in the density
- When multiple sliding windows overlap the window containing the most points is preserved. The data points are then clustered according to the sliding window in which they reside.

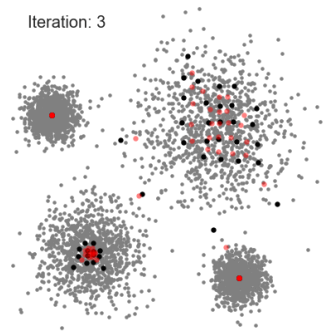
Density Models

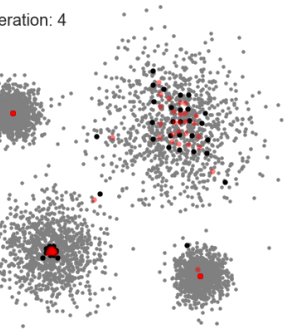


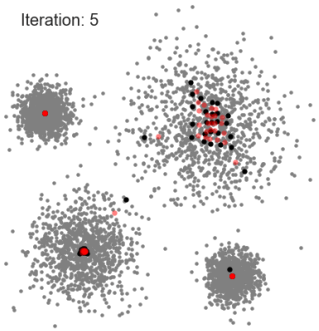


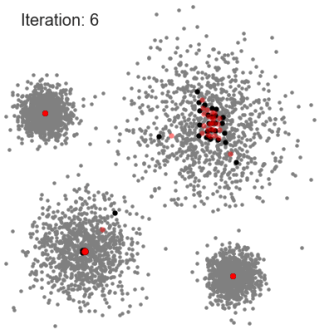


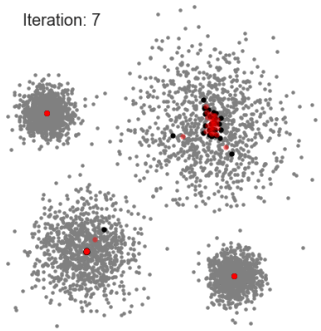


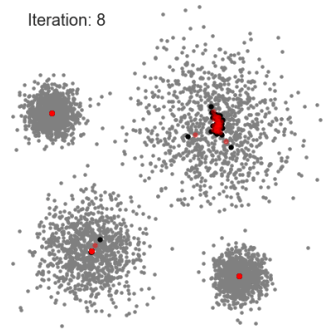


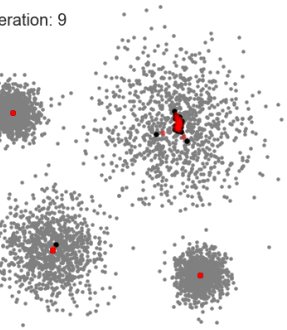


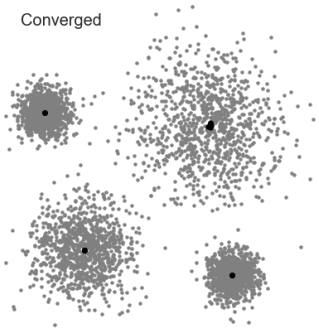


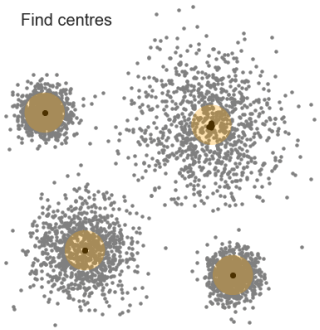


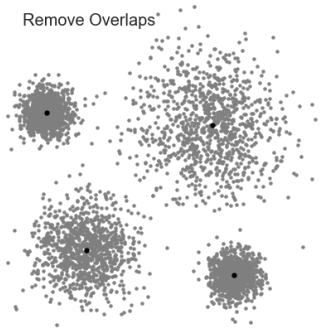












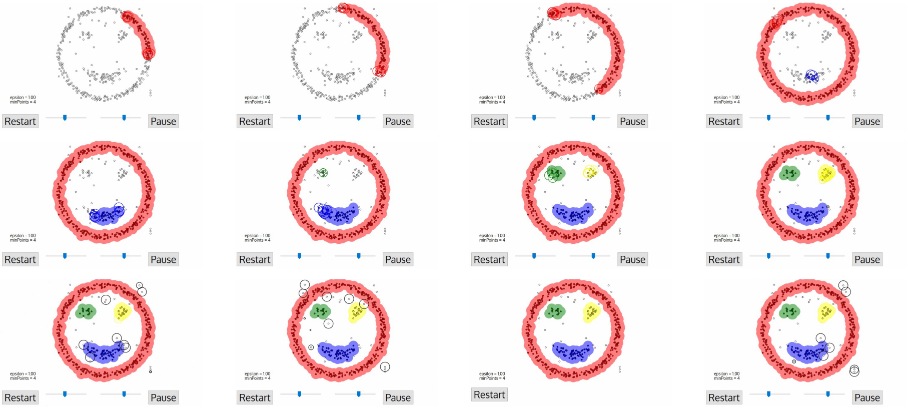
Density-Based Spatial Clustering of Applications with Noise-DBSCAN

-label all data point to be unvisited. For all unvisited points:

1. All points which are within the ϵ distance are neighborhood points (part of the same cluster)
2. If neighborhood points ¿= minPoints, then the clustering process starts and the current data point becomes the first point in the new cluster - Otherwise, mark the point as noise - In both cases that point is marked as "visited"
3. repeated for all of the new points in the cluster group
4. next an new unvisited point is retrieved and processed

Since at the end of this all points have been visited, each point will have been marked as either belonging to a cluster or being noise.

Density Models



Hierarchical DBSCAN - HDBSCAN

content...

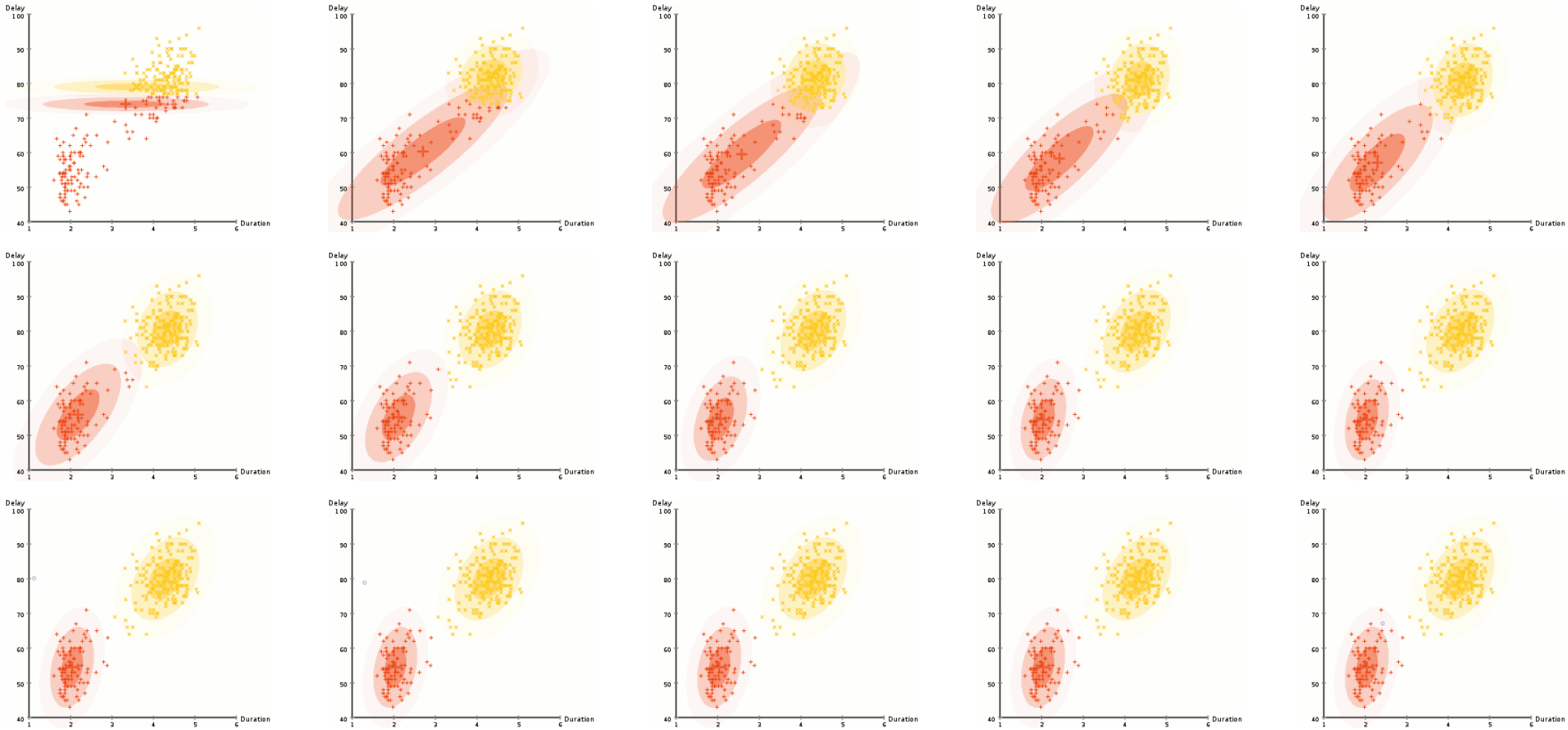
Gaussian Mixture Models (GMMs)

Assumption: the data points are Gaussian distributed (parameters: the mean and the standard deviation)! Each Gaussian distribution is assigned to a single cluster. To find the parameters of the Gaussian for each cluster, use an optimization algorithm called Expectation–Maximization (EM).

Expectation–Maximization (EM) using GMM

choose num of clusters
compute the probability that each data point belongs to a particular cluster. With a Gaussian distribution we are assuming that most of the data lies closer to the center of the cluster.
From probabilities $\rightarrow$ recompute set of parameters such that we maximize the probabilities of data points within the clusters
We compute these new parameters using a weighted sum of the data point positions, where the weights are the probabilities of the data point belonging in that particular cluster.
Repeat till convergence

Distribution Models



K means

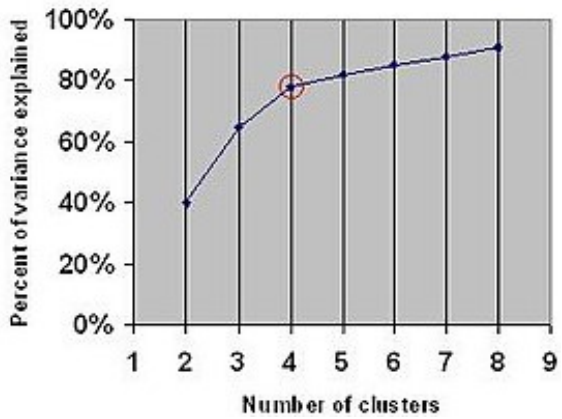
iterative clustering algorithm that aims to find local maxima in each iteration

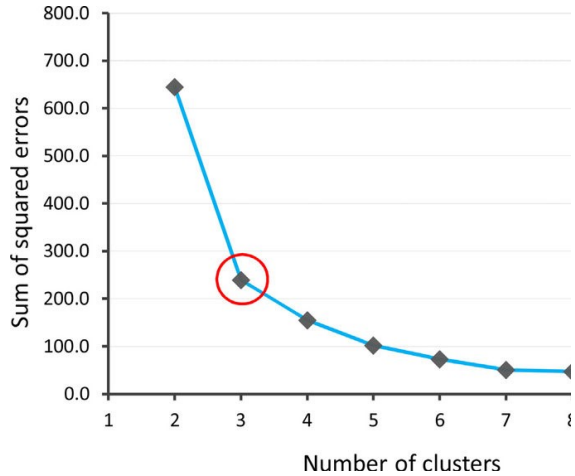
take a quick look at the data and choose k (num clusters)

assign data points to the cluster $\leftrightarrow$ compute cluster centroid

repeat to reduce variation error

elbow method



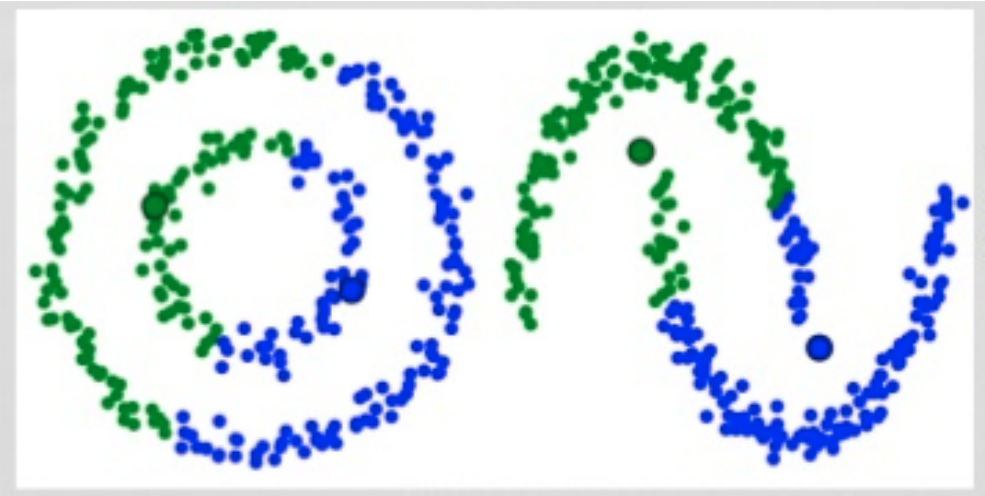


k-median

K median vs kmeans: instead of recomputing the group center points using the mean (like in K-Means) we use the median vector of the group. This method is **less sensitive to outliers** (because of using the Median) but is **much slower for larger datasets as sorting is required** on each iteration when computing the Median vector

mean-shift vs kmeans: Instead of selecting the number of clusters as mean-shift automatically discovers this (advantage), the selection of the window size/radius "r" can be non-trivial.

kmeans fail



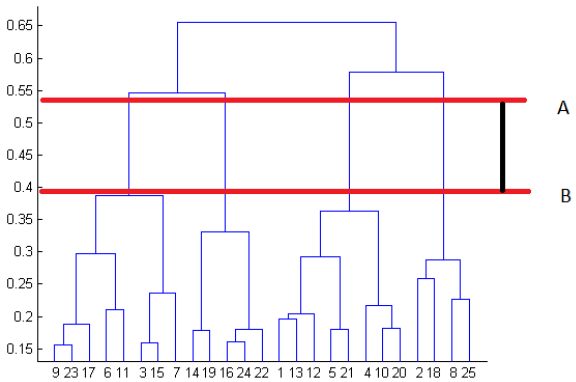
K-Means is actually a special case of GMM in which each cluster's covariance along all dimensions approaches 0

Agglomerative Hierarchical Clustering

The decision of dividing into or merging **two** clusters is taken on the basis of closeness of these clusters. Metrics for deciding the closeness of two clusters:
Euclidean distance: $||a-b||_2 = \sqrt{\sum(a_i - b_i)}$
Squared Euclidean distance: $||a-b||_2^2 = \sum(a_i - b_i)^2$
Manhattan distance: $||a-b||_1 = \sum |a_i - b_i|$
Maximum distance: $||a-b||_{INFINITY} = \max_i |a_i - b_i|$
Mahalanobis distance: $\sqrt{(a-b)^T S^{-1}(-b)}$
Maybe, use **average linkage** which defines the distance between two clusters to be the average distance between data points in the first cluster and data points in the second cluster.

hierarchical agglomerative clustering (HAC) or bottom-up

1. Each data point as a single cluster
2. select a distance metric
3. Iterate till convergence
   - combine two clusters with the smallest average linkage

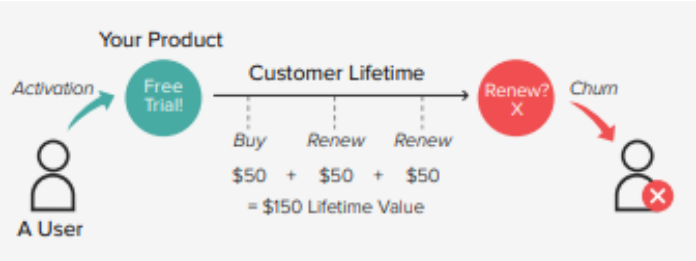


The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space. take 4 clusters as the red horizontal line in the dendrogram covers maximum vertical distance AB.

Churn or Retention Analysis

Customer Retention Rate: The percentage of customers who repurchase in a given time period compared to an equal and preceding time period

Churn Rate: The inverse of Customer Retention Rate, or the percent of users who did not repurchase or whom you lost



proactive churns: losing customers due to cancellations

passive churn: failures to renew

Cohort Analysis and Life Time Value (LTV)

LTV: The expected amount of profit/revenue from a user
CLV = NPV (net present value) of the sum of all future revenues from a customer, minus all costs associated with that customer
Why LTV: - Tracking your LTV to Customer Acquisition Cost (CAC) ratio: Companies typically use the 3:1 CAC ratio or Cost Per Acquisition (CPA)
- Evaluating your most valuable marketing channels
- Focus on retaining your most valuable customers
Historic CLV: sum of the gross profit from all historic purchases for an individual customer

Avg Order Value, AOV $= \frac{Revenue}{Orders}$; Avg Purchase Rate $= \frac{Orders}{NumCustomers}$

Avg Customer Value $= \frac{AvgPurchaseVal}{AvgPurchaseRate}$

Avg Customer Lifespan $= \frac{SumCustomerLifespans}{NumCustomers}$ gives LTV

LTV =
- Avg Customer Value X Avg Customer Lifespan
- ARPU X $\sum_{n=0}^{N}(1 - CR)^t$ .... [N=num months to examine]
- ARPU/CR_n ...[ARPU=Avg rev per User, for n months]
- ARPU/CR_n x DR ..... [for variable churn & n months]
- ASP/CR + m(1-CR)/CR^2 ... [for account expansion]

CLV =
- AGM x $\sum_{0}^{numTransactions}$ *Transaction* [AGM=Avg Gross Margin]
- $((T_{avg} x AOV)$AGM)ALT =GML(gross margin per user lifespan)
- GML(R/(1+D-R)); [account expansion]

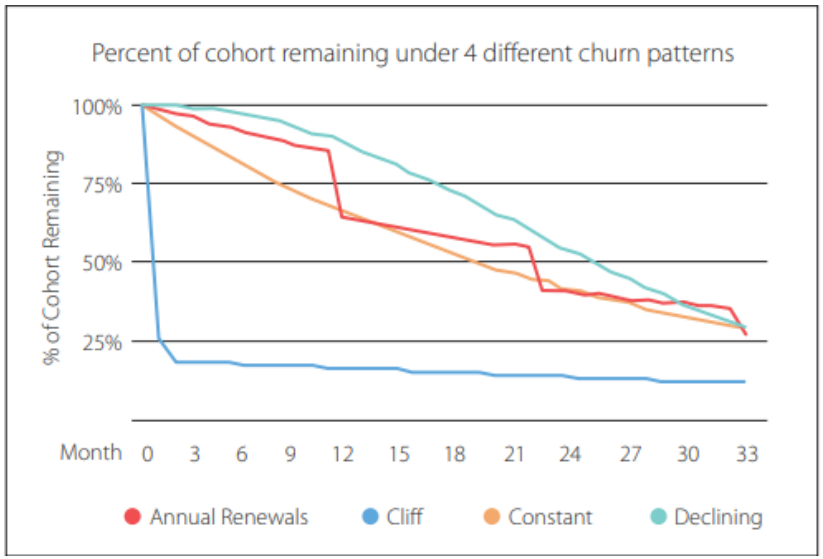
CR=Churn rate; ASP=Avg Selling Price; m=↑ARPU/user/month;
T_avg = avg monthly transactions; ALT=avg User Lifespan (in months)
D=monthly discount rate; R=monthly retention rate; DR=Discount Rate
to adjusts for mix churn (Annual Renewals, Constant, Declining and Cliff

Steps to LTV:

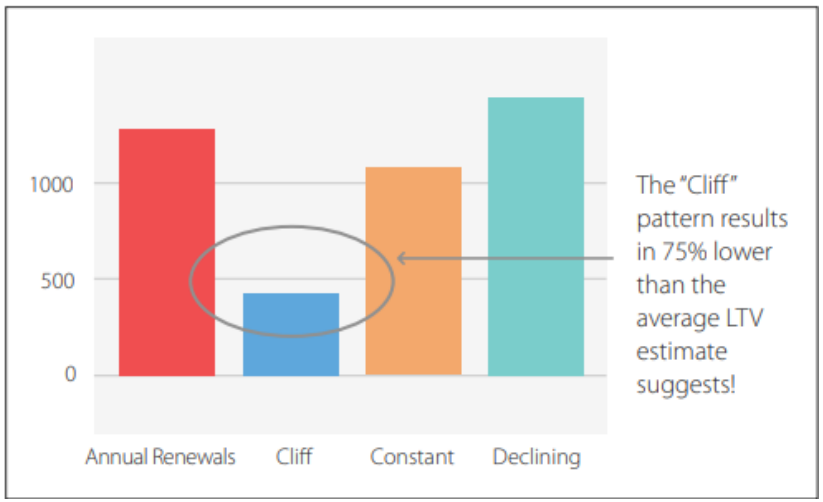
- Normalizing to Acquisition Date: Bin users into buckets like Day 0, Day 1, Day 2 or Week 0, Week 1, Week 2, and so on.
- Normalized to a Closed Time Limit: Broader questions "What is my total CLTV," should be replaced with "What is our 3-year or 5-year LTV?" & should be based on:
  1. Average Customer Lifespan
  2. Customer Retention Rate
  3. Churn Rate: The inverse of Customer Retention Rate.
  4. Time to General Profitability Against Acquisition Costs: If your business is a "Loss Leader" Model this time may be a longer length than businesses with lower acquisition costs and lower profitability.
  5. Rate of Discount

Types of churn and LTV

- Annual Renewals: larger churn at each contract renewal.
- Cliff churn: majority of the churn within the first month, and then a small constant churn thereafter.
- Constant: steady, constant churn rate (shown as 3.5%).
- Declining: churn rate starts at zero, increases each month.



Effect on LTV:



Types

content...

The Kaplan-Meier curve

content...

The log-rank test

content...

Cox proportional hazards regression

content...

Parametric models

content...

Frailty models

content...

Competing risk models

content...

Discrete Time Model using logistic regression

content...

Social Network Analysis

content...

Trend Analysis

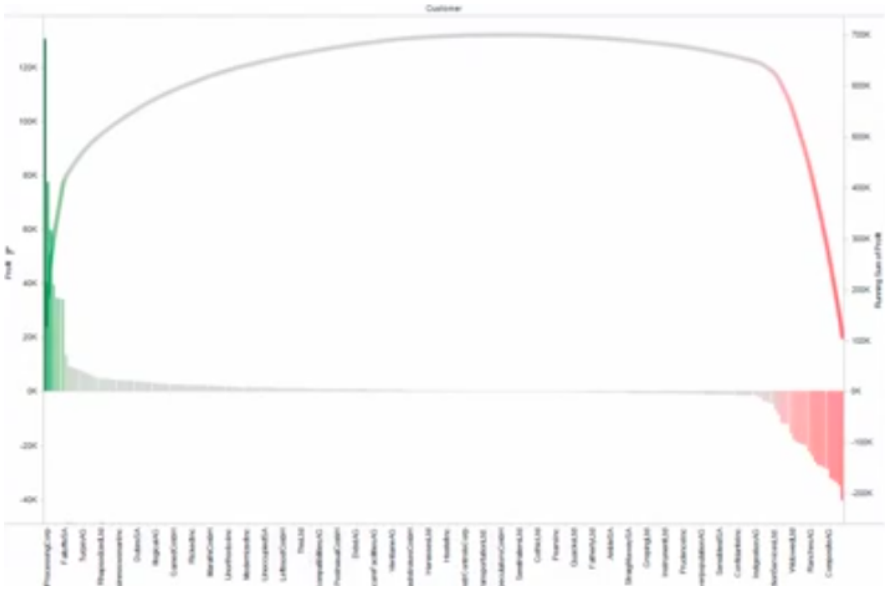
content...

Whale Curve Analysis

Technique to visualize the data

Sort the data before plotting

Pareto Principle: for many events, roughly 80% of the effects come from 20% of the causes



"Loss Leader" Model

"Loss Leader" Model, where you introduce new customers at a high cost in the hope of building a customer base or securing future revenue?

Market Basket Analysis

content...

Propensity of Cross-sell

content...

Thank You!