- Split data accord. to some criterion, classify new (o) accord. to same criterion - P(new (o) belong to each group) calc New (o) assigned to group with highest prob. Discriminant Analysis (Supervised learning)	Classification		
- New (o) assigned to group with highest prob. Discriminant Analysis (Supervised learning) Groups specified (can identify from available data) Find hidden assoc. structures within unlabeled data Logistic regression Clustering Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Freliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	- Split data accord. to some criterion, classify new (o) accord. to same criterion		
Groups specified (can identify from available data) Groups specified (can identify from available data) Groups specified, need to find from data itself Find hidden assoc. structures within unlabeled data Logistic regression Clustering Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Group similar objects but finds degree and nature of similarity by itself Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods 2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters don't change (i.e. convergence) Weather the from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Spilt data into groups			
Groups specified (can identify from available data) Logistic regression Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters don't change (i.e. convergence) 5. Final k clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups			
Logistic regression Logistic regression Clustering Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Clustering Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) Recompute centroids (cluster mean) for each updated cluster Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups			
Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Freliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	Groups specified (can identify from available data)		
Training data: data used to build model Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known			
Misclassification: predict wrongly (Goal: min. this error) Quote prob. rather than binary 0-1 decision Labels of response variable known Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters from their respective centroids Versum of til. dist. of (o) in each cluster from their respective centroids Uses a tree structure to create a rule for classifying data Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups			
Quote prob. rather than binary 0-1 decision Labels of response variable known Similarity by itself - Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted Determine groups for discriminant analysis		Survival analysis: Failures of Samsung phones	
Labels of response variable known similarity by itself - Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted - Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods 2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids - Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups			
- Preliminary classifier: rough labelling of (o) before thorough model-based analysis attempted - Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods 2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	,		
thorough model-based analysis attempted Determine groups for discriminant analysis K-means clustering 1. Choose value of k, guess for k cluster centriods 2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters of Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree Uses a tree structure to create a rule for classifying data Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	Labels of response variable known		
1. Choose value of k, guess for k cluster centriods 2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		thorough model-based analysis attempted	
2. Compute distance of each pt from k centroids (Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		K-means clustering	
(Assign each pt. to cluster with closest centroid) 3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		Choose value of k, guess for k cluster centriods	
3. Recompute centroids (cluster mean) for each updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		2. Compute distance of each pt from k centroids	
updated cluster 4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		(Assign each pt. to cluster with closest centroid)	
4. Con't. step 2 & 3 until composition of clusters don't change (i.e. convergence) 5. Final k clusters Choice of k Min. WSS (Within sum of squares) of clusters Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree Uses a tree structure to create a rule for classifying data Sequence of decisions specifying consequences Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		3. Recompute centroids (cluster mean) for each	
don't change (i.e. convergence) 5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		updated cluster	
5. Final k clusters Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		4. Con't. step 2 & 3 until composition of clusters	
Choice of k - Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		don't change (i.e. convergence)	
- Min. WSS (Within sum of squares) of clusters - Measures sum of ttl. dist. of (o) in each cluster from their respective centroids - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		5. Final k clusters	
- Measures sum of ttl. dist. of (o) in each cluster from their respective centroids - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		Choice of k	
Decision Tree - Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups from their respective centroids		- Min. WSS (Within sum of squares) of clusters	
Uses a tree structure to create a rule for classifying data - Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups		- Measures sum of ttl. dist. of (o) in each cluster	
 Uses a tree structure to create a rule for classifying data Sequence of decisions specifying consequences Used for prediction Build: Full data with all its var. and (o) Sequence of if-then statements: Split data into groups 		from their respective centroids	
- Sequence of decisions specifying consequences - Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	Decision Tree		
- Used for prediction Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	- Uses a tree structure to create a rule for classifying data		
Build: 1. Full data with all its var. and (o) Sequence of if-then statements: 2. Split data into groups	- Sequence of decisions specifying consequences		
 Full data with all its var. and (o) Sequence of if-then statements: Split data into groups 	- Used for prediction		
Sequence of if-then statements: 2. Split data into groups	Build:		
2. Split data into groups	1. Full data with all its var. and (o)		
, and the second	Sequence of if-then statements:		
3 Fach group split into subgroups	2. Split data into groups		
5. Lach group split litto subgroups	3. Each group split into subgroups		
4. Splitting con'ts till certain stopping rule is reached	4. Splitting con'ts till certain stopping rule is reached		
or each (o) becomes a group of its own	or each (o) becomes a group of its own		
Classification Tree Regression Tree	Classification Tree	Regression Tree	
- o/p var.s usually applied to be categorical in nature - o/p var.s can be numeric or cts	- o/p var.s usually applied to be categorical in nature	- o/p var.s can be numeric or cts	
- binary decisions like 'yes' or 'no'	- binary decisions like 'yes' or 'no'		

Association rules (Unsupervised learning)

- Uncover r/s of form $X \rightarrow Y$: when X (o), item Y (o)
- Descriptive, not a predictive method
- Itemsets: items bought in the same transaction, webpages visited in the same sitting etc.
- k-itemset: { item 1, item 2, ..., item k }

Apriori algorithm

- Given collection of items, explores all subsets of items
- Provides subsets which appear more than some pre-defined frequency
- Need transaction database D, min. length support threshold , max. length an itemset could reach N (optional)

1.	Support	Given itemset L, ~ of L: proportion of transactions that
		contain L (0 < support < 1)
2.	Frequent itemset	~ has items that appear tgt. often enough
		(satisfy a min. support criterion, usually set at 0.5)
3.	Downward closure property (Apriori property)	If itemset considered frequent, any subset of freq. itemset
		also frequent.

Steps:

- A) Bottom-up approach: Starting from 1-itemset, merges pairs of current itemsets to create new itemsets
- B) Each stage: Algorithm checks if all itemsets satisfy min. support criterion (If not, dropped and not considered in rest of algorithm)
- C) Process continues till it runs out support or itemsets reach a predefined max. length

SQL (Structured Query Language)

In-database analytics: describes the processing of data within its repository

- ©: database can update regularly → produce most up-to-date results
 - : Eliminate need to move data from place to place
 - : Well-protected/ Security -> need to extract to another portal to make changes/ perform analysis

Relational database: part of RDBMS (Relational Database Management system)

- c: organizes data in tables with established r/s between tables
 - : splitting data into tables \rightarrow no need to store entire table of info
 - : can change a specific part \rightarrow no need to change everything and store in same place
 - 1. Smaller memory to handle them
 - 2. Various tables stored on diff. machines
 - 3. Changes & corrections easily made
 - 4. Relational database: number of duplicates reduced

Parallel Computation

Distributed: ~ of computation happens at processor level

: compilers optimised to distribute computation whenever

Parallel: run a function repeatedly in diff. processors

: snowFT \rightarrow run full function in parallel

Residual bootstrap

Given dataset (x1, x2, ... xn)

- Calculate mean, var.
- Resample the data of size n with replacement (means pts can be repeated)
- Recalc. Data
- Do for a lot of times (thousands?)

Can show that sample distribution will be close to the true distribution of the r.v. looking at