TECHNICS IN ANALYTICS: BIRD'S EYE VIEW

Analytics

Techniques

Machine Learning

Supervised	Unsupervised
Observations from past - Known outcome	Observations from past - No associated outcome
Examples - Age 40, high blood pressure => diabetes - Additional income, Earn $ => spam	Examples - Power forward, 25 ppg - Income $50,000, single

Classification

- Example
  - John Doe: 40 age, 180lbs, no exercise = > diabetes
  - John Doe Jr.: 25 age, 150 lbs, exercise => no diabetes
  - John Doe Sr.: 40 age, 210 lbs, no exercise => diabetes
- John Doe Tiny
  - 38 age, 150 lbs, exercise
  - How likely to have diabetes?
  - Should he be considered for treatment?

Classification

- Email
  - Email 1 => spam
  - Email 2 => no spam
  - Email 3 => spam
- Extract words from emails
  - Word 1 => spam
  - Word 2 => no spam
- New email with words
  - Spam or not spam?

Loss Functions

- X = set of historical data
  - $(x, y) \in X$
  - x = feature vector
  - y = label
- Given x, f(x;w) is prediction
  - w unknown parameters
- Goal
  - $f(x; w) \approx y$
  - $l(f(x; w), y)$ = loss function
    - Penalize the discrepancy

$$\min_w \sum_{(x,y) \in X} l(f(x; w), y)$$

$$\min_w \sum_{(x,y) \in X} \hat{l}(x, y; w)$$

$$\min_w \sum_i \hat{f}_i(x_i; w)$$

$$\min_w E_{X,Y}(\hat{l}(x, y; w))$$

Linear Regression Loss Function

- $f(x; w) = w_1 x + w_o$
  - $w = (w_1, w_0)$
  - All are vectors
- $l(u, v) = (u - v)^2$
- $l(f(x; w), y) = (f - y)^2 = (w_1 x + w_0 - y)^2$
- $\hat{l}(x, y; w) = (w_1 x + w_0 - y)^2$

$$\min_{w_1, w_0} \sum_i (w_1 x_i + w_0 - y_i)^2$$

Feature Preparation

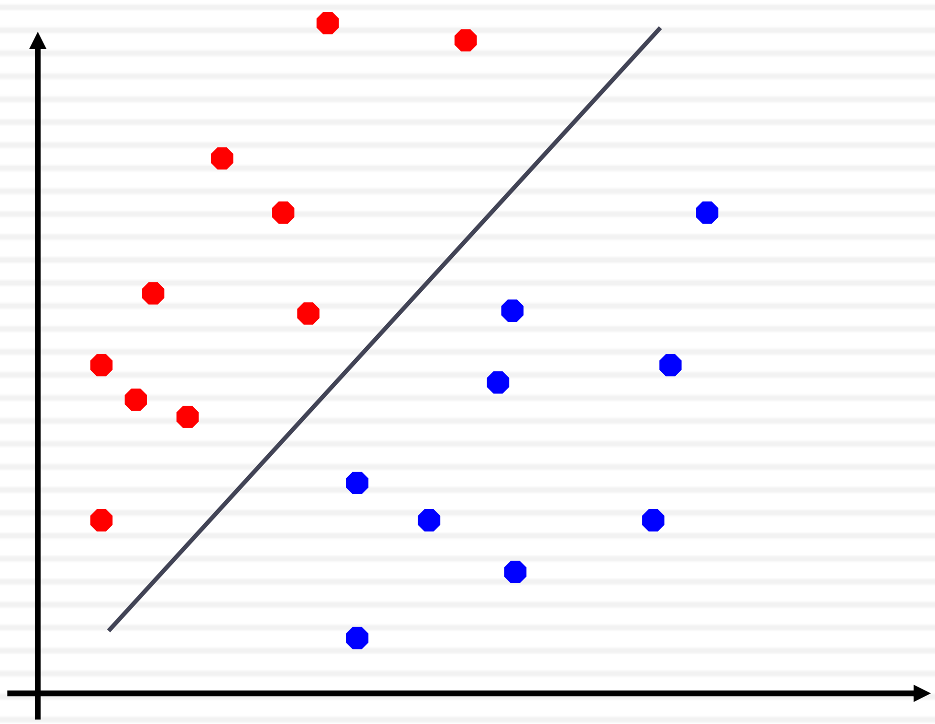
- Large/big feature values pose problems
  - Normalization
- For each feature
  - Subtract the mean and
  - Divide by standard deviation
- Standardization
  - Transform to [0,1]
  - Divide by the maximum value after the minimum value is translated to 0

Terminology

- ML synonyms
  - Record, sample, observation, row
  - Ground truth, label
- Classes
  - Labels belong to classes
- Parametric models have loss functions
  - Non-parametric do not

Support Vector Machines

- Binary classification
- Maximize the distance from the line



Support Vector Machine Loss Function

- Margin $= \frac{2}{||w||}$
  - Want to maximize
- Complete separation

$$\min_{w,b} ||w|| \text{ subject to } y_i(w^T x_i + b) \geq 1 \; \forall \; i$$

- General
- $\min_{w,b} ||w|| + C \sum_i \max(0, 1 - y_i(w^T x_i + b))$
- $f(x; w, b) = \text{if } w^T x + b \geq 0$

$$y_i = 1$$

$$\text{else } y_i = -1$$

- $\hat{l}(x, y; w, b) = \frac{||w||}{n} + C \max(0, 1 - y(w^T x + b))$

Classification

- Training set
- Fit function
  - From training set tune parameters of function
- New entity
  - Evaluate the function
  - Classify entity

Evaluation

- Data set
  - Split to training and test
  - Calibrate model on training
  - Evaluate on test
    - How many predictions are correct on test data
- Measure is regression error
  - Is it possible to achieve zero error?

10-fold validation

- Split the data into 10 chunks
  - Train on 9
  - Validate on 1
  - Repeat for all 10 test chunks/folds
- Model selection
  - Not parameter/weight selection
  - Not one set of parameters/weights on all 10

Precision and Recall

- Confusion matrix

True Positive

False Negative

		Predicted Class		
		+	-	
Actual Class	+	tp	fn	C = tp+fn
	-	fp	tn	
		A = tp+fp		T = tp+fn+fp+tn

False Positive

True Negative

Precision and Recall

- Accuracy
  - Proportion of correctly predicted

    all correct/all = (tp+tn)/(tp+tn+fp+fn)
- Recall
  - Proportion of + correctly predicted

    true positive/all positive = tp/(tp+fn)
- Precision
  - Proportion of + among all predicted +

    true positive/ predicted positive = tp/(tp+fp)
- F-measure
  - Harmonic mean of precision and recall

    2 recall * precision / (recall + precision)

Example

Results from Classification Algorithm

ID	Actual Class	Predicted Class
1	+	+
2	+	+
3	+	+
4	+	+
5	+	-
6	-	+
7	-	+
8	-	-

Confusion matrix

		Predicted Class		
		+	-	
Actual Class	+	4	1	C = 5
	-	2	1	
		A = 6		T = 8

- **True positive = 4**
- **False positive = 2**
- **True Negative = 1**
- **False Negative =1**

Categorical Features

- Label encoding
  - Assign a number to each category
  - What number?
- One-hot-encoding
  - Large number of features
  - Tight models
    - Does not scale

Advanced Target Encoding

- One category
  - Percentage of samples of this category in a class
- Issues
  - Data leakage
    - Get the values based on a hold-out set
  - What if only a few samples of a given category
    - Spurious correlation

Smoothed Tae Encoding

- Combine 'local' estimates with 'global' estimates
- Global ratio of two classes
- One category

$$\frac{ratio_{category} + K\, ratio_{global}}{n + m}$$

n = number of samples in this category

m = total number of samples

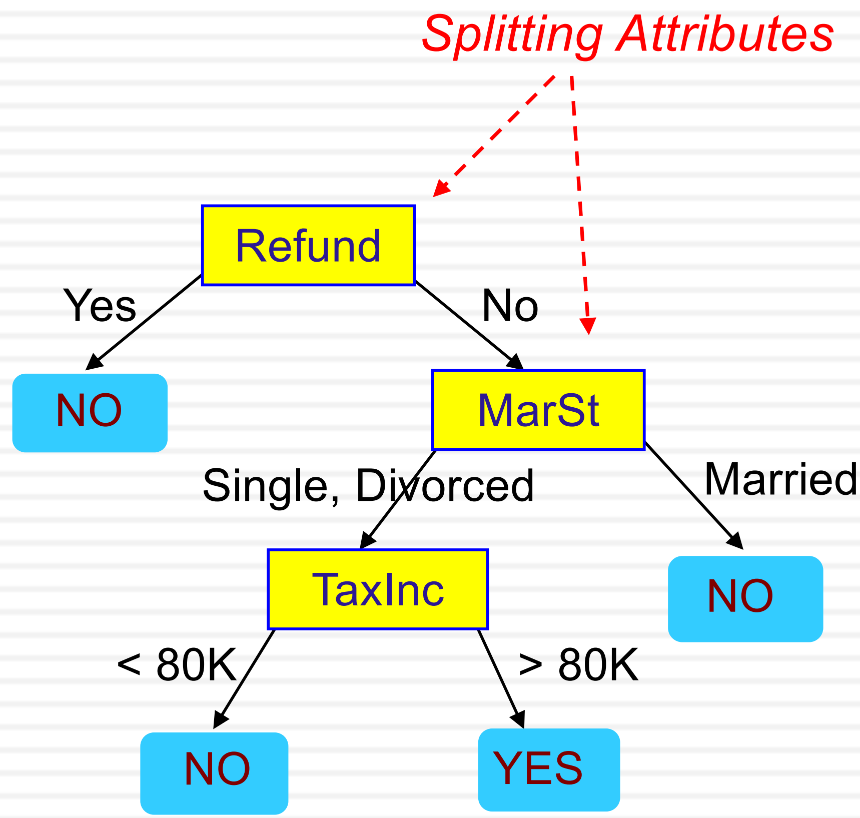
K = hyper parameter

Decision Tree

categorical categorical continuous class

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Training Data



Model: Decision Tree

Logistic Regression

- Categorical predictors
  - Predicting cancer from smoking and eating tomatoes.
  - We don't know what happens when non-smokers eat tomatoes because we have no data
- Probability of cancer?

Do you smoke?	Do you eat tomatoes?	Do you have cancer?
Yes	No	Yes
Yes	Yes	Yes
No	No	Yes
No	Yes	??????

Discrete Choice

- Observations
  - Purchase made
  - AND available options
- Assumption
  - Customers maximize their own utility

Choice made = the choice that maximizes utility

u(selected) = max u(options)

  - Utility = linear combination of attributes

Discrete Choice

- Travel
  - Online purchasing
  - Utility of flight = coeff * departure time + coeff * elapsed time + coeff * price + coeff * loyalty
  - Coefficients fitted based on historical observations
- Market basket
  - Available goods in the store
  - Utility of purchase = coeff * price + coeff * brand loyalty + coeff * home inventory

Discrete Choice vs Logistic Regression

- Observe actions
  - Purchase, no-purchase
  - Existing/left customers
  - Choices not available
- Observe choices
  - Still capture actions
  - Record all available options
  - Customer select maximum utility option
- Probability of selection

Logistic regression

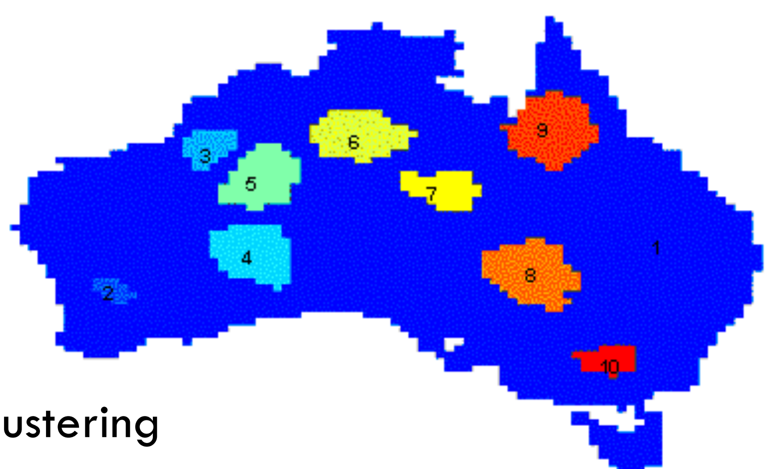
Discrete choice

Applications of Cluster Analysis

- **Understanding**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

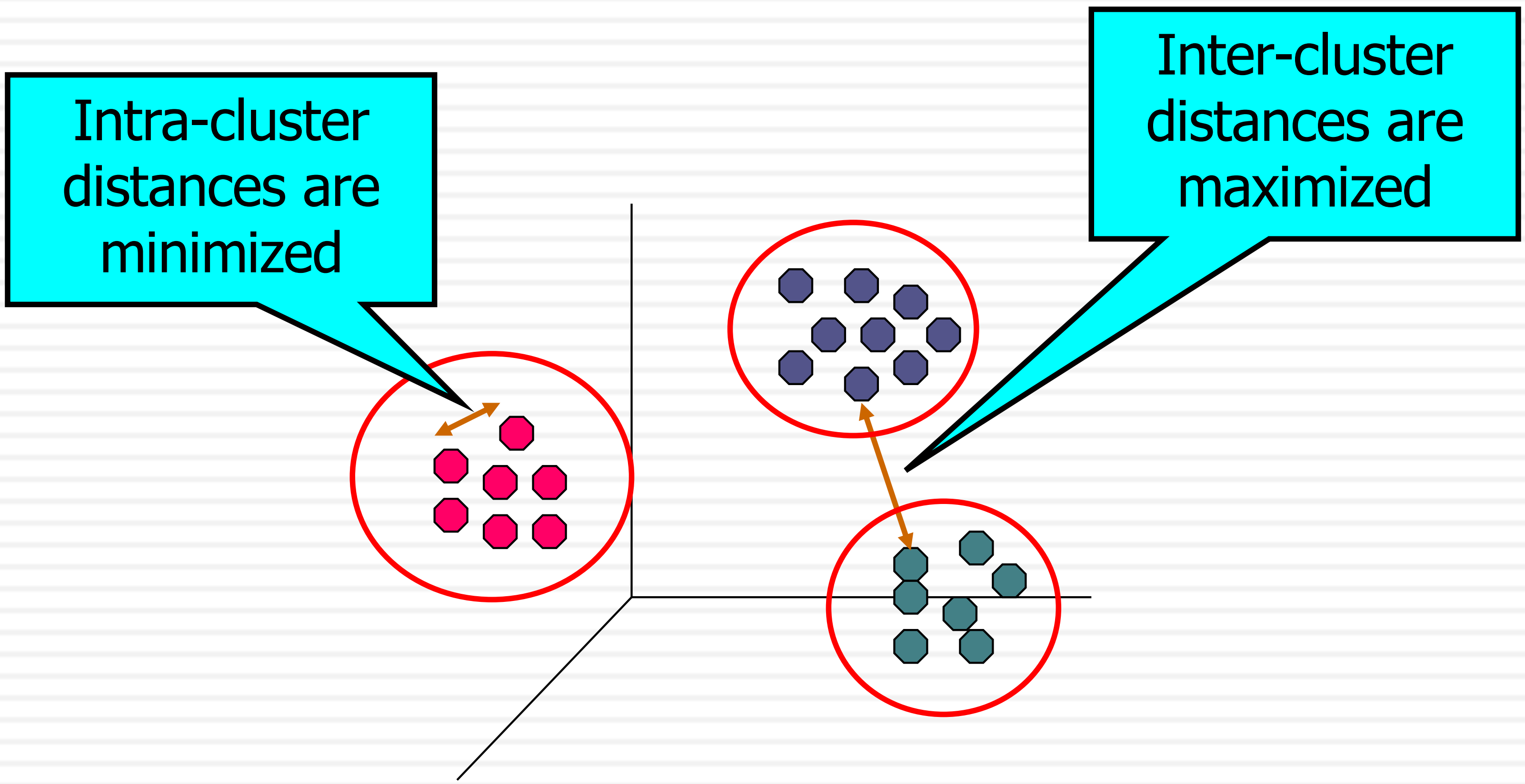
- **Summarization**
  - Reduce the size of large data sets



Clustering precipitation

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Example of Association Rules

{Diaper} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs,Coke},
{Beer, Bread} $\rightarrow$ {Milk},

Implication means co-occurrence, not causality!

Other Examples

- Market basket
  - If A, then bought B
  - How to use in promotions?
- Click sequence on the web
  - If visited X, then visited Y
  - If visited Z, then checkout
- Amazon.com
  - Recommendations

Machine Learning

Supervised learning	Unsupervised learning
Regression	Data clustering
Classification	Principal component analysis
Logistic regression	Independent component analysis
Random forests	Association rules
Neural networks	

Collaborative Filtering

- Personalized preferences
  - Amazon.com
- Market basket
  - Shopping carts
    - No order of purchases
  - Association rules
- Web click sequences
  - Order of selection
  - Aggregation trees

Collaborative Filtering

- Ratings are available
  - What people buy together
  - How they value goods
- User-user
  - Score/difference between two users
    - Recommend the choices of the ‘closest’ friend
  - Not many common reviews
  - Normalization
  - User rates the same item differently

Collaborative Filtering

- Item-item
  - Distance between two items
    - Based on evaluations of users of both items
  - Challenge
    - Two users like books
    - Odds of two books are low
- Resolution
  - Split to ‘categories’
  - Category-category and item-item

Survival Analysis

- Breakdowns
  - Machines
  - Computers
- What is the probability
  - Breakdown a months from now

    Historical observations

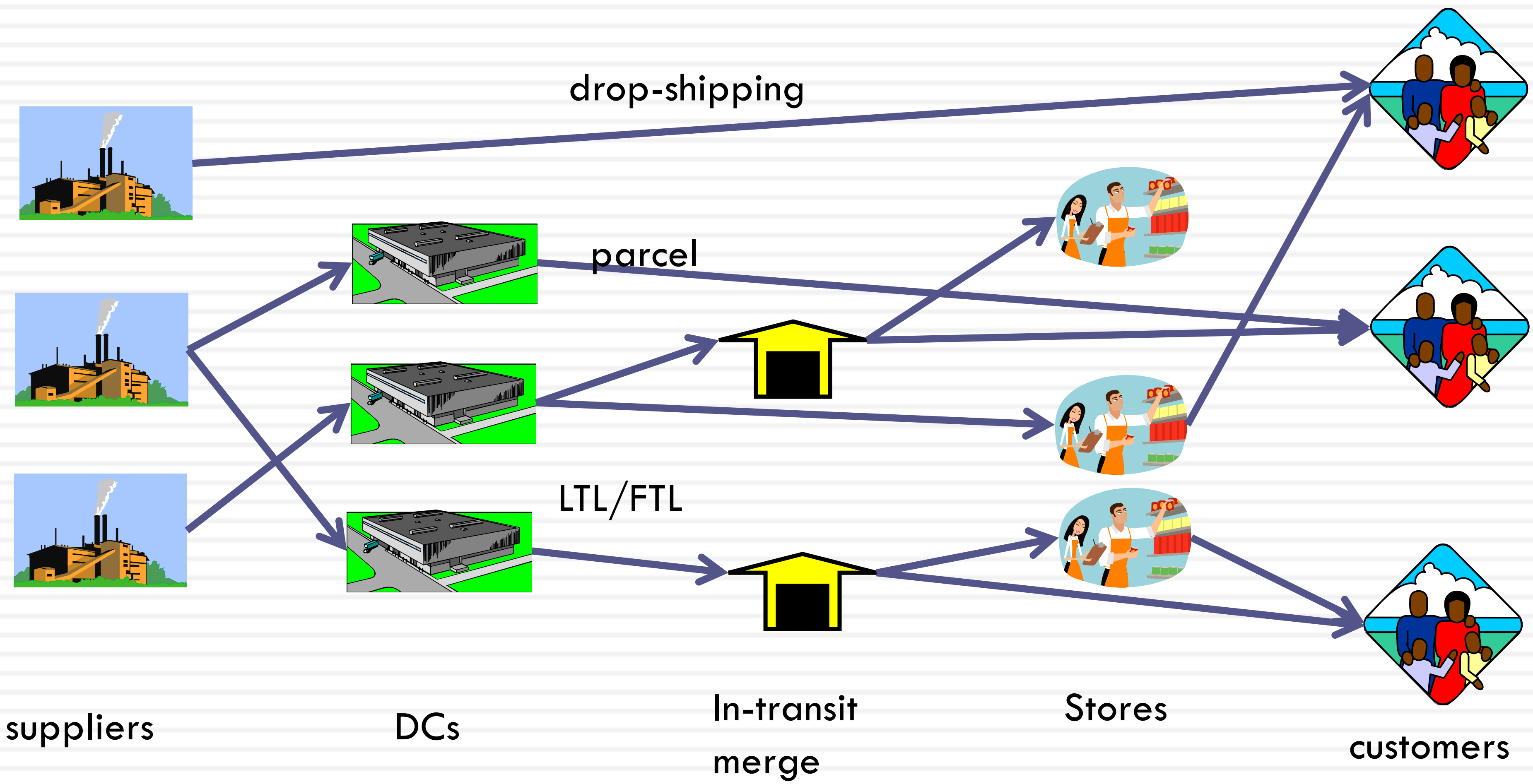
    Breakdown 1 month ago

    Breakdown 3 months ago

    Distribution fitting

    Probability[breakdown >= 1 month]

SC Network Design



By Industry

Marketing

- Churn analysis
  - Attrition rate
  - Probability that a customer will default
  - Logistic regression
- Customer segmentation
  - By demographics
  - By spend
  - Clustering

Propensity

- What is the likelihood the customer will next buy item X
  - Historical purchases
  - Logistic regression
- Yield of campaigns
- Budget allocation

Web Analytics

- Study paths on site made by customers
  - Redesign flow on sites
  - Why customers do not buy
  - Aggregation trees
  - Markov chains
- Many steps to get to paths
  - Associating clicks to ‘users’
- Digital ads placement
  - Where to place a banner ad
  - Which one
  - Optimization

Text Analytics

- Sentiment analysis
  - Tweet, blog, forum
  - Does it say something positive about a product
- List of good words
- List of bad words
- Count
  - If count of good – count of bad <= number
    - Positive sentiment
- Many challenges

Text Analytics

- Presidential campaign
  - How many positive tweets about President Obama
- Stock predictions
  - How many tweets make positive comments about the IBM stock
- Similar documents
  - Group similar documents
  - Manually inspect the documents in one group
  - Lawyers

Social Networking Analysis

- Six (or less) degrees of separation
  - Facebook
- Recommend friends
- Placement of ads
- Improved churn analysis
  - Take friendship into account
- Challenge
  - Sheer size of data

Other industries

- Healthcare
  - Too many applications
- Telecommunication
  - Churn
  - Product bundling
  - Network design
- Finance
  - Fraud detection
  - Churn

Other industries

- Sport analytics
  - Performance measures
- Service sector
  - Sales force optimization
  - Match employees to projects
- Transportation
  - Too many
  - Airlines one of the earliest users of analytics