Data Science and Technology Entrepreneurship

Linear Classifiers, Naive Bayes Classifier, Startup Technology and Scaling Issues

Sameer Maskey

Week3

Announcements

- TA Office hours
  - Thursdays 1pm to 3pm
  - Get help if needed! Need to setup webserver for next week
- Classroom
  - 313 FayerWeather (behind Avery)
  - Classes will be in 313 FayerWeather unless announced

Topics for Today

- Linear Classifiers
- Guest Lecture:
  - Startup Technology
  - Scaling Issues

Guest Lecture

- Hrishi Dixit
- CTO, LearnVest

Assignment II - Due Monday

- Fill up Lean Canvas
  - https://drive.google.com/a/parakhi.com/previewtemplate?id=16uOd158UzJM9oqGWgJOtbppzGNPmZ4fWMSV6_xBz3Z8&mode=public&pli=1#
- Field Assignment
  - Prepare minimum of 8 questions
  - Talk to at least 5 potential customers
    - 2 can be your friends
    - 3 has to be strangers
- Due Next Monday 18th @ 6pm

Extra Classes : Web Programming 101

- Starts next week
- Fridays @3:30 pm

Team Name

1. BuzztheBar
2. GymLogger
3. Intellidata
4. Kammunity
5. Pitch Perfect
6. PerFit
7. PsychSymptoms
8. SourceBase
9. Sochna
10. Soldthru
11. TertiaryMarket

Initial Pitch Day - Next Week

- Judge Panel
  - Amol Sarva
    - Co-founder, Peek, Virgin Mobile USA
  - David Lerner
    - Angel Investor, Entrepreneur in Residence
  - Ben Sisovick
    - General Partner, IA Ventures
  - Paul Tumpowsky
    - Chairman, inSITE

Preparing for Initial Pitch Day

- 2.5 hrs
- 12 Teams
- 12 min each
- 6 min presentation
- 6 min in feedbacks and QA

Preparing for Initial Pitch Day

- Presentation Guidelines
  - MadLib Template 1 line pitch
  - Discuss components of Lean Model Canvas
  - Focus on these topics
    - Customer Segments, Problem, Value Proposition of your solution
    - Know your competitive advantage, channels, market and revenue model
  - Customer Validation
    - Were your assumptions valid?
    - Analysis of customer interviews
      - What did you find out from questions you asked and data you collected
- Mockup
  - User Interface mockup
  - Describe user interaction with your mock up
- Website
  - Have a running website in AWS or other webservers
  - Add your mock up and any relevant information in your website

Website for Your Startup

- You need to have a running webserver
- Webserver can be password protected
  - You need to tell us how to access the site
- Mockups should be uploaded to the website
- Website can also be made information with other relevant information on it
- You can get a domain if you want

- Get help from Morgan tomorrow if you don't know how to do this!

Setting Up AWS Webserver

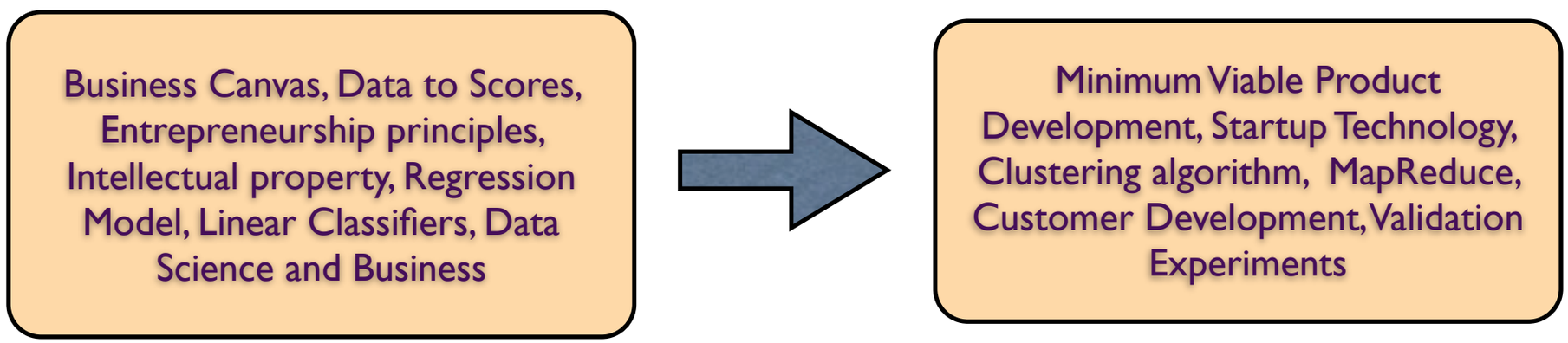
- We will help you with setting up AWS webserver
- Look for DSTE image in public image search in AWS
- DSTE image comes with
  - Apache
  - PHP
  - MySQL
  - few other basic software pre-installed

Course Stages

Stage 1 (3 weeks – Jan 30 – March Feb 20) Problem definition, Data collection, Customer development, Business Model Canvas, Minimum Viable Product development

Stage 2 (4 weeks – Feb 13 – March 10) Minimum Viable Product development, Quantifying customer feedback with classification and clustering techniques

(2 lectures after pitch event will be more Machine Learning with focus on classification and clustering algorithms)



Machine Learning and Business

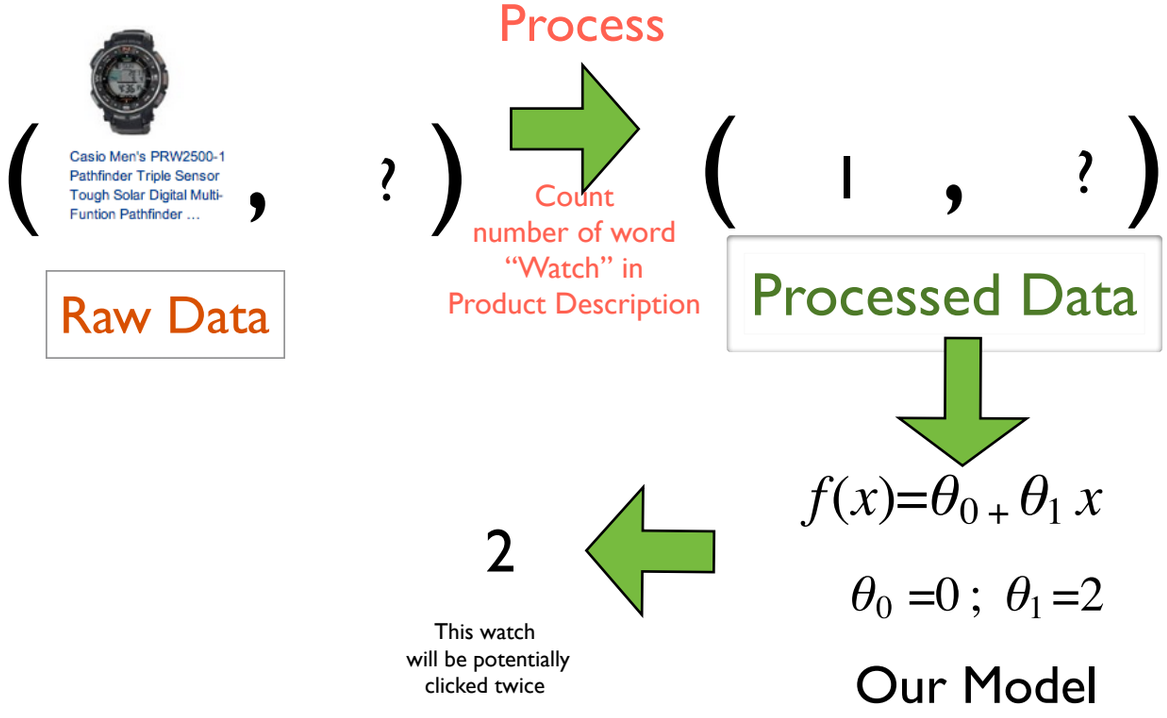
- Methods to analyze data that are all useful in decision making for businesses in general
- Data to Scores
- Data to Classes
  - Discriminative Methods
  - Generative Methods
- Data to Clusters

Machine Learning and Business

- Methods to analyze data that are all useful in decision making for businesses in general

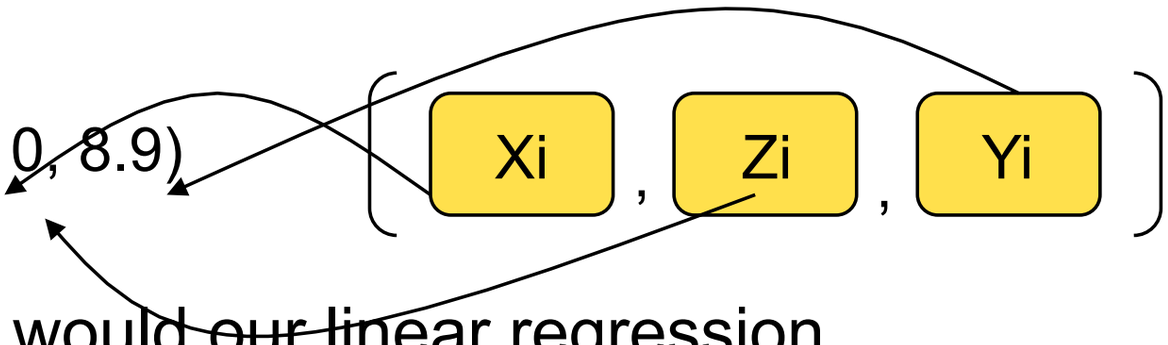
- Data to Scores
- Data to Classes
  - Discriminative Method
  - Generative Methods
- Data to Clusters

Data to Predicted Scores



More Features in Regression Model

- Adding one more feature Z_i
  - (1, 3, 4)
  - (0, 6,1.8)
  - .
  - .
  - .
  - (2, 0, 8.9)



- What would our linear regression function would look like

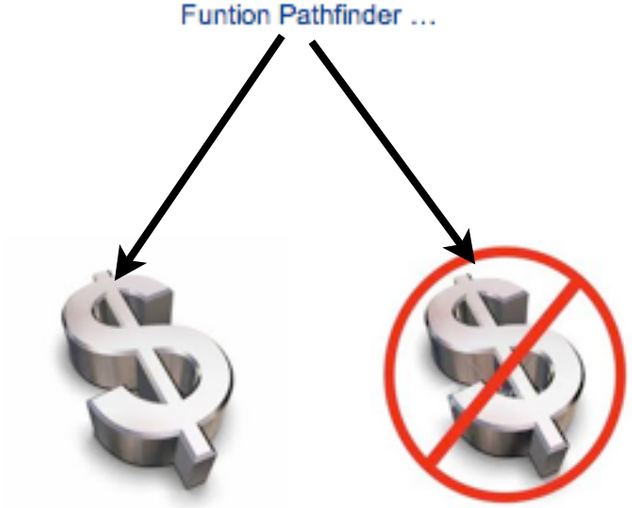
Machine Learning and Business

- Methods to analyze data that are all useful in decision making for businesses in general
- Data to Scores
- Data to Classes
  - Discriminative Methods
  - Generative Methods
- Data to Clusters

Machine Learning and Business

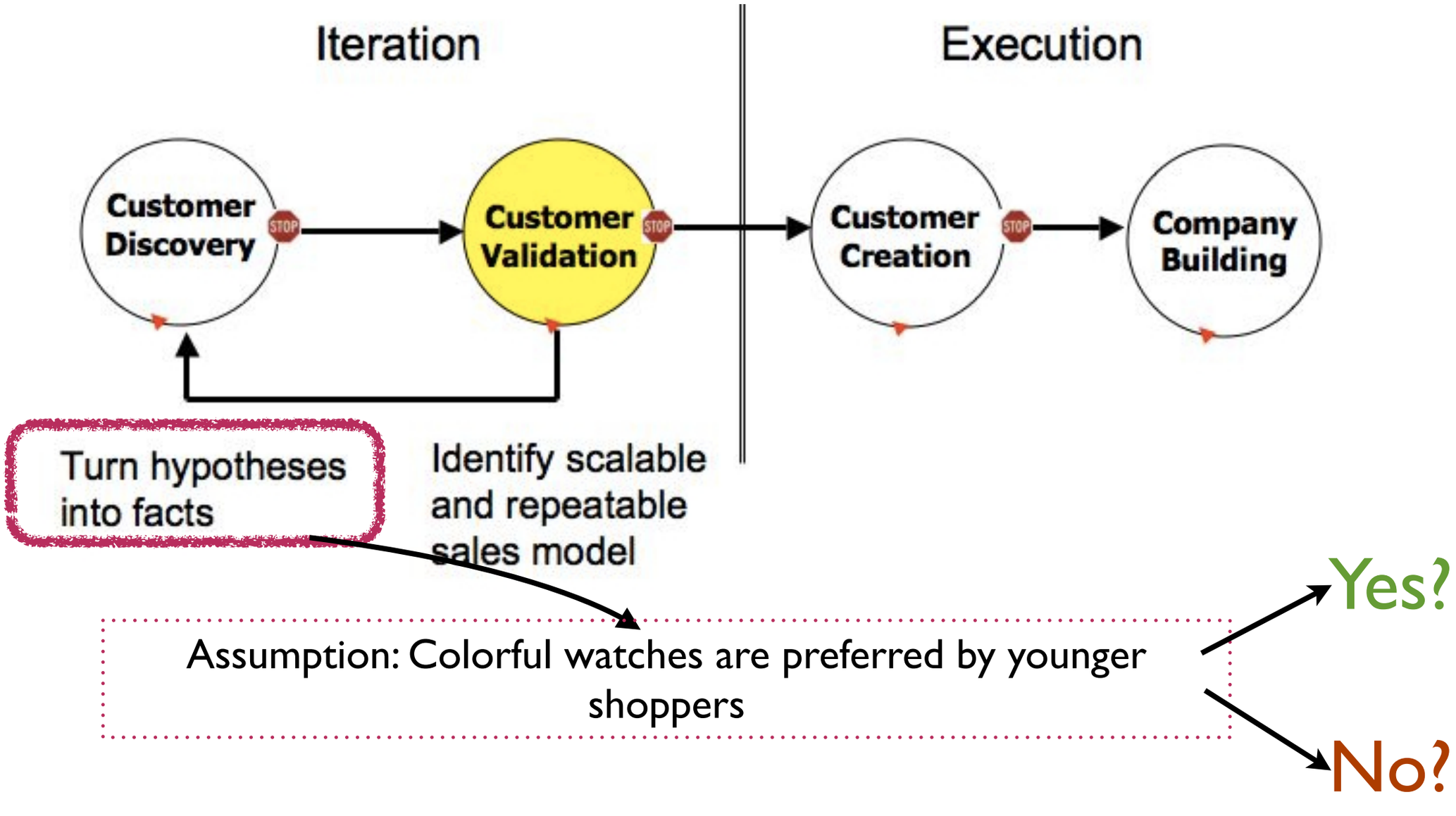
- Methods to analyze data that are all useful in decision making for businesses in general
- Data to Scores
- Data to Classes
  - Discriminative Methods
  - Generative Methods
- Data to Clusters

Casio Men's PRW2500-1 Pathfinder Triple Sensor Tough Solar Digital Multi-Funtion Pathfinder …



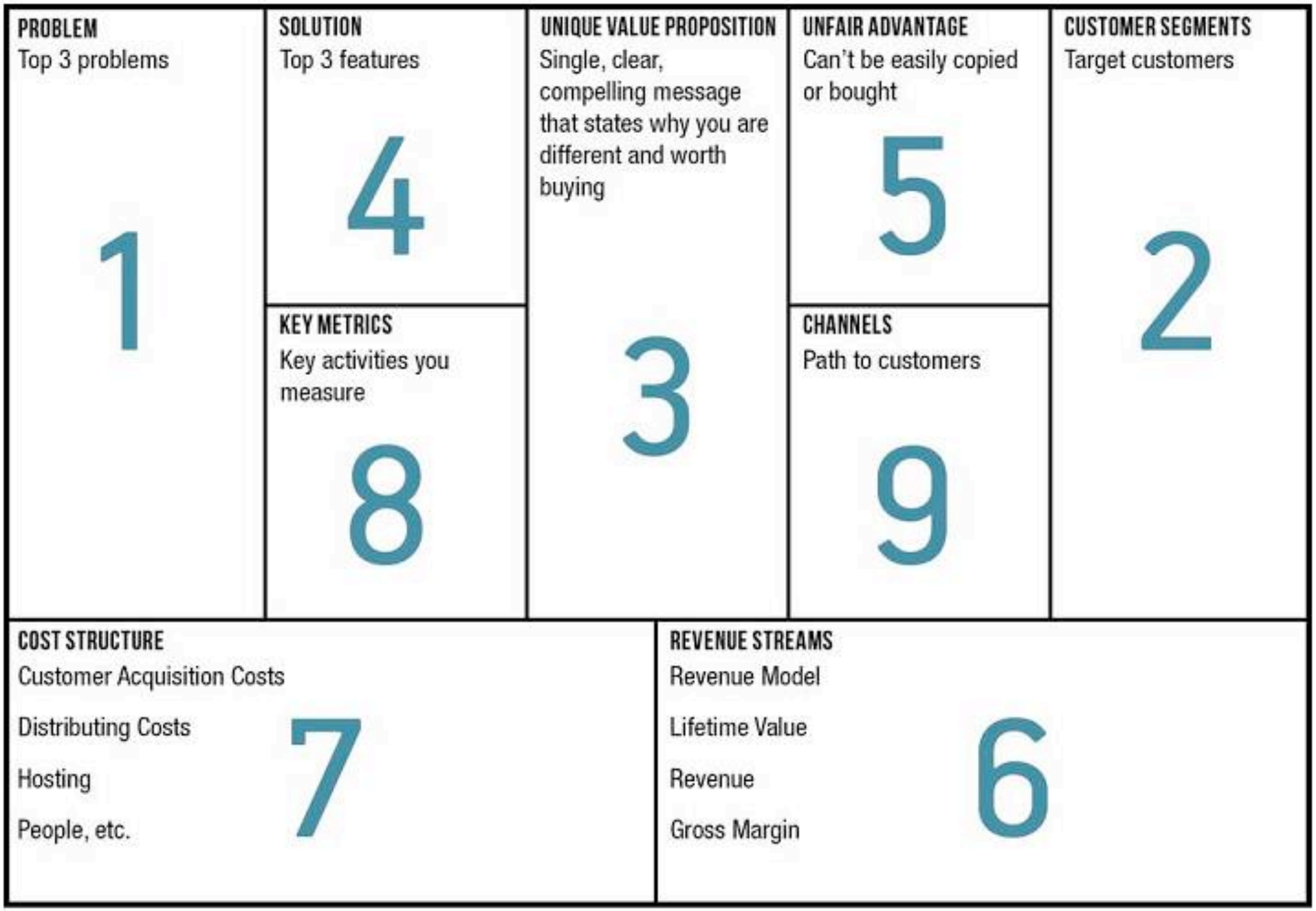
Customer Discovery Process

- Source : Startup Owner's Manual - Steve Blank and Bob Dorf



Lean Canvas [Maurya, A]

https://docs.google.com/drawings/d/1RCczinVGbE1FJ0geyOwpGWWm5FYkvmLSXnRenf9dY_o/edit



Business Model Assumptions

- The company Zoolaster sells Zoola watches online
- Zoolaster buys watches from wholesaler for cheaper price and sells them online
- Zoolaster assumes certain types of Zoola watches sell well
- Zoolaster executives want to quantify which Zoola watches may sell well so that they just buy those kind from the wholesaler

Sales Data

Product Buy?

( **Stauer Compendium Hybrid Watch**
$29.00
+ No Shipping Info
STAUER , 1)

( **Michael Kors Watches Bradshaw Gold - Michael**
$250.00
+ Free Shipping
eBags. , 1)

( **Men's Pro Diver Chronograph Black**
$199.99
+ Free Shipping
World of Watches , 0)

1 = Bought
0 = Didn't Buy

Sales Prediction Model

Product Buy?

( , 1)

Stauer Compendium Hybrid Watch
$29.00
+ No Shipping Info
STAUER

( , 1)

Michael Kors Watches Bradshaw Gold - Michael
$250.00
+ Free Shipping
eBags.

( , 0)

Men's Pro Diver Chronograph Black
$199.99
+ Free Shipping
World of Watches

Customer Will Buy?

Zoolaster can potentially buy more watches from wholesaler that have higher potential of selling online

Data to Classification

- Given a set of features

$$X=(x_1, x_2, x_3, \ldots, x_n)$$

- we want to predict Y

How about x?

How do we get them?

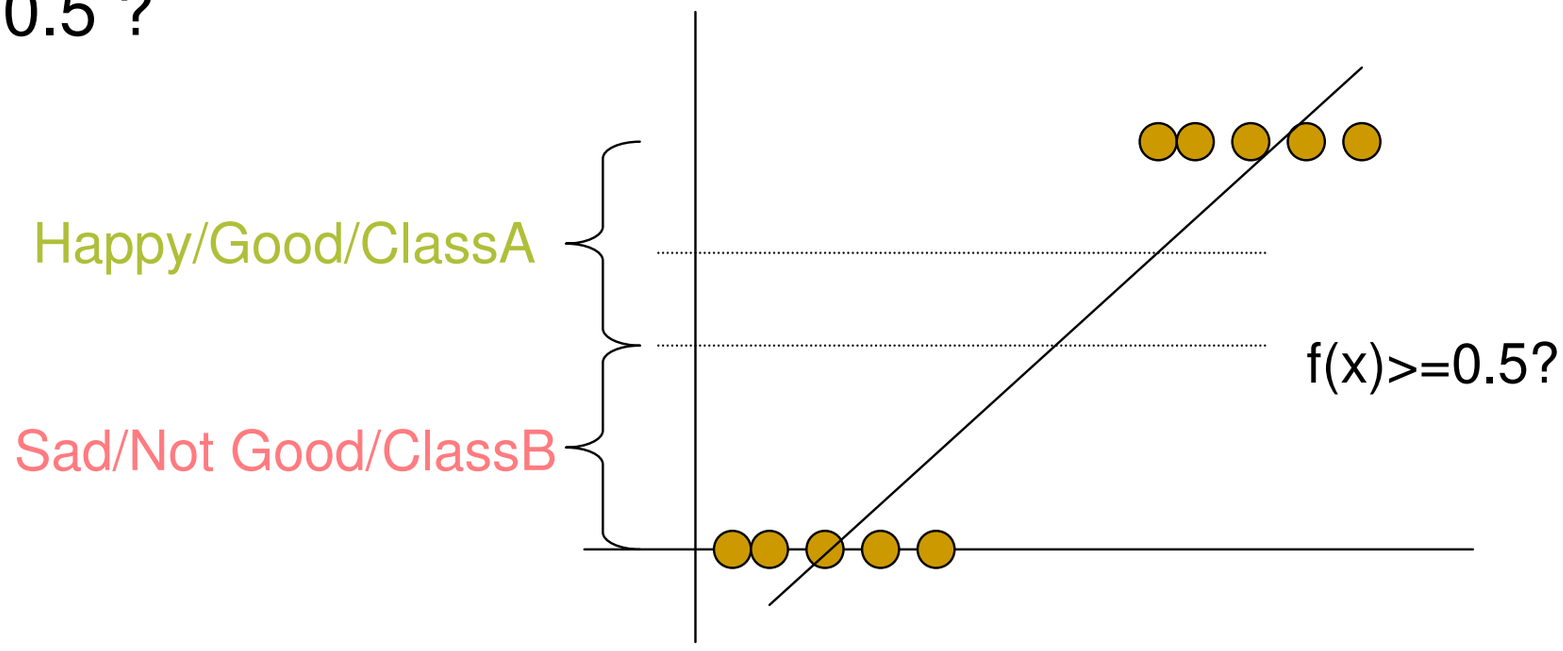
$$Y=\{0,1\}$$

(
Stauer Compendium Hybrid Watch
$29.00
+ No Shipping Info
STAUER

, { will not buy OR will buy })

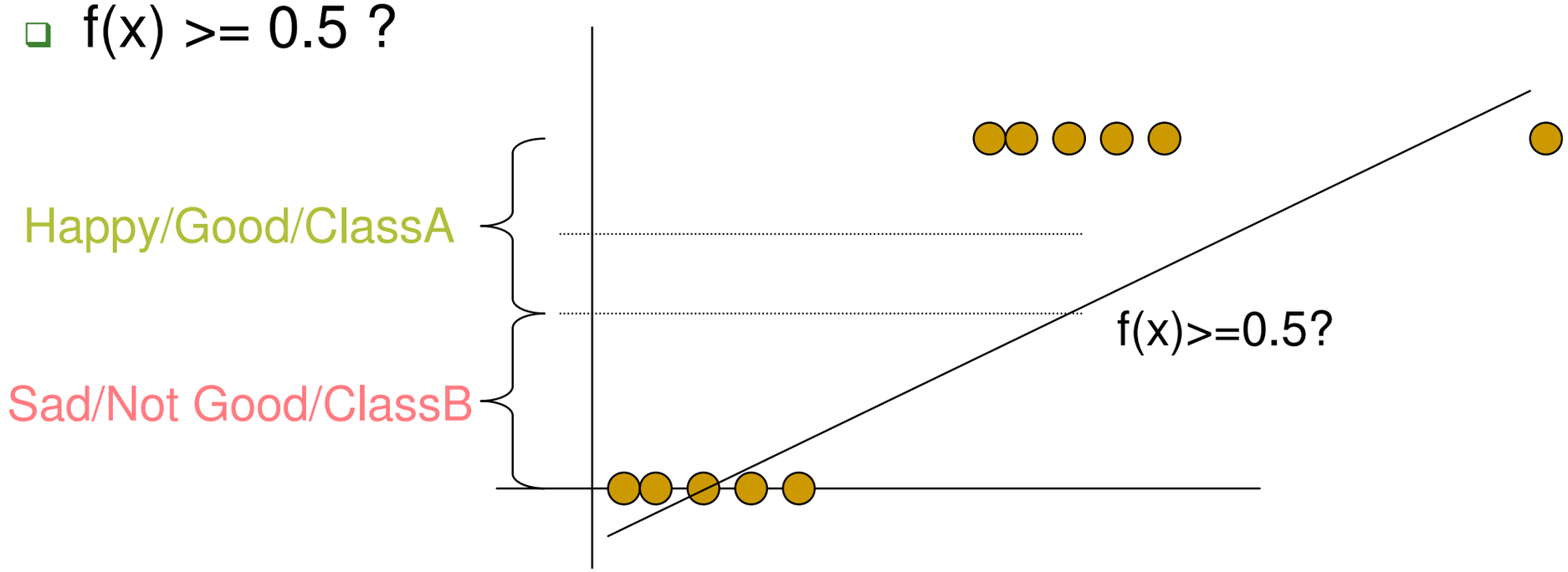
Data to Classification

- Can we build a regression model to model such binary classes?
- Train Regression and threshold the output
  - If f(x) >= 0.7 CLASS1
  - If f(x) < 0.7 CLASS2
  - f(x) >= 0.5 ?



Regression to Classification

- Can we build a regression model to model such binary classes?
- Train Regression and threshold the output
  - If f(x) >= 0.7 CLASS1
  - If f(x) < 0.7 CLASS2
  - f(x) >= 0.5 ?

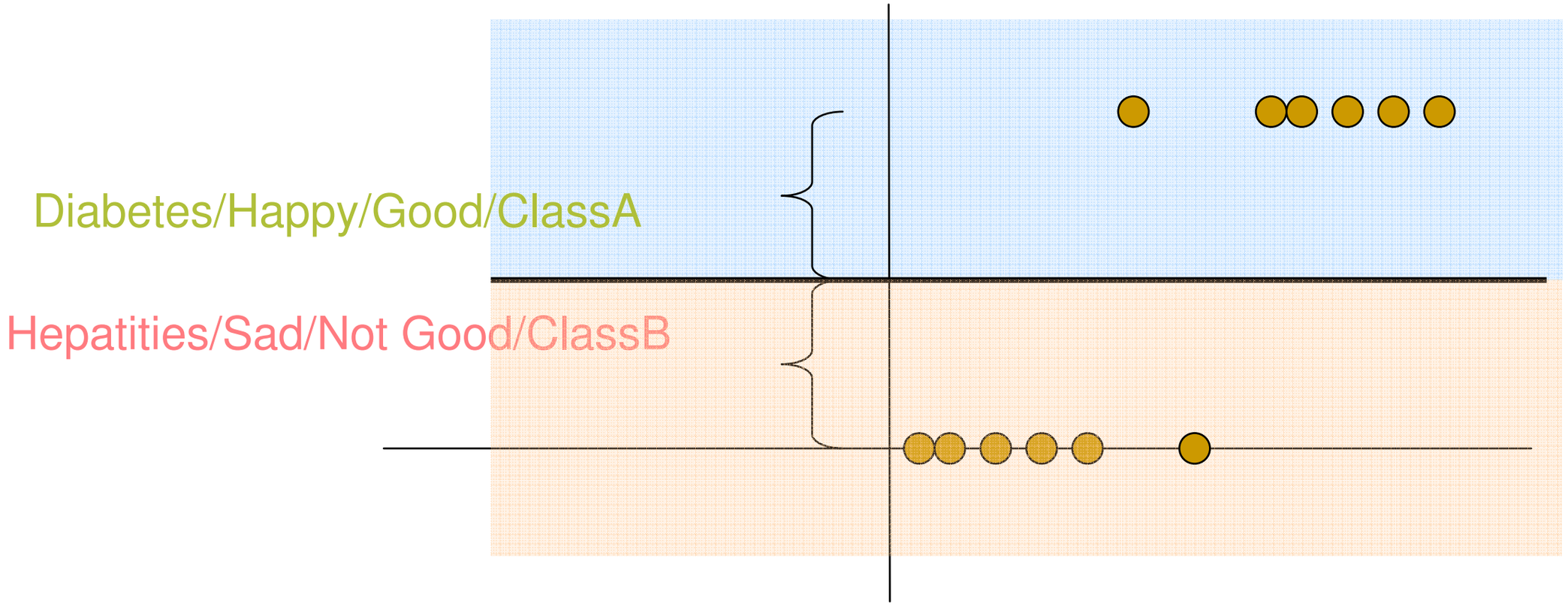


Regression to Classification

- Thresholding on regression function does not always work
- Gaussian assumption on noise
- When the output is binary class, we may want to dry a different technique of modeling than regression
- Many modeling techniques that will better produce class category values we want for Y
- Using Linear Classifiers is one such method

Half Plane and Half Spaces

- Half plane is a region on one side of an infinite long line, and does not contain any points from other side
- Half space n-dimensional space obtained by removing points on one side of hyperplane (n-1 dimension)
  - What would it look like for a 3 dimensional space



Decision Surface

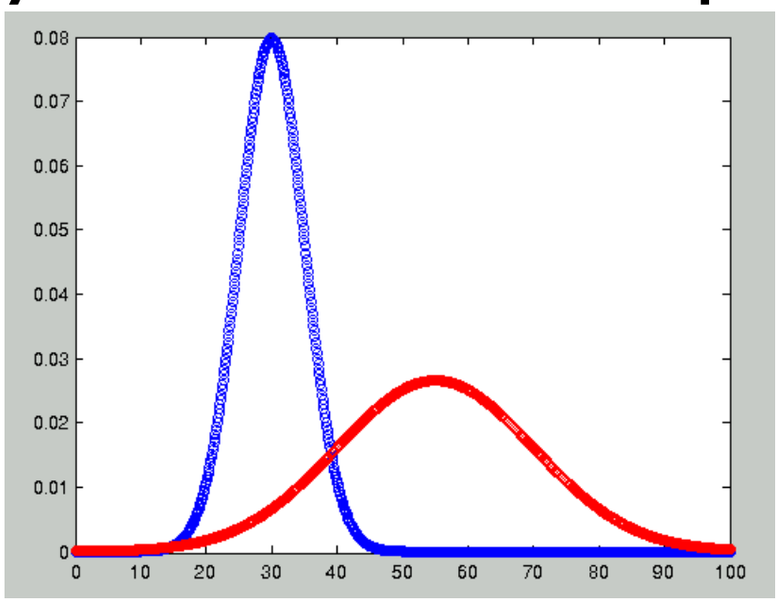
- We want to find a decision surface that will classify our data better
- Fisher's Linear Discriminant
  - Dimensionality reduction, project data on a line and classify
- Naive Bayes
  - Compute p(y|x) using conditional independence assumption
- Perceptron
  - Linear Discrimination with a hyperplane in (d-1) dimension

Generative vs. Discriminative Classifier

- Generative Classifier
  - Model joint probability p(x,y) where x are inputs and y are labels
  - Make prediction using Bayes rule to compute p(y|x)
- Discriminative Classifier
  - Try to predict output directly
  - Model p(y|x) directly

Generative Classifier

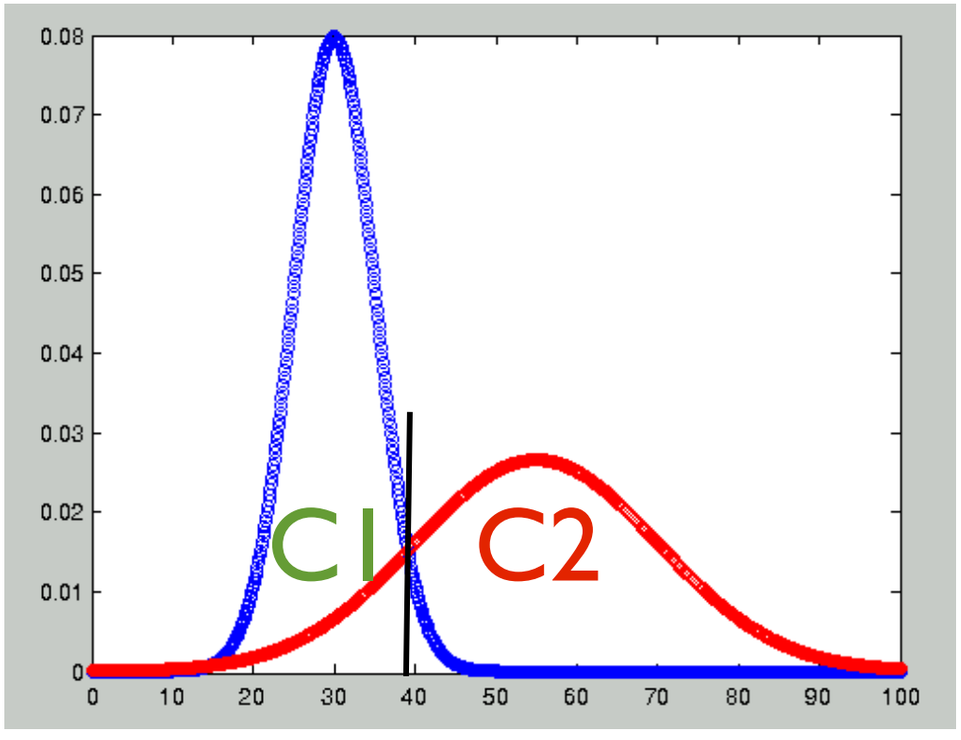
- We can model class conditional densities using Gaussian distributions
- If we know class conditional densities
  - p(x| y=C1)
  - p(x|y=C2)
- We can find a decision to classify the unseen example



Bayes Rule

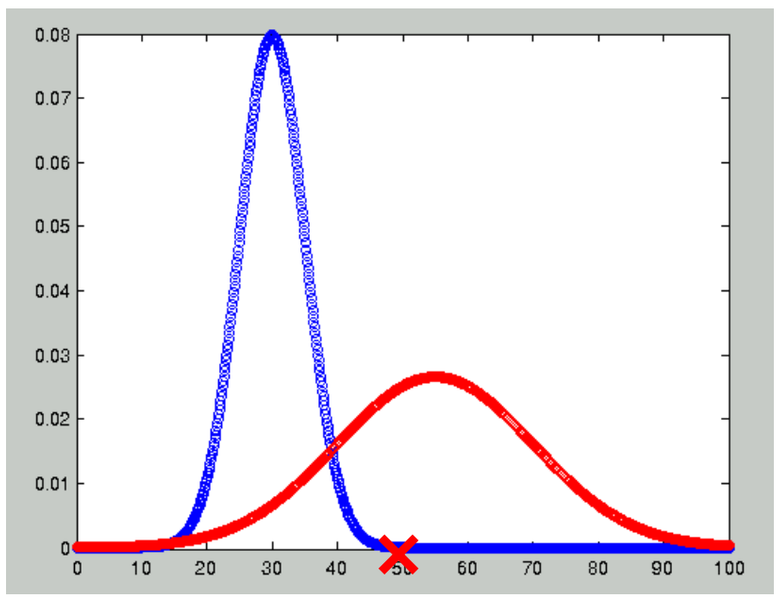
$$P(Y|X) = \frac{P(X|Y)\ P(Y)}{P(X)}$$

C1 = Buys
C2 = Doesn't Buy



Generative Classifier

- Given a new data point find out posterior probability from each class and take a log ratio
- If higher posterior probability for C1, it means new x better explained by the Gaussian distribution of C1



$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$p(y = 1|x) \propto p(x|\mu_1, \sum_1)p(y = 1)$$

Naive Bayes Classifier

- Naïve Bayes Classifier a type of Generative classifier
- Compute class-conditional distribution but with conditional independence assumption
- Shown to be very useful for many classification tasks

Naive Bayes Classifier

- Conditional Independence Assumption

$$P(X_1, X_2, ..., X_N|Y) = \Pi_{i=1}^{N} P(X_i|Y)$$

Naive Bayes Classifier

$$P(Y_k, X_1, X_2, ..., X_N) = P(Y_k)\Pi_i P(X_i|Y_k)$$

Prior Probability of the Class

Conditional Probability of feature given the Class

Naive Bayes Classifier

$$P(Y = y_k|X_1, X_2, ..., X_N) = \frac{P(Y{=}y_k)P(X_1,X_2,..,X_N|Y{=}y_k)}{\sum_j P(Y{=}y_j)P(X_1,X_2,..,X_N|Y{=}y_j)}$$

$$= \frac{P(Y{=}y_k)\Pi_i P(X_i|Y{=}y_k)}{\sum_j P(Y{=}y_j)\Pi_i P(X_i|Y{=}y_j)}$$

$$Y \leftarrow argmax_{y_k} P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$