Data Science and Technology Entrepreneurship

Data Science for Your Startup
Classification Algorithms
Minimum Viable Product Development

Sameer Maskey
Week6
Announcements

‣ No class for next 2 weeks
  ‣ March 11 week - NO Class - MBA students not on campus
  ‣ March 18 week - NO Class - Spring break

‣ Extra Lectures
  ‣ This Friday’s lecture is cancelled
Topics for Today

- Big Data
- Data Science for your Startup
- Linear Classifiers
  - Naive Bayes
  - Perceptron
- Minimum Viable Product Development
Feedback

http://www.surveymonkey.com/s/BFQJY79
$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion
potential annual value to Europe’s public sector administration—more than GDP of Greece

$600 billion
potential annual consumer surplus from using personal location data globally

60%
potential increase in retailers’ operating margins possible with big data
Big Data in Various Fields

- Healthcare
- Government
- Ecommerce
- Marketing
- Manufacturing
- Retail
Value for Different Fields

Big data can generate significant financial value across sectors

**US health care**
- $300 billion value per year
- ~0.7 percent annual productivity growth

**Europe public sector administration**
- €250 billion value per year
- ~0.5 percent annual productivity growth

**Global personal location data**
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

**US retail**
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

**Manufacturing**
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

Source - McKinsey Report
Some sectors are positioned for greater gains from the use of big data

Historical productivity growth in the United States, 2000–08

1 See appendix for detailed definitions and metrics used for value potential index.


Source - McKinsey Report
A heat map shows the relative ease of capturing the value potential across sectors.

Source - McKinsey Report

1 See appendix for detailed definitions and metrics used for each of the criteria.
Visualization
Republicans Vs. Democrats

‣ Can we predict which congressman is republican or democrat?

‣ Can we predict what is the likelihood that a congressman will vote yes in the upcoming vote?
Data

1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)
5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)
Predict who is Republican or Democrat?

2. handicapped-infants: 2 (y, n)
3. water-project-cost-sharing: 2 (y, n)
4. adoption-of-the-budget-resolution: 2 (y, n)
5. physician-fee-freeze: 2 (y, n)
6. el-salvador-aid: 2 (y, n)
7. religious-groups-in-schools: 2 (y, n)
8. anti-satellite-test-ban: 2 (y, n)
9. aid-to-nicaraguan-contras: 2 (y, n)
10. mx-missile: 2 (y, n)
11. immigration: 2 (y, n)
12. synfuels-corporation-cutback: 2 (y, n)
13. education-spending: 2 (y, n)
14. superfund-right-to-sue: 2 (y, n)
15. crime: 2 (y, n)
16. duty-free-exports: 2 (y, n)
17. export-administration-act-south-africa: 2 (y, n)
Data

'n', 'y', 'n', 'y', 'y', 'y', 'n', 'n', 'n', 'y', '?', 'y', 'y', 'y', 'n', 'y', 'republican'
'n', 'y', 'n', 'y', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'y', 'y', 'y', 'n', '?', 'republican'
', 'y', 'y', '?', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'y', 'y', 'y', 'y', 'n', 'n', 'democrat'
'y', 'y', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'y', '?', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'democrat'
'y', 'y', 'y', 'y', 'n', 'n', 'n', 'y', '?', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'democrat'
'y', 'y', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'n', 'n', 'y', '?', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'democrat'
'n', 'y', 'n', 'y', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'n', 'n', '?', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'y', 'republican'
'n', 'y', 'n', 'y', 'y', 'y', 'n', 'n', 'n', 'n', 'n', 'n', 'n', 'n', 'y', 'y', 'y', 'y', 'n', 'y', 'republican'
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

\[ P(Y|X) = \frac{P(X|Y) \, P(Y)}{P(X)} \]

\( C_1 = \text{Buys} \)
\( C_2 = \text{Doesn't Buy} \)
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.

\[
p(y | x) \propto \frac{p(x | y)p(y)}{p(x)}
\]

\[
p(y = 1 | x) \propto p(x | \mu_1, \Sigma_1)p(y = 1)
\]
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

\[
P(X_1, X_2, \ldots, X_N | Y) = \prod_{i=1}^{N} P(X_i | Y)
\]
Naive Bayes Classifier

\[ P(Y_k, X_1, X_2, \ldots, X_N) = P(Y_k) \prod_i P(X_i | Y_k) \]

- Prior Probability of the Class
- Conditional Probability of feature given the Class
Naive Bayes Classifier

\[
P(Y = y_k|X_1, X_2, \ldots, X_N) = \frac{P(Y = y_k)P(X_1, X_2, \ldots, X_N|Y = y_k)}{\sum_j P(Y = y_j)P(X_1, X_2, \ldots, X_N|Y = y_j)} \\
= \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}
\]

\[Y \leftarrow \text{argmax}_{y_k} P(Y = y_k)\prod_i P(X_i|Y = y_k)\]
Naive Bayes Classifier for Text

Given the training data what are the parameters to be estimated?

\[ P(Y) \quad P(X|Y_1) \quad P(X|Y_2) \]

- Diabetes : 0.8
- Hepatitis : 0.2

- the: 0.001
diabetic : 0.02
blood : 0.0015
sugar : 0.02
weight : 0.018
...

- the: 0.001
diabetic : 0.0001
water : 0.0118
fever : 0.01
weight : 0.008
...
Implementing Naive Bayes

\[ P(X|Y_1) = \Pi_i P(X = x_i|Y = y_1) \]

\[ \theta_{i,j,k} \equiv P(X_i = x_{ij}|Y = y_k) \]

MLE Estimation of the parameters

\[ \hat{\theta}_{i,j,k} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}} \]

\#D\{x\} = number of elements in the set D that has property x
Dimensionality reduction is one way of classification

We can also try to find they discriminating hyperplane by reducing the total error in training

- Perceptrons is one such algorithm
We want to find a function that would produce least training error

\[ R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(y_i, f(x_i; w)) \]
Given training data \( < (x_i, y_i) > \)

We want to find \( w \) such that

\[
(w \cdot x_i) > 0 \text{ if } y_i = -1 \text{ is misclassified} \\
(w \cdot x_i) < 0 \text{ if } y_i = 1 \text{ is misclassified}
\]

- We can iterate over all points and adjust the parameters

\[
w \leftarrow w + y_i x_i \\
\text{if } y_i \neq f(x_i; w)
\]

- Parameters are updated only if the classifier makes a mistake
We are given \((x_i, y_i)\)

Initialize \(w\)

Do until converged

\[
\text{if } \text{error}(y_i, \text{sign}(w.x_i)) == \text{TRUE} \\
\quad w \leftarrow w + y_i x_i
\]

end if

End do

If predicted class is wrong, subtract or add that point to weight vector
Training Perceptron

Another Version

\[ Y_j(t) = f[w(t) \cdot x_j] \]

\[ w_i(t + 1) = w_i(t) + \alpha (d_j - y_j(t)) x_{i,j} \]

Y is prediction based on weights and it's either 0 or 1 in this case.

Error is either 1, 0 or -1.

Example from Wikipedia
Weka

- Publicly available free software that includes many common ML algorithms that are used in Natural Language Processing
- GUI and Commandline Interface
- Feature Selection, ML algorithms, Data filtering, Visualization
Weka Download and Setup

- [http://sourceforge.net/projects/weka/files/weka-3-4/3.4.17/weka-3-4-17.zip/download](http://sourceforge.net/projects/weka/files/weka-3-4/3.4.17/weka-3-4-17.zip/download)

  ```
  >> unzip weka-3-4-17.zip
  >> java -jar weka-3-4-17/weka.jar
  >> Click on Explorer
  ```
Data needs to be in ARFF format

Prediction Class at the end of feature list

Weka

Filter Features

Visualize data
Building ML Models with Weka

Classifier Choice

Model Testing

Results
Model Evaluation with Weka

Tasks

Modal Load/Save

Visualize Model
10-fold Cross Validation

- 10 fold cross validation
  - Assuming we have 100K data points
    - Train on 90K (1 to 90,000)
    - Test on 10K (90,001 to 100,000)
  - But we can do this 10 times if we select different 10K of test data point each time

  Exp1 | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k

  Exp2 | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k

  ...  

  Exp10 | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k | 10k

- 10 experiments, build model and test times with 10 different sets of training and test data
- Average the accuracy across 10 experiments
- We can do any N-fold cross validation to test our model
## Interpreting Weka Results

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
Precision, Recall, F-Measure

Precision \( \frac{TP}{(TP+FP)} \)

Recall \( \frac{TP}{(TP+FN)} \)

F-Measure \( \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \)

Accuracy \( \frac{(TP+TN)}{(TP+TN+FP+FN)} \)
## Confusion Matrix

- Assume we are classifying text into two categories Hepatitis (H) and Others (B).
- Let’s assume we had 1000 documents such that 500 are H and 500 are B.
- Assume we got given predictions.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>H</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>400</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.6667</td>
</tr>
<tr>
<td>Recall</td>
<td>0.8000</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.7273</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.7000</td>
</tr>
</tbody>
</table>
Commandline for Weka

- Make sure CLASSPATH variable is setup; can also give the path explicitly using –cp parameter
  - `>> export CLASSPATH=$CLASSPATH:/home/smaskey/soft/weka-3-4-17/weka.jar`

- Try to see if java can access the classes for classifiers
  - `>> java weka.classifiers.bayes.NaiveBayes`

- Try to build a model from commandline
  - `>> java weka.classifiers.trees.J48 -i -t data/weather.arff`

- Try other examples from Weka wiki
  - `>> java weka.classifiers.bayes.NaiveBayes -K -t soybean-train.arff -T soybean-test.arff -p 0`
Data Science for Your Startup

PerFit
FlyJets
GymLogger
PsychSymptoms
NomadTravel
BuzztheBar
Pitch Perfect
Karmmunity
Sochna
Intellidata
SourceBase
SoldThru
Minimum Viable Product Development

- Build MVP with minimum number of feature sets that allows you to do test your customer

- All MVPs are not the same
  - Physical product MVP
  - Web Application can be tested faster

**Goal of MVP is to have a prototype that allows you to figure out if you understand the customer problem and if your product potentially solves it**
Customer Discovery with MVP

Phase 1: Set of Hypotheses about your business (Problem?, Solution? Value Proposition?)

Phase 2: Set of Hypotheses about your business (Test your hypotheses by talking to customers)

Phase 3: Build MVP and test MVP with customers (Does your MVP solves the problem customer want?)

Phase 4: Analyze results of your Phase 3 (Ready to signup paying customers?)
Multiple MVPs

- Multiple MVPs can be used to test competing hypotheses

Example:

- MVP with pay per use model
- MVP with pay per month model

If it is not difficult to build multiple MVPs then build them and test them with customers