SUPER BOWL XLIX

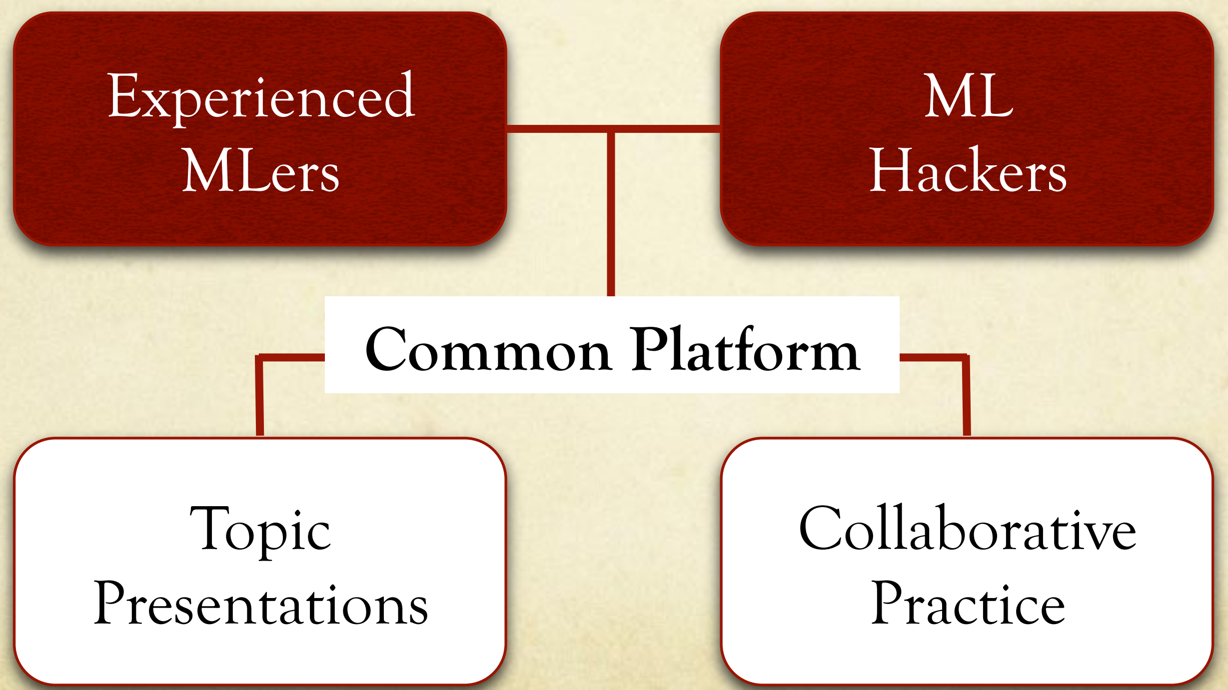
ARIZONA 02.01.15

Naïve Bayes for the Superbowl

John Liu
Nashville Machine Learning Meetup
January 27th, 2015

NashML Goals

Create a hub for like-minded people to come together, share knowledge and collaborate on interesting domains.



Platform

- IPython Notebook (Project Jupyter)
- Java, Scala, Python (others?)
- Scikit-learn
- PyLearn2/Theano
- iTorch
- AWS/Mahout/Spark/Mllib?

Rev. Thomas Bayes

"An Essay towards solving a Problem in the Doctrine of Chances" published posthumously 1763

I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper…

Bayes' Theorem

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Statistical Inference

Likelihood

Prior

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Posterior

(What we want to find)

Evidence

Intuition Behind Bayes

A Priori	Initial Belief (model)
Evidence	See the Data
Likelihood	How likely to see Data given Belief
A Posteriori	Updated Belief after seeing Data

$$posterior = \frac{prior \bullet likelihood}{evidence}$$

Example

Blackjack Insurance Bet:

What is the probability that a dealer with an Ace showing has Blackjack?

Example

Prior	P(Dealer has Blackjack)	32/663
Evidence	P(Ace showing)	1/13
Likelihood	P(Ace showing|has Blackjack)	1/2
Posterior	**P(Blackjack|Ace showing)**	**16/51**

16/51 = 31% or less than 1/3 of the time.

Are you a Bayesian?

You read Burton Malkiel's book "A Random Walk down Wall Street" and believe in the Efficient Market Hypothesis. Your broker gives you a tip to buy Tesla. You ignore the broker and Tesla rises 100 days in a row.

As a Bayesian, do you believe:

A) The stock is long due for a correction

B) It is possible for Tesla to rise another 100 days in a row

C) You were fooled by randomness

Naïve Bayes

You observe outcome Y with some n features X_1, X_2, ..X_n. The joint density can be expressed using the chain rule:

$$P(Y, X_1, X_2, ..X_n) = P(X_1, X_2, ..X_n | Y)\ P(Y)$$

$$= P(Y)\ P(X_1,Y)\ P(X_2 | Y,X_1)\ P(X_3 | Y,X_1,X_2)...$$

This is messy, but simplifies if we naively assume independence,

$$\left.\begin{aligned} P(X_2 | Y,X_1) &= P(X_2 | Y) \\ P(X_3 | Y,X_2,X_1) &= P(X_3 | Y) \\ P(X_n | Y,X_n...X_2,X_1) &= P(X_n | Y) \end{aligned}\right\} \textbf{Naïve Bayes Assumption}$$

Naïve Bayes Classifier

Let K classes be denoted c_k. The (conditional) probability of class c_k given that we observed features x_1, x_2,..x_n is:

$$\mathrm{P}(\mathrm{c_k} \mid \mathrm{x_1}, \mathrm{x_2}, ..\mathrm{x_n}) = \mathrm{P}(\mathrm{c_k}) \prod_{i=1}^{n} P(x_i \mid c_k)$$

A Naïve Bayes classifier simply chooses the class with highest probability (maximum a posteriori):

$$\mathrm{c}_{NB} = \underset{k \in K}{\mathrm{argmax}}\ \mathrm{P}(\mathrm{c_k}) \prod_{i=1}^{n} P(x_i \mid c_k)$$

Gaussian Naïve Bayes

When features x_i are continuous valued, typically make the assumption they are normally distributed:

$$P(x_i \mid c_k) = \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{ki}}{\sigma_{ki}}\right)^2}$$

$$c_{NB} = \underset{k \in K}{\operatorname{argmax}} \ \mathrm{P}(\mathrm{c}_\mathrm{k}) \prod_{i=1}^{n} P(x_i \mid c_k)$$

Variance σ_{ki} can be independent of x_i and/or c_k.

Multinomial Naïve Bayes

When features x_i are the number of occurrences of n possible events (words, votes, etc...)

p_{ki} = probability of *i-th* event occuring in class k

x_i = frequency of *i-th* event

The multinomial Naïve Bayes classifier becomes:

$$c_{NB} = \underset{k \in K}{\operatorname{argmax}} \left(\log P(c_k) + \sum_{i=1}^{n} x_i \bullet \log p_{ki} \right)$$

Naïve Bayes Intuition

- Assumes all features are independent with each other
- Independence assumption decouples the individual distributions for each feature
- Decoupling can overcome curse of dimensionality
- Performance robust to irrelevant features
- Very fast, low storage footprint
- Good performance with multiple equally important features

Naïve Bayes Applications

- Document Categorization
- NLP
- Email Sorting
- Collaborative Filtering
- Sports Prediction
- Sentiment Analysis

Example: Doc Classification

Want to classify documents into k topics. Document d consisting of words w_i is assigned to the topic c_{NB}:

$$c_{NB} = \underset{k \in K}{\operatorname{argmax}} P(c_k) \prod_i P(w_i \mid c_k)$$

$P(c_k)$ = topic frequency = $\dfrac{N_{docs(topic=c_k)}}{N_{docs}}$

$P(w_i \mid c_k)$ = word w_i frequency in all topic c_k docs

$$= \frac{N_{word=w_i(topic=c_k)}}{\sum_i N_{word=w_i(topic=c_k)}}$$

Laplace (add-1) Smoothing

What happens with $P(w_i \mid c_k) = 0$ for a particular i, k?

$$\mathrm{P}(\mathrm{c_k} \mid \mathrm{x_1}, \mathrm{x_2}, ..\mathrm{x_n}) = \mathrm{P}(\mathrm{c_k}) \prod_{i=1}^{n} P(x_i \mid c_k) = 0!$$

Solution is to add 1 to numerator & denominator:

$$P(w_i \mid c_k) = \frac{N_{word=w_i(topic=c_k)} + 1}{\sum_i \left(N_{word=w_i(topic=c_k)} + 1\right)}$$

NBC Application Roadmap

- Read Dataset
- Transform Dataset
- Create Classifier
- Train Classifier
- Make Prediction

Sports Prediction

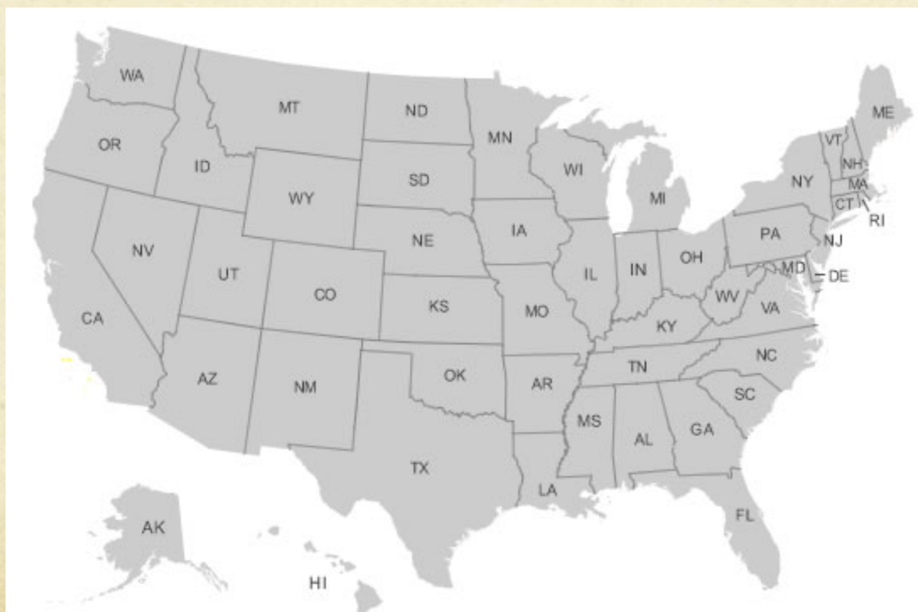
Who is favored to win the Superbowl?

Given the 2014 season game statistics for two teams, how can we make a prediction on the outcome of the next game using a Naïve Bayes classifier?

Sentiment Analysis

Which team is most favored in each state?

What if we analyzed tweets by sentiment and location using a Naïve Bayes Classifier?



Starter Code

Repo with Starter Code at:

https://github.com/guard0g/NaiveBayesForSuperbowl

IPython notebook:	NB4Superbowl.ipynb
Datasets:	SeattleStats.csv
	NewEnglandStats.csv