BC COMS 2710: Computational Text Analysis

Lecture 15 – Machine Learning: Text Classification (Naive Bayes)

Copyright © 2016 Barnard College

Announcements – Assignments

- Readings 04:
  - link posted to course site
  - due Sunday

- HW 02:
  - Due Wednesday night (last night)

- HW 03:
  - Released today
  - Due next Wednesday night

Copyright © 2016 Barnard College

Final Project – Deliverables

- **Project ideation – Friday May 28st**
  - https://www.overleaf.com/read/yzpgxcgsqdvp

- roughly 250 word overview of what you are interested in

Final Project – Deliverables

- Project ideation – Friday May 28st
  - 5 points
- Project proposal – ~~Friday June 4th~~ Sunday June 6th
  - 9 points
- Project presentations – Monday June 14th
  - 6 points
- Project submissions – Friday June 18th
  - 15 points
- http://coms2710.barnard.edu/final_project

Copyright © 2016 Barnard College – Adam Poliak

Notes – Pandas comment

When computing the same thing across a row or column, what should we do?

1. Define a function
2. apply the function

Looping through a dataframe is not ideal

Copyright © 2016 Barnard College

Slow Jupyterhub & your code

"A computer program does what you tell it to do, not what you want it to do."

Be careful when looping and adding to lists

Copyright © 2016 Barnard College

Machine Learning

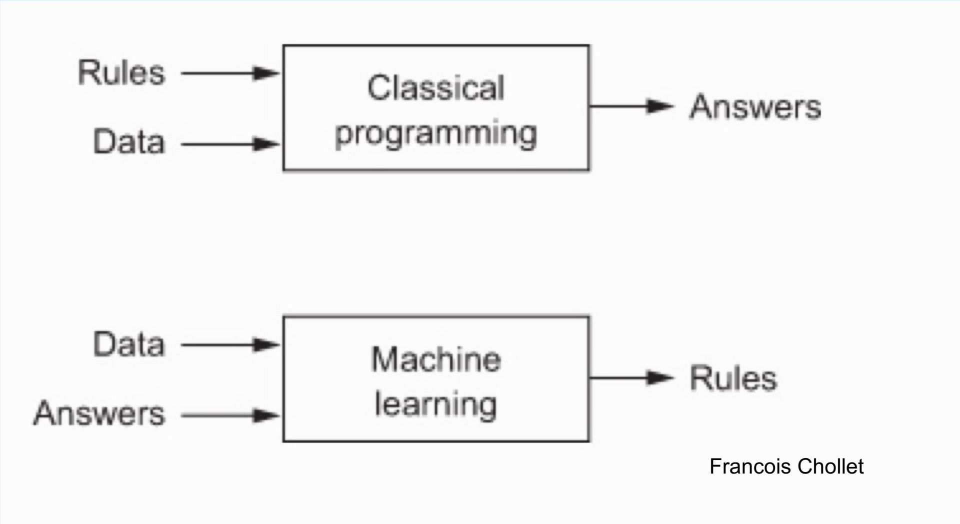
Copyright © 2016 Barnard College

Machine Learning Algorithm

A mathematical model

calculated based on sample data ("training data")

makes predictions or decisions without being explicitly programmed to perform the task



Copyright © 2016 Barnard College

Different Types of Machine Learning

- Supervised Learning
  - Learn rule from data and answers
- Unsupervised Learning
  - Learn a rule for patterns from data
- Reinforcement Learning
  - try your rule on a piece of data, and get feedback on how good your rule was

Slide from Tony Liu

Copyright © 2016 Barnard College

Prediction

Copyright © 2016 Barnard College

Guessing the Value of an Attribute

- Based on incomplete information

- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

Copyright © 2016 Barnard College

Two types of predictions: Classification & Regression

Classification = Categorical

Regression = Numeric

Predicting sentiment:

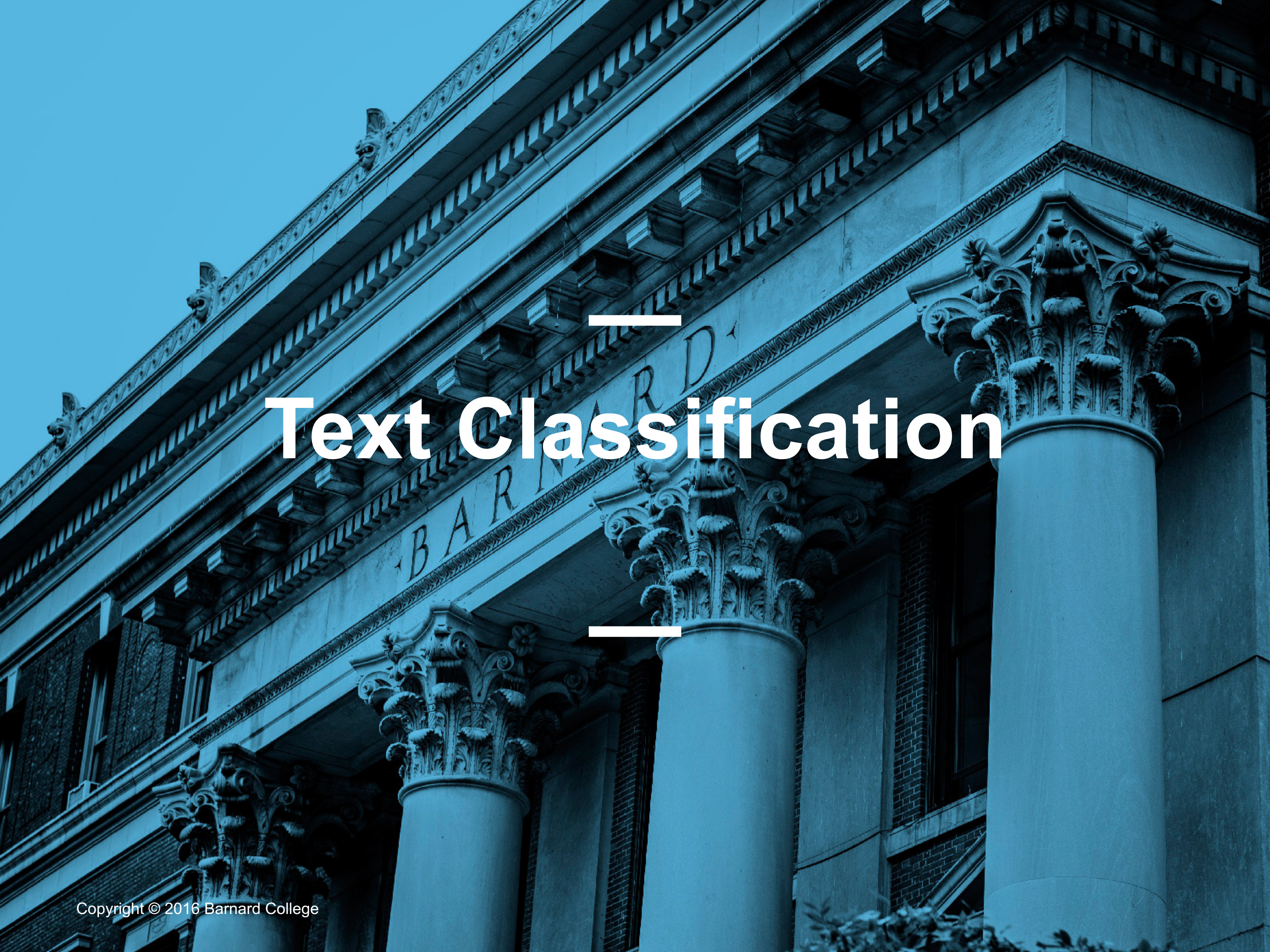
- Classification

  👍 👎

- Regression:

  [-1, …, 1]

Prediction Example: Hot dog or not Hot dog?



Text Classification

Copyright © 2016 Barnard College

Spam or Not Spam?

David, Adam 6	Tennis this week? - in playing tennis on Tuesday. It >>>> will b...
Citi Alerts	Your Citibank account statement is available online - com to y...
Humane Rescue Allia.	Your HRA E-Newsletter - Read news and events updates from ...
SLEEP NUMBER	Check out these limited-time Weekend Specials - PLUS get fre...
aishagaddafi11119	Inquiry for Investment. - Inquiry for Investment. Assalamu Alai...

Copyright © 2016 Barnard College

What is this medical article about?

MEDLINE Article


Available online at www.sciencedirect.com

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Susanne Gahl, Lise Menn, Gail Ramsberger, Daniel S. Jurafsky, Elizabeth Elder, Molly Rewega, and L. Holland Audrey

1. Introduction

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Sentiment Analysis

+ *...zany characters and richly applied satire, and some great plot twists*

– *It was pathetic. The worst part about it was the boxing scenes...*

+ *...awesome caramel sauce and sweet toasty almonds. I love this place!*

– *...awful pizza and ridiculously overpriced...*

Broad applications of sentiment analysis

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

Text Classification

Input:

- a document d
- a fix set of classes $C = \{c_1, c_2, \ldots, c_n\}$
- A training set of n labeled documents $(d_1, c_1), (d_2, c_2), \ldots, (d_n, c_n)$

Output:

- A learned classifier f
  - f is a mapping from $d \rightarrow c$

Copyright © 2016 Barnard College

Classifiers

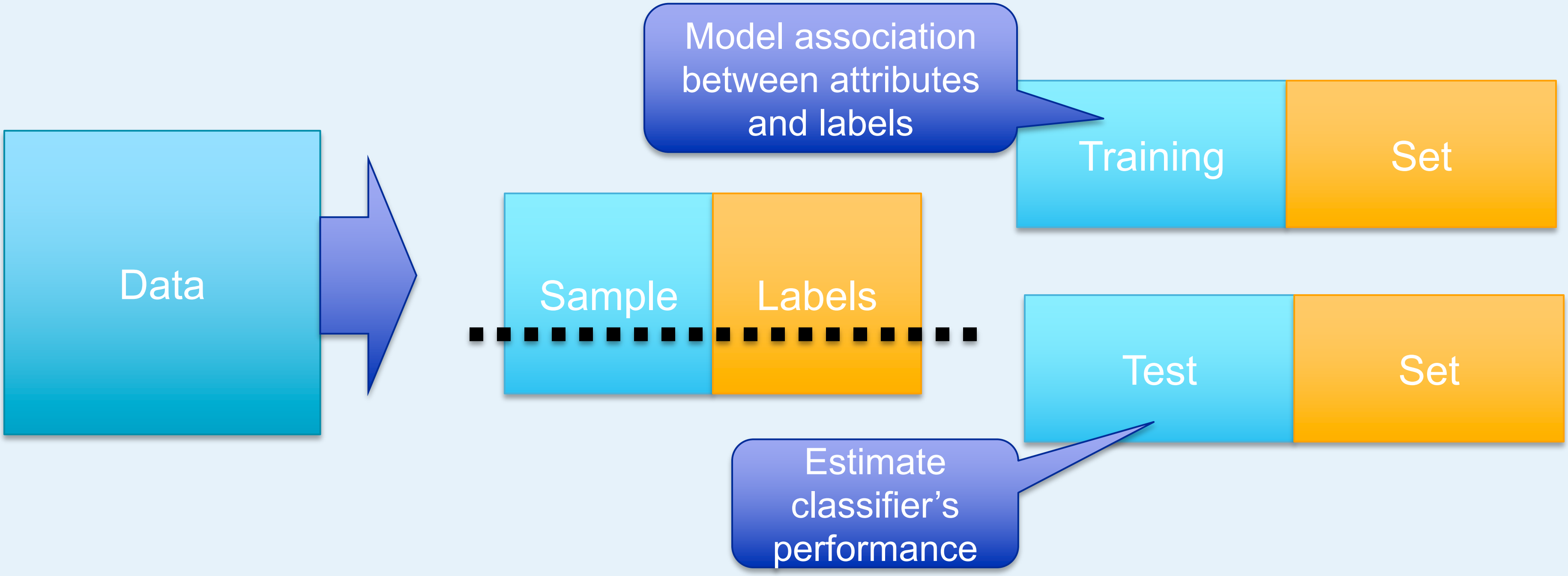
Copyright © 2016 Barnard College

Training a Classifier

Attributes (features) of an example

Predicted label of the example

Setup for training and evaluating a classifier



Copyright © 2016 Barnard College

Scikit-Learn

Copyright © 2016 Barnard College

Scikit-learn train and predict

scikit-learn uses a standard set of functions for all models

The two main ones for our purposes

- model.fit(X, y) — train the model with the given data set
- model.predict(X_test) — get predictions for the given test set

Slide from Jorge Mendez

Copyright © 2016 Barnard College

Different types of classifiers

- Neural Networks
- K-Nearest Neighbors
- Logistic Regression
- Naive Bayes
- ….

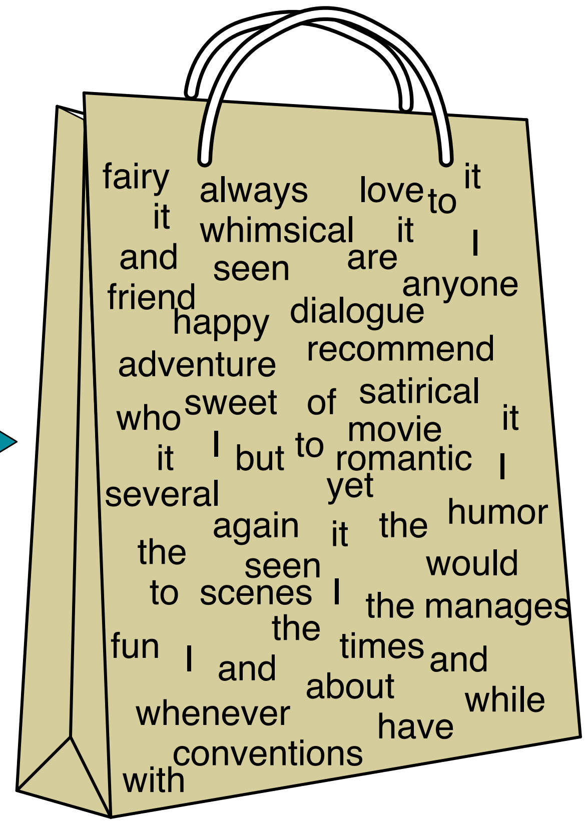
Copyright © 2016 Barnard College

Naive Bayes

Copyright © 2016 Barnard College

Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Slide from Dan Jurafsky

Classify document based on BoW

What is the probability of the class given the BoW

f(

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

)=c

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Bayes Rule

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Bayes Rule Derivation

Copyright © 2016 Barnard College

Bayes Rule for documents and classes

Given document *d*, what is the probability of category *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Naive Bayes Classifier

Choose category *c* that has the highest probability given document *d*

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

Slide from Dan Jurafsky

Naive Bayes Classifier

Choose category *c* that has the highest probability given document *d*

"Likelihood" "Prior"

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

How do we represent document *d*

Answer: Bag of Words

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Slide from Dan Jurafsky

Naive Bayes Independent Assumption

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Bag of Words assumption**: Assume position doesn't matter
- **Conditional Independence**: Assume the probabilities $P(x_i|c_j)$ are independent given the class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Plugging this into our prediction equation:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Computing probabilities

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Count Frequencies in training data

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Computing probabilities

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Count Frequencies in training data

$$\hat{P}(c_j) =$$

$$\hat{P}(x_i | c_j) =$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Computing probabilities

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Count Frequencies in training data

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(x_i | c_j) =$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Computing probabilities

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Count Frequencies in training data

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(x_i|c_j) = \frac{count(x_i, c_i)}{\sum_{x \in V} count(x, c_i)}$$

fraction of times word x_i appears
among all words in documents of topic c_i

Slide from Dan Jurafsky

Computing probabilities

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Count Frequencies in training data

Maximum Likelihood Estimation

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(x_i | c_j) = \frac{count(x_i, c_i)}{\sum_{x \in V} count(x, c_i)}$$

fraction of times word x_i appears among all words in documents of topic c_i

Slide from Dan Jurafsky

Question

What if we have seen no training ***positive*** documents with the word ***fantastic***?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

Probability of class will be 0, regardless of other words

$$c_{NB} = \operatorname*{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x \mid c)$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Smoothing

$$\hat{P}(x_i|c_j) = \frac{count(x_i, c_i) + 1}{\sum_{x \in V}(count(x, c_i) + 1)}$$

$$= \frac{count(x_i, c_i) + 1}{(\sum_{x \in V} count(x, c_i)\ ) + \ |V|}$$

Laplacian smoothing (add 1)

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Learning a Naive Bayes Classifier

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Predict with Naive Bayes Classifier

Give a document of composed of words ***X***

choose the class ***c***

that maximizes the Naive Bayes equation

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Scikit-learn train and predict

scikit-learn uses a standard set of functions for all models

The two main ones for our purposes

- model.fit(X, y) — train the model with the given data set
- model.predict(X_test) — get predictions for the given test set

Copyright © 2016 Barnard College

Consideration in Naive Bayes

Unknown Words

- words that are not in our training data but are in our test data
- **Ignore them**
  - Pretend they are not in our test

Stop Words

- For NB, removing them doesn't usually help

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Naive Bayes Example

Copyright © 2016 Barnard College

What label should we predict for test?

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

Copyright © 2016 Barnard College

Slide from Dan Jurafsky

Procedure

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

1. Prior from training:

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

2. Drop "with"

3. Likelihoods from training:

$$p(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

4. Scoring the test set:

Hint: for this example, do we care about words not in test?

Copyright © 2016 Barnard College

Slide from Dan Jurafsky