Is this spam?

---------- Forwarded message ----------
From: **E-Lotto** <infoalert@wp.pl>
Date: Fri, Sep 16, 2016 at 3:24 PM
Subject: Ref#.EL16BDXAUL16-03
To:

Hi,

E-Lotto congratulates you as the winner of $2,500,000.00. Email Rep (Aaron Martins) at aaronmarts@excite.com
with Ref#.EL16BDXAUL20

UMass Amherst

CS 585

Fall 2016

9/20 + 9/22

Classification Lecture
NB + LogReg + Eval

Slides from J&M 3rd ed. website

Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods

James Madison

Alexander Hamilton

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

#1: Saigon. Grey, Anthony

#2: Jerusalem the Golden. Drabble, Margaret

From shlomo argamon

Positive or negative movie review?

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

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

6

What is the subject of this article?

MEDLINE Article

?

MeSH Subject Category Hierarchy

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

7

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

Text Classification: definition

Prediction for a document

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR ("dollars" AND"have been selected")
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

~~Cons~~

Problems

- Less Explcit Lang.
- Language Evolves
- False Positive
- Multiple word senses
- Lexical Coverage

Classification Methods: Supervised Machine Learning

① Train

- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1, c_1), \ldots, (d_m, c_m)$
- *Output:*
  - a learned classifier $f: d \rightarrow c$

11

② Predict on new data

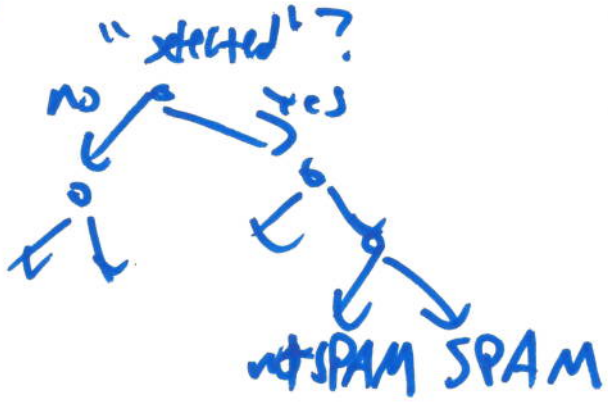
Models for Sup. learning

⇒ Naive Bayes

- k - Nearest Neighbors $f(d) = c$

⇒ Neural Networks

- Decision Trees



⇒ Support V. M.

⇒ Logistic Regression

The 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!

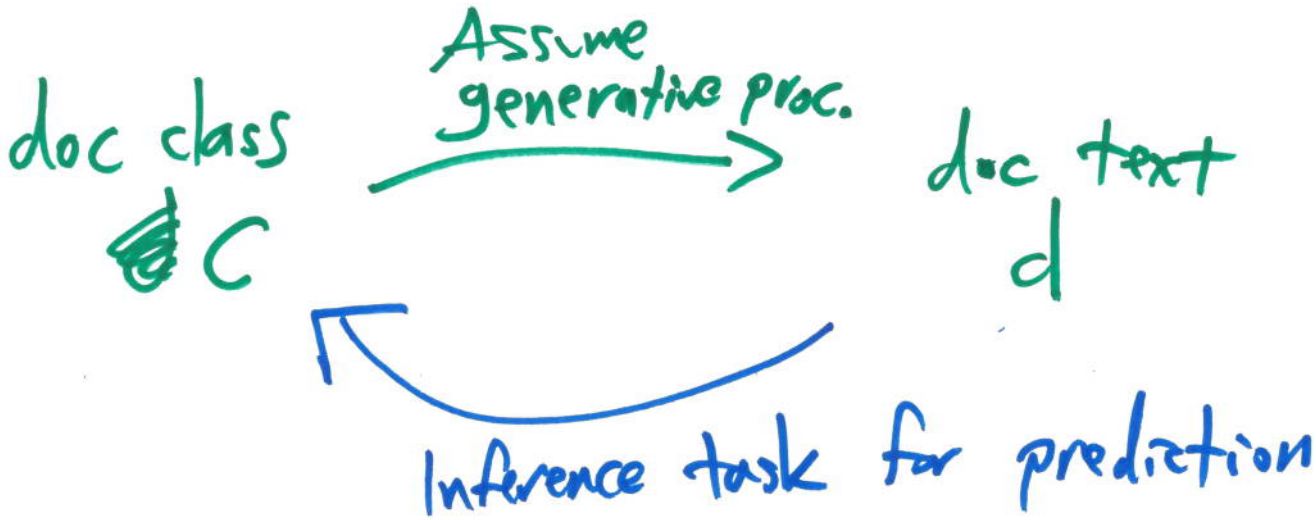
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

15

Classifier with BOW

$f(\ldots) = c$

Bayes Rule for Doc Classif.



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

Likelihood

Prior

Ignore ???

$$= \sum_{c \in C} P(d|c) P(c)$$

$$\max_{x \in \{-1,0,1,2\}} x^2 = 4$$

$$\operatorname{argmax}_{x \in \{-1,0,1,2\}} x^2 = 2$$

"argmax" notation

NB Classifier

"MAP" = max. posterior

$$c_{MAP} = \underset{c \in C}{\arg\max}\ P(c|d)$$

BR

$$= \underset{c \in C}{\arg\max}\ \frac{P(d|c)\,P(c)}{P(d)}$$ ← Constant

$$= \underset{c \in C}{\arg\max}\ P(d|c)\,P(c)$$

$$= \underset{c \in C}{\arg\max}\ P(\underbrace{x_1, x_2 \cdots x_N}_{\text{Doc Features}}|c)\,P(c)$$

Learning?

$\Downarrow$

$\simeq O(|X|^N C)$

Too many Params

$\Downarrow$

Just count!

Easy!

Multinomial Naïve Bayes Independence Assumptions

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

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

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

Can estimate

Multinomial Naïve Bayes Classifier

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

↓ Cond. indep. assump.

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

↓

Unigram LM for class c

Naïve Bayes as a Language Model

- Which class assigns the higher probability to s?

Model pos	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

P(I love this film | c)

P(text | c=pos) = P(I|c=p) P(love|c=p) . . .

0.1 x 0.1 x 0.01 x 0.05 x 0.1

I love this fun film

-vs-

P(text | c=neg) = 0.2 x 0.001 x .01 x .005 x a1

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Total # tokens in c_j

Parameter estimation

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic *j* by concatenating all docs in this topic
  - Use frequency of *w* in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive (*thumbs-up)***?

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

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Laplace (add-1) smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

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

$$\frac{count(w_i, c) + \alpha}{(\sum_{w} \ldots) + \alpha |V|}$$

Multinomial Naïve Bayes: Learning

- 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 $Text_j$

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

size ? => $|V| \cdot |C|$

4

100k

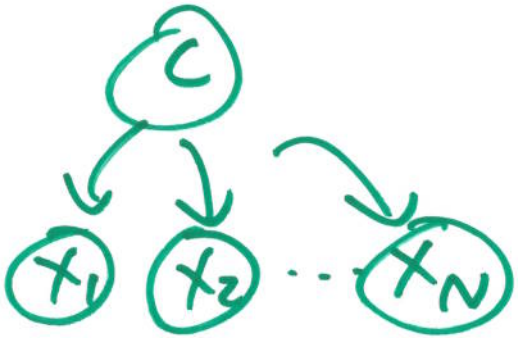
Summary: Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Robust to Irrelevant Features

  Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features

  Decision Trees suffer from *fragmentation* in such cases – especially if little data
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
  - **But we will see other classifiers that give better accuracy**

Multinom. Naive Bayes



To classify

$$\text{argmax}_c \; P(c,d)$$

Conf. Indep. Assump.

$$= \cdots \; P(c)\,P(d|c)$$

$$= \cdots \; P(c)\prod_{i=1}^{N} P(x_i|c)$$

NB $\Leftrightarrow$ Unigram LM only if features = words

other Features

- Is the word an emoticon?
  - Happy emo? $P(:) \mid c)$
  - Sad emo? $P(:] \mid c)$
  - IS_HAPPY_EMO $P(:-)))) \mid c) = ??$

Lists

Might generalize better $P(\text{IS\_HAPPY\_EMO} \mid c)$

Naïve Bayes in Spam Filtering

- SpamAssassin Features:
  - Mentions Generic Viagra
  - Online Pharmacy
  - Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - Phrase: impress ... girl
  - From: starts with many numbers
  - Subject is all capitals
  - HTML has a low ratio of text to image area
  - One hundred percent guaranteed
  - Claims you can be removed from the list
  - 'Prestigious Non-Accredited Universities'
  - http://spamassassin.apache.org/tests_3_3_x.html

Features: Domain Knowledge

Feature Engineering

NB as a Log-Linear Model

Pred Time $P(c|d) = \frac{1}{P(d)} P(c) \prod_{i=1}^{N} P(x_i|c)$

$\log P(c|d) = \underbrace{-\log P(d)}_{\text{Constant w.r.t. } c} + \underbrace{\log P(c)}_{w_{c,0}} + \sum_{i=1}^{N} \underbrace{\log P(x_i|c)}_{w_{x_i,c}}$

Log-Lin Model

$$P(c|d) = \frac{e^{\sum_{j=1} w_{j,c} f_j}}{\sum_{c'} e^{\sum_j w_{j,c'} f_j}}$$

$\log P(\text{dog}|\text{pos}) \equiv w_{\text{dog,pos}}$

For MNB equivalence

f_j = Num. times word j occurs

MuHinom. Logistic Regression

NB vs Log. Reg

NB: Generative Model $P(c,d) = P(c)\,P(d|c)$

Train: maximize likelihood ↑ $\Rightarrow$ Counting

Log. Reg: Discriminative Model $P(c|d)$ [Has no opinion about $P(d|c)$]

Train: maximize likelihood ↑

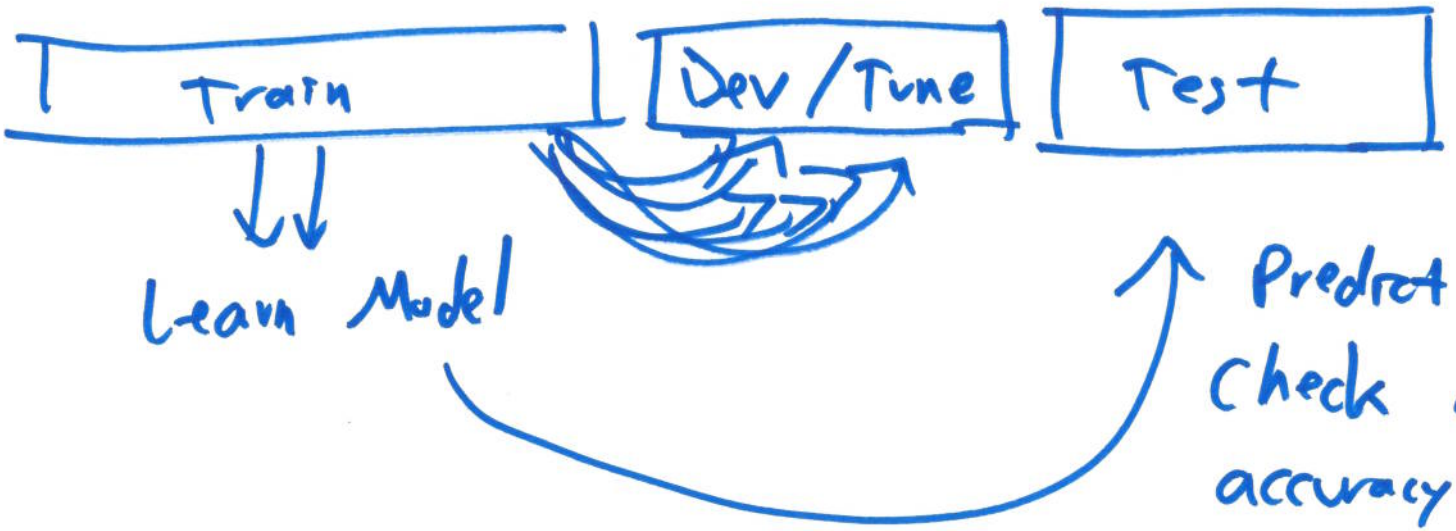
Algo: iterative gradient descent

$\Rightarrow$ Returns $\vec{w}$

Why? No CI assumptions in LR

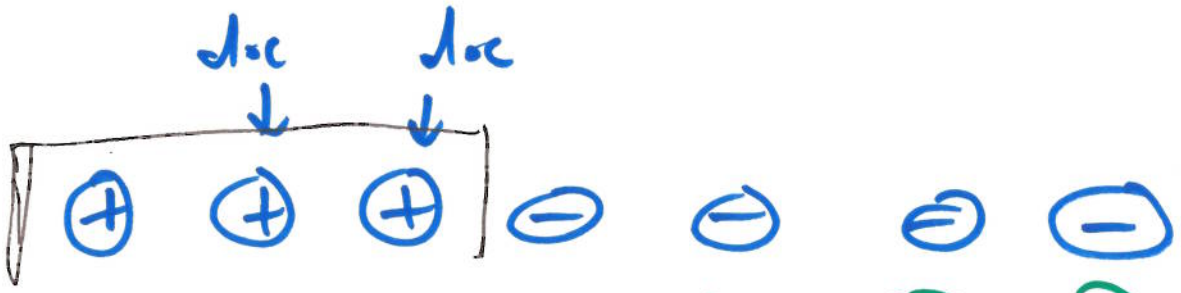
Evaluation

Data Splits : All annotated data



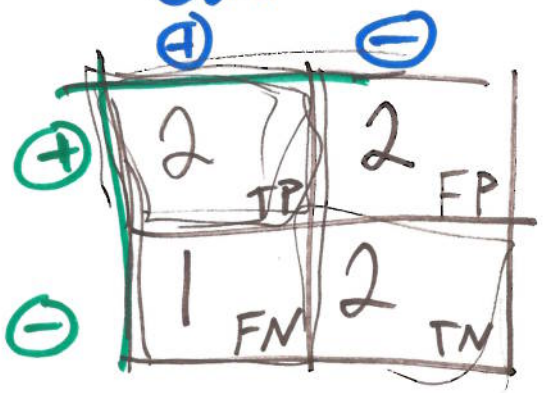
Tuning hyperparameters? ⇒ Dev
- Pseudocount α

Metrics



Acc Rate: % correct = $\frac{4}{7}$

Confusion Matrix



Prec & Recall

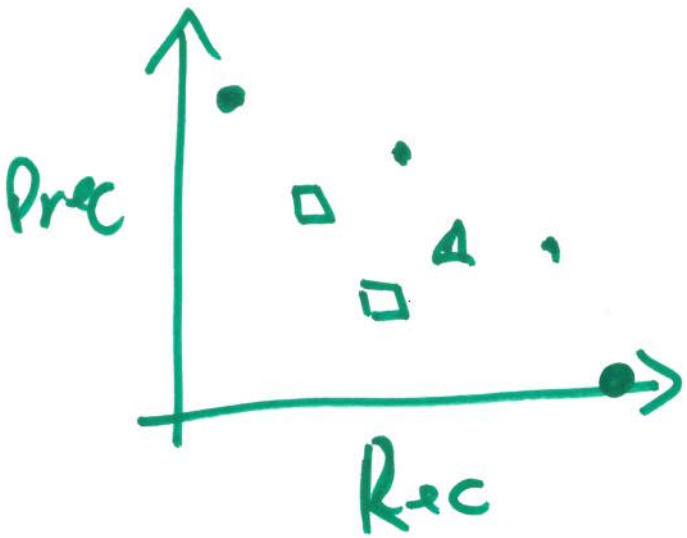
Prec = % correct of POS preds

$= \frac{TP}{TP+FP} = \frac{2}{4} = \frac{1}{2}$

Recall = % of gold POS we found

$= \frac{TP}{TP+FN} = \frac{2}{2+1} = \frac{2}{3}$

Prec & Recall Trade off



Single Number from P, R

$$F\text{-}Score := \frac{2PR}{P+R}$$

"Harmonic Mean"

Development Loop - Feature Engr