Text Classification 1

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

January 12, 2017

Based on slides from Nathan Schneider, Noah Smith, Dan Klein and everyone else they copied from.

Text Classification 1

Introduction to Text Classification

Naive Bayes Classification

Course Projects

Text Classification

Introduction to Text Classification

Naive Bayes Classification

Course Projects

Sentiment Analysis

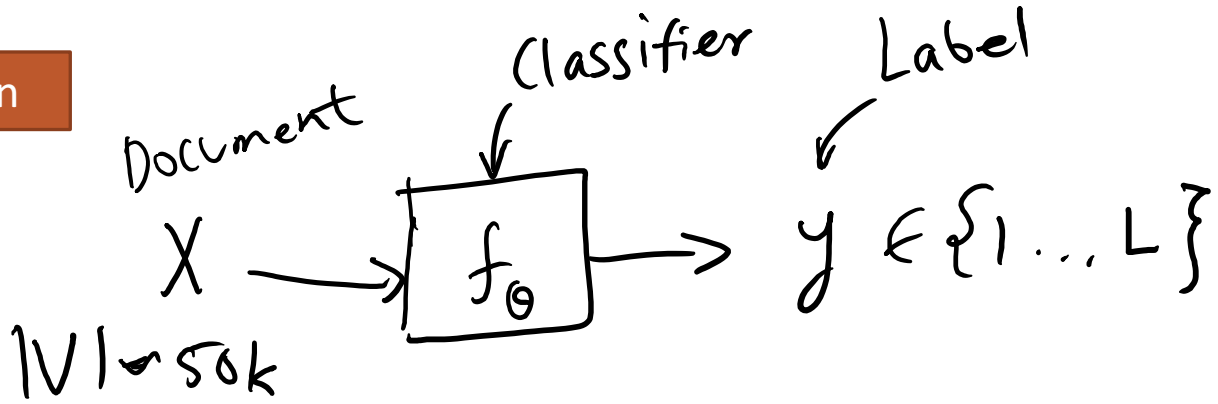
Filled with horrific dialogue, laughable characters, a laughable plot, ad really no interesting stakes during this film, "Star Wars Episode I: The Phantom Menace" is not at all what I wanted from a film that is supposed to be the huge opening to the segue into the fantastic Original Trilogy. The positives include the score, the sound …

Other Examples

- Reviews of films, restaurants, products: positive vs. negative
  - Amazon reviews data, IMDB reviews data Yelp
- Library-like subjects (e.g., the Dewey decimal system)
- News stories: politics vs. sports vs. business vs. technology ...
  - 20 newsgroup data
- Author attributes: identity, political stance, gender, age, ...
- Email: spam vs. not
  - Gmail: important, promotion, updates, social media, …
- What is the reading level of a piece of text?
  - Automatic graders?
- How influential will a scientific paper be?
- Advertisement recommendations …
- Will a piece of proposed legislation pass?
  - Identify the presidential candidate from speeches
- Post recommendations / Fake news detection
  - Can majorly influence the world!

Formal Setup

Classification

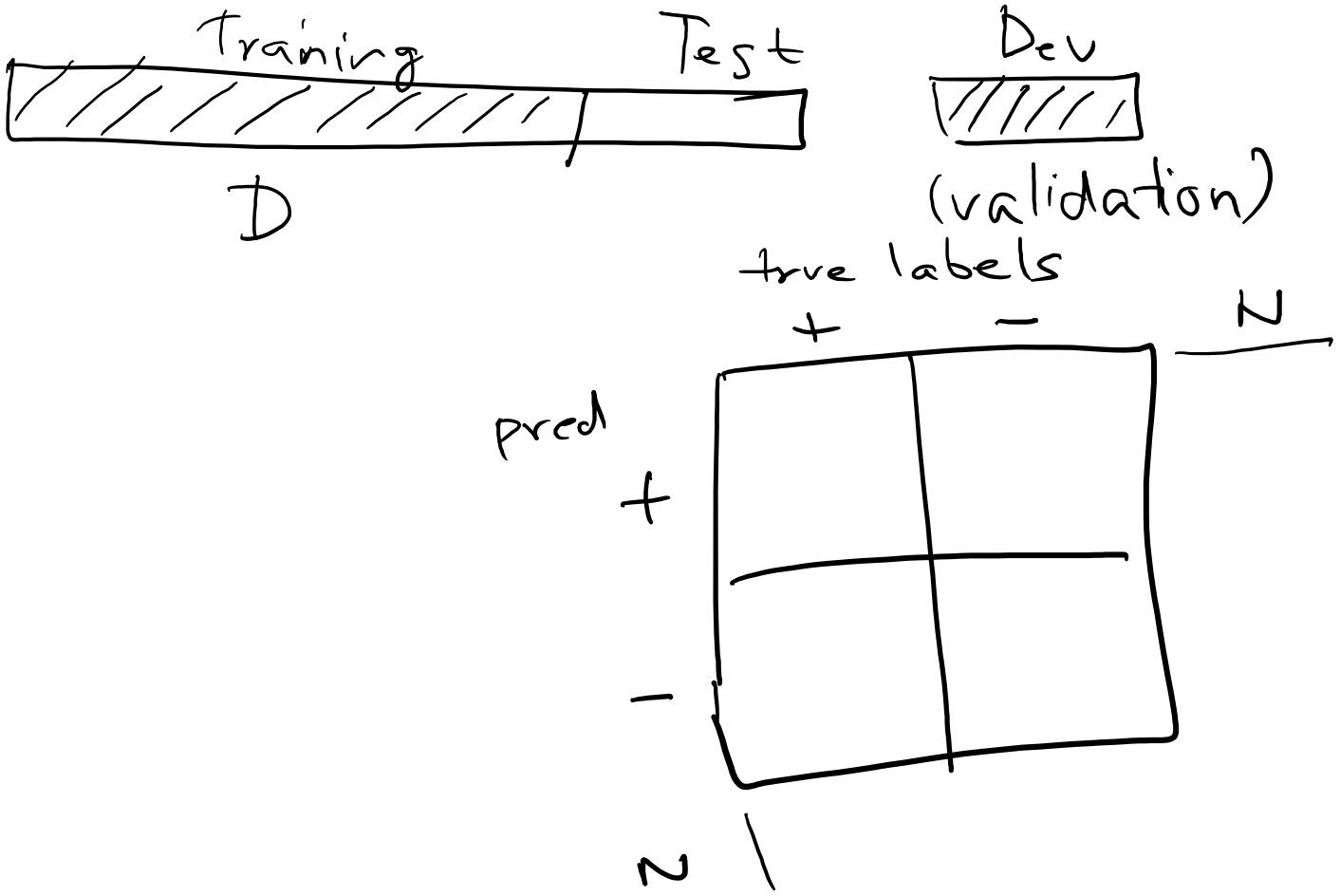


Supervised Learning

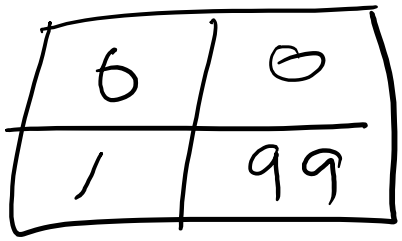
Dataset $D = \{x_i, y_i\}_{1:N}$

Confusion Matrix

Evaluation: Contingency Table

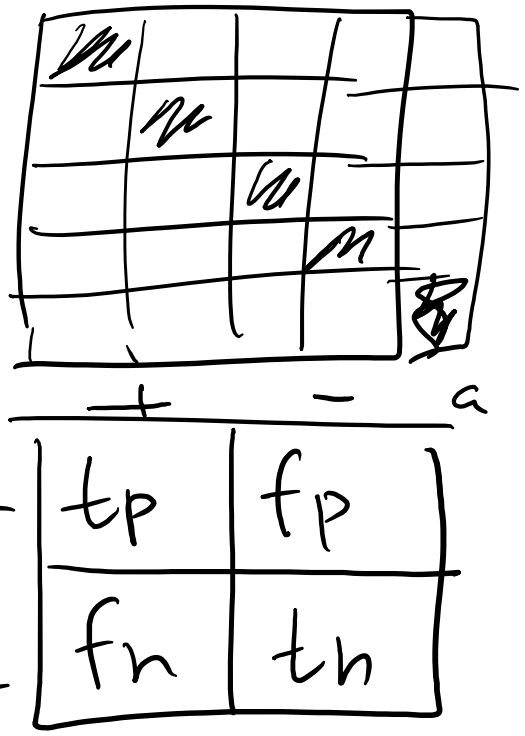


Accuracy



$$\frac{\#\text{correct}}{\#\text{total}}$$

$$\frac{t_p + t_n}{t_p + t_n + f_n + f_p}$$



Problem

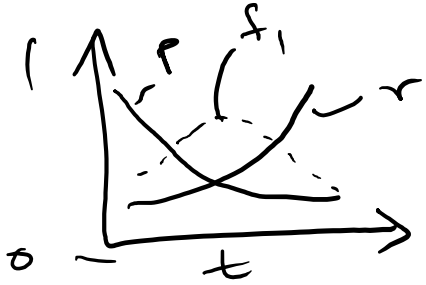
- Class imbalance hurts..
- Getting one class right matters more than the other (retrieval)

Precision and Recall

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

	+	−
+	tp	fp
−	fn	tn



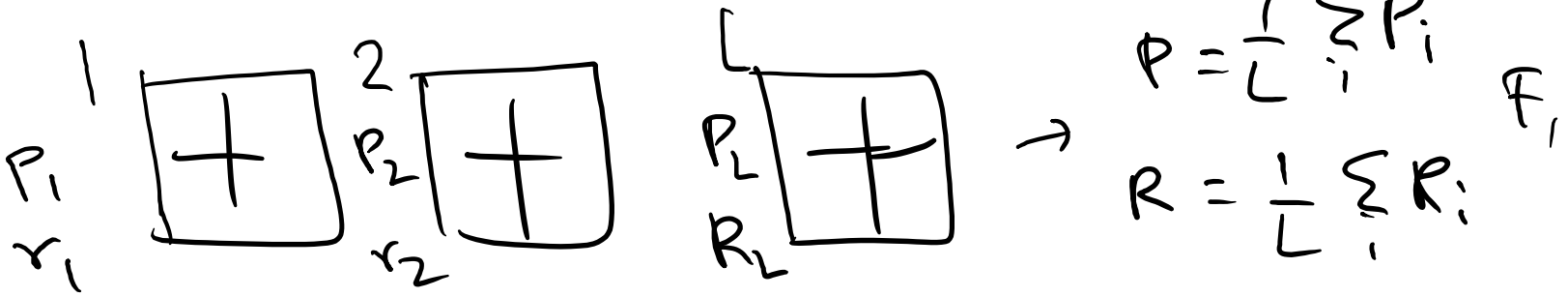
$$F_{score} = \frac{2pr}{p + r}$$

↖ Harmonic mean P and R

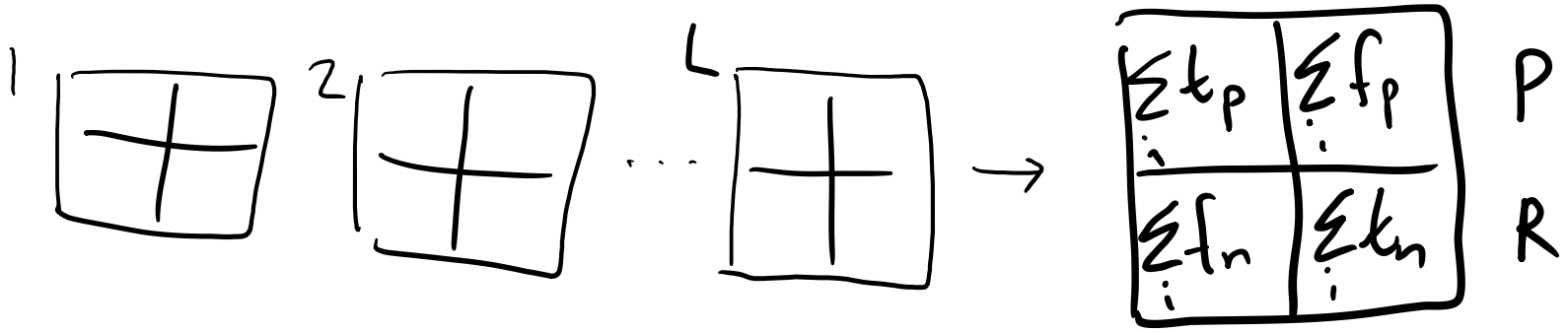
$$F_\beta = (1 + \beta^2)\frac{pr}{(\beta^2 p) + r}$$

>2 Classes?

Macro-averaged Measures

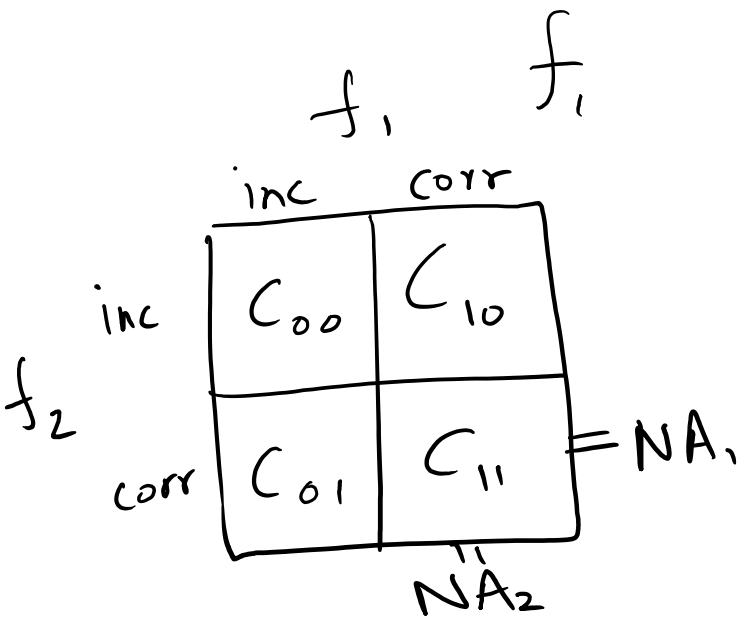


Micro-averaged Measures



McNemar's Test, *Psychometrika,* (1947)
More tests in Smith book, appendix B

Statistical Significance



$k = \min(C_{01}, C_{10})$

Binomial (N, P) where $N = C_{01} + C_{10}$, $P \to 0.5$

$$p\text{-value} = \frac{1}{2^{C_{01}+C_{10}-1}} \sum_{j=0}^{k} \binom{C_{01}+C_{10}}{j}$$

← "Choose"

Text Classification

Introduction to Text Classification

Naive Bayes Classification

Course Projects

Classification using Joint Prob

$f_\theta : X \rightarrow y$

$$f_\theta(X) = \underset{y}{\operatorname{argmax}}\ p(y/X)$$

$$= \underset{y}{\operatorname{argmax}}\ \frac{p(X,y)}{p(X)} \quad \text{Bayes Rule}$$

$$= \underset{y}{\operatorname{argmax}}\ p(X,y)$$

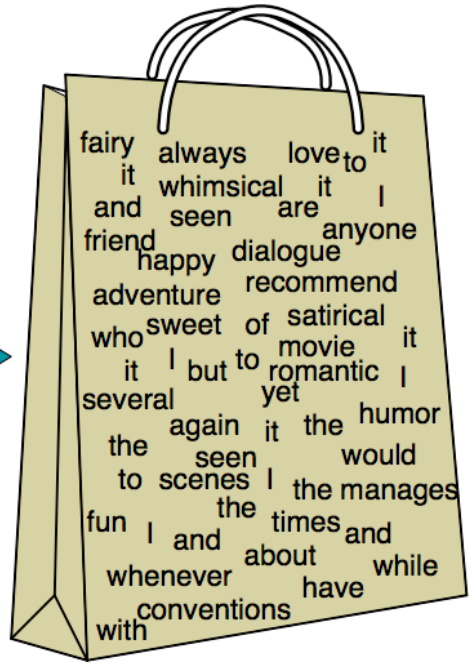
$$= \underset{y}{\operatorname{argmax}}\ p(y)\,p(X/y) \quad \text{Bayes Rule}$$

Naïve Bayes Classifier

Two assumptions

- Word ordering does not matter (**Bag of Words**)

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!



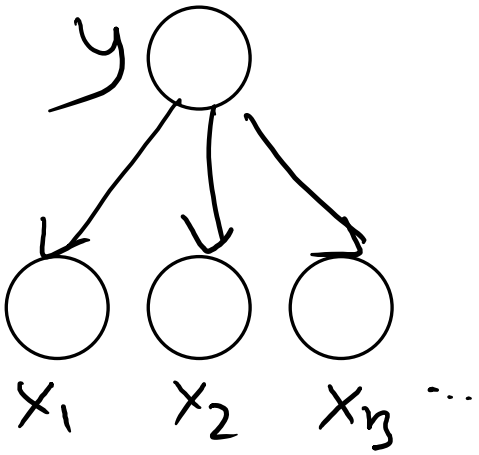
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

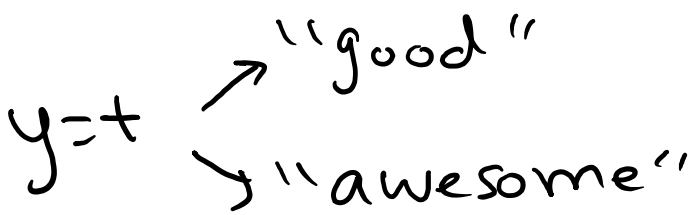
Naïve Bayes Classifier

Two assumptions

- Word ordering does not matter (**Bag of Words**)
- Words are independent given category

$$P(X|y) = \prod_i P(x_i|y)$$





Estimation of Parameters

$$f_\theta = \underset{y}{\operatorname{argmax}}\ P(y) \prod_i p(x_i \mid y)$$

$P(y)$: L — $p(x_i \mid y)$: $V \times L$

$$P(y) = \frac{\#\, y_j = y}{N}$$

$$P(x_i \mid y) = \frac{\#\,(x_{ji} = x_i \wedge y_j = y_i)}{\#\, y_j = y}$$

$$D = \{x_j, y_j\}$$

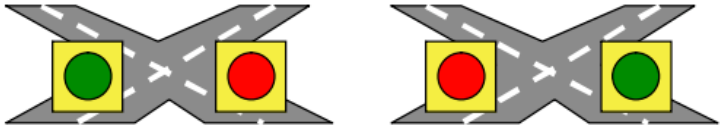
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_j P(x_j, y_j)$$

$$= \underset{\theta}{\operatorname{argmax}} \log \prod_j P(x_j, y_j)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_j \log P(y) + \sum_i \log P(x_i \mid y)$$

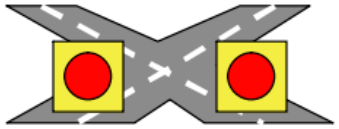
Problem with Naïve Bayes

Lights Working



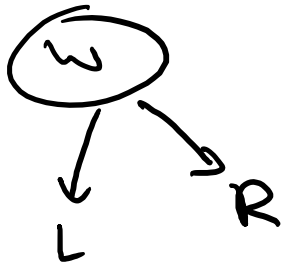
P(g,r,w) = 3/7 P(r,g,w) = 3/7

Lights Broken



P(r,r,b) = 1/7

w?



$P(w) = 6/7$ $P(b) = 1/7$

$P(r|w) = 1/2$ $P(r|b) = 1$

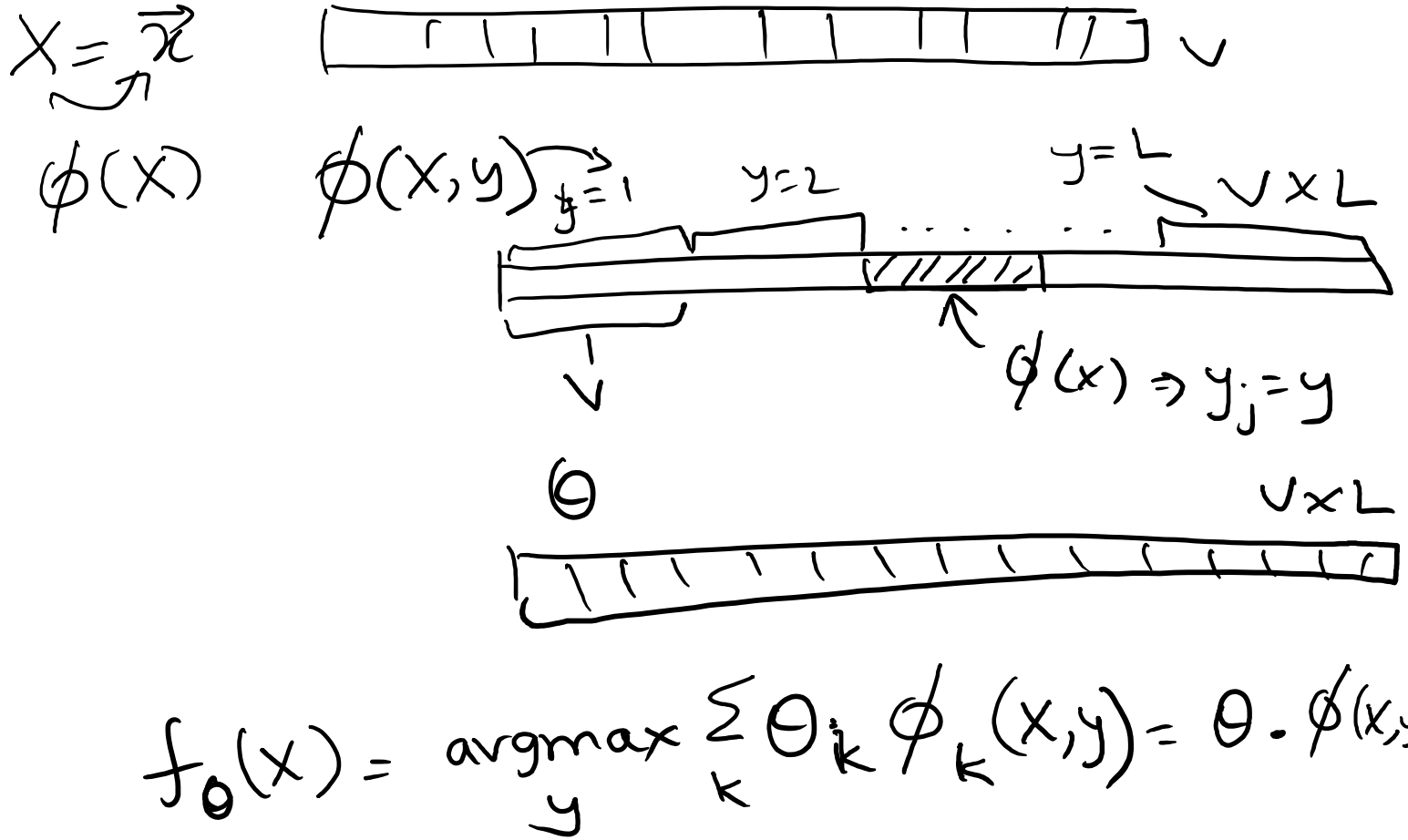
$P(g|w) = 1/2$ $P(g|b) = 0$

$P(w,r,r) = 6/7 \times \frac{1}{2} \times \frac{1}{2} = \frac{6}{28}$

$P(b,r,r) = \frac{1}{7} \times 1 \times 1 = \frac{4}{28}$

$P(w|r,r) = \frac{6}{10}$

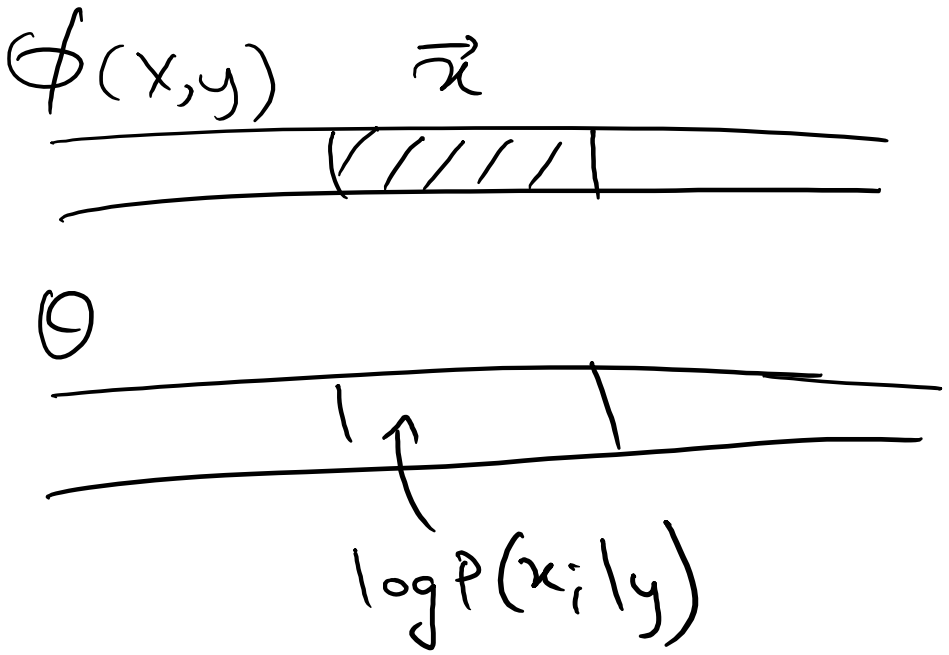
Linear Models



$$f_\theta(X) = \arg\max_y \sum_k \theta_k \phi_k(X, y) = \theta \cdot \phi(X, y)$$

Naïve Bayes as a Linear Model

$$f_\theta = \underset{y}{\operatorname{argmax}} \log P(y) + \sum_i \log P(x_i | y)$$



Text Classification

Introduction to Text Classification

Naive Bayes Classification

Course Projects

Group Projects

Groups for the Project

- Ideal team size is 3
  - Absolute maximum of 4
  - <3 if I approve (ongoing work)

Submit Four Reports

- First two reports are very short (1 page)
- Final report matters the most

How do I know it's NLP?

- Output is *any* phrase or sentence, definitely!
- Input is any phrase or sentence
  - Output is a sequence or structure (yes!)
  - Classification: only if over words or phrases
- Output is linguistic classes/structures (yes!)

Scope of Work

Novelty

- New Task/Data
- New Method/Models
- New Application of Existing Method to Existing Task

But not too much!

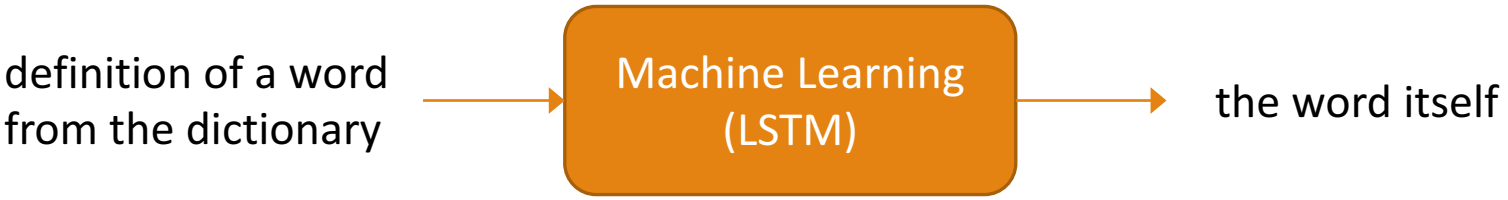
- You do not have much time!
- Aim to have the whole pipeline done soon
- Keep the "scale" of the data small, sub-sample if needed
- Better to have a complete finished report
  - than grand ideas that did not work

Reuse

- You do not have to code everything
- Exploit existing code, datasets, libraries, web services
- Do not reinvent all the wheels!

Example 1: What's the word..

What's the word for someone using pretentious words? lexiphanic



This can be a cool Twitter bot!

Evaluation

- Accuracy of guessing the word, using definitions from different dictionary?
- Baselines: Google, reversedictionary.org, …

Example 2: SQuAD

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

https://rajpurkar.github.io/SQuAD-explorer/

How many siblings did Tesla have?
four
What was Tesla's brother's name?
Dane
What happened to Dane?
killed in a horse-riding accident

Rank	Model	Test EM	Test F1
1	r-net (ensemble) (Microsoft Research Asia)	74.5	82.0
2	BiDAF (ensemble) Allen Institute for AI & University of Washington (Seo et al. '16)	73.3	81.1
3	Dynamic Coattention Networks (ensemble) Salesforce Research (Xiong & Zhong et al. '16)	71.6	80.4
4	BiLSTM-Fusion German Research Center for Artificial Intelligence	70.8	78.9

Datasets and Papers

Data

- Search Kaggle, Quora, etc for large text datasets
- See recent papers in NLP for released datasets
- Look for “shared tasks”, “challenges”, workshops
- Links to some existing datasets coming to website soon

Papers

- NLP Conferences: ACL, EMNLP, NAACL
- ML Conferences: NIPS, ICML, ICLR, AAAI
- Data focused venues: TREC/TAC, SemEval, CONLL
- Workshops at these conferences: interesting directions
- More papers coming soon to the website

Writing the Pitch

Team

- Team name and members
- Single sentence description for each member
  - (approximately) what they will do
- Single sentence on what makes your team diverse

Project

- Motivation and Problem Description
- Planned approach: tentative
- Evaluation: usually, most important

Appointment

- If 1 or 2, meet me before/on January 17 (o.w. no need)
- Every group has to meet afterwards to discuss the project

Upcoming...

Homework

- Homework 1 is up!
- Next lectures will continue with more details
- Sign up for the Kaggle account (@uci.edu email)
- Due: January 26, 2017

Project

- Project pitch is due January 23, 2017!
- Start assembling teams now! (use Piazza)
- Start looking at papers, data, etc. for ideas