Text Classification & Linear Models

CMSC 723 / LING 723 / INST 725

Marine Carpuat

Slides credit: Dan Jurafsky & James Martin, Jacob Eisenstein

Logistics/Reminders

Homework 1 – due Thursday Sep 7 by 12pm.

Project 1 coming up

• Thursday lecture time: project set-up office hour in CSIC 1121

Recap: Word Meaning

2 core issues from an NLP perspective

- **Semantic similarity**: given two words, how similar are they in meaning?
- Key concepts: vector semantics, PPMI and its variants, cosine similarity
- Word sense disambiguation: given a word that has more than one meaning, which one is used in a specific context?
- Key concepts: word sense, WordNet and sense inventories, unsupervised disambiguation (Lesk), supervised disambiguation

Today

- Text classification problems
 - and their evaluation
- Linear classifiers
 - Features & Weights
 - Bag of words
 - Naïve Bayes

Text classification

Is this spam?

```
From: "Fabian Starr"
<Patrick_Freeman@pamietaniepeerelu.pl>
Subject: Hey! Sofware for the funny prices!
Get the great discounts on popular software today
for PC and Macintosh
http://iiled.org/Cj4Lmx
70-90% Discounts from retail price!!!
All sofware is instantly available to download - No
Need Wait!
```

What is the subject of this article?

MEDLINE Article



Available on the at www.cciencedirect.com

ton analyzing a part to an



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

Susanne Gohl, " Liee Monn," Gall Rameherger," Daniel S. Jurafeky, " Elizabeth Elder," Molly Reseas," and L. Halland Audrey."

> Named Sciences, Sandrale and Other among of finingly dealer fro, until security of Atlanta, Partie of State AMOUNT THE REE

This easily investigates there is since they have been argued to define "soundied form" in account accomplishment. Accounts errors, attack etc. and happing of sage. We be married to dain the account conditing examinate only years. Effect is entry to a free of peaks arrows. Using a pleasibility judgment using an about the a robal group of a placing performed algorithments in the construction of the construction. We then come the observation the greatment appearing bother than contact the construction of the cons arrents arrange. Nation of realizables on to one shallbanks make that readers of an including order. Were served to one, their electronics region in interference in the second collection of the second region in a collection of the second region and indicate the second region and re tie de no ment. These indige egges the "exemisal form" where impurey and initial blass. # 200 Danier Ing All rights regress

1. Introduction

The simplicity of Paramologi Serry' or Paramologic word order," for named and aphada comprehension has after from taken as sufferitient in the sentence comprehension literature. However, as has been publish out by Menn (2005), the printingal status of severtal form half reads explanation. Different definitions of "securing law" yield belong different productions. One approach to the definition of severing emission Seen to that Implicit in Basis, Property, and Wallace (1855, how all). Bein et al. rate that gentrous with Application. Didn't order recepts the seconds. word order for English. A second approach is beard on grissis "moreon" and see and differ a newsyrankel any word order that diverges from the (_XX). (offerbAP) and position assemble for the stepdisplace of English perimons. Reset on this position standing of secondary, Kepl (1998) expensions are many with progressing upon check by differing prome for article policie, in particular for patients with "agreementary," for masters that are enalogous in-

compared to active. Although the preside definition of communicipie control of toring, Levin & Represent Here, 1915, unansative term are generally units. stead to be interesting upon place (softer) subjets represent Undergoor organisms. Exemples of unwestpaties serie include serie like early and black. Under the transferrational analysis assured in Kepl (HHF), the soften colificia of communities series are listed six management in all and addition in the engineering. Disputers, alon feeder lake the my are difficulty as pushe serieses, asserting to Keph analysis, and should be as hard as positive for actuals considers.

the feature giving that is the greater difficulty of peachers

A different approach to paramial form has been monant by Mann et al. (1998) who suggest that masaried from relies on the most frequent contacts from for a place spek. Chaire this along polyatic problems with producing and understanding peoples derive from the Set that, for most impositive raths, positive caper has Inspecting that patient. One mediation of this premium. also advanced by Gald (2012), is that comprehension effectly disease very with the bested found for words

196-1965 - An for the Will State for All Specimen

ACCUPATION AND DESCRIPTION AND

MeSH Subject Category Hierarchy

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

Text Classification

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

Text Classification: definition

- Input:
 - a document d
 - a fixed set of classes $Y = \{y_1, y_2, ..., y_J\}$
- Output: a predicted class $y \in Y$

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

Classification Methods: Supervised Machine Learning

Input

- a document d
- a fixed set of classes $Y = \{y_1, y_2, ..., y_J\}$
- a training set of m hand-labeled documents $(d_1, y_1), \dots, (d_m, y_m)$

Output

• a learned classifier $d \rightarrow y$

Aside: getting examples for supervised learning

- Human annotation
 - By experts or non-experts (crowdsourcing)
 - Found data

- How do we know how good a classifier is?
 - Compare classifier predictions with human annotation
 - On held out test examples
 - Evaluation metrics: accuracy, precision, recall

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Precision and recall

• **Precision**: % of selected items that are correct

Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

A combined measure: F

 A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F1 measure
 - i.e., with β = 1 (that is, α = ½):

$$F = 2PR/(P+R)$$

Linear Classifiers

Bag of words





$$\mathbf{w}_1 = \{\text{great}, \text{sunset}, \text{tonight}, \ldots\}$$

$$\mathbf{w}_2 = \{ \text{ugly, skies, buford, } \ldots \}$$

	અર્વે	ignaj.	acits 	γê	hind	Vil	jord	96	mds	OJ.		ાઉ	£.
$\mathbf{x}_1^T =$							00						
$\mathbf{x}_{2}^{T} =$	0	0	00	0	00	1	00	0	00	0	00	1	0.

```
\mathbf{x}_1 = \{ \text{great} : 1, \text{sunset} : 1, \text{tonight} : 1, \ldots \}
\mathbf{x}_2 = \{ \text{ugly} : 1, \text{skies} : 1, \text{buford} : 1, \ldots \}
```

Defining features

Suppose $y \in \mathcal{Y} = \{\text{pos}, \text{neg}, \text{neut}\}$. Then,

$$\mathbf{f}(\mathbf{x}, y = \mathsf{pos}) = [\mathbf{x}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, 1]^\mathsf{T}$$

$$\mathbf{f}(\mathbf{x}, y = \mathsf{neg}) = [\mathbf{0}^\mathsf{T}, \mathbf{x}^\mathsf{T}, \mathbf{0}^\mathsf{T}, 1]^\mathsf{T}$$

$$\mathbf{f}(\mathbf{x}, y = \mathsf{neut}) = [\mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{x}^\mathsf{T}, 1]^\mathsf{T}$$

Defining features

Suppose $y \in \mathcal{Y} = \{\text{pos}, \text{neg}, \text{neut}\}$. Then,

$$\mathbf{f}(\mathbf{x}, y = \mathsf{pos}) = [\mathbf{x}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, 1]^\mathsf{T}$$
$$\mathbf{f}(\mathbf{x}, y = \mathsf{neg}) = [\mathbf{0}^\mathsf{T}, \mathbf{x}^\mathsf{T}, \mathbf{0}^\mathsf{T}, 1]^\mathsf{T}$$
$$\mathbf{f}(\mathbf{x}, y = \mathsf{neut}) = [\mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{x}^\mathsf{T}, 1]^\mathsf{T}$$

The feature vector is composed of individual feature functions, e.g.:

$$f_{176}(\mathbf{x}, y) := x_{176} \times \delta(y = \text{pos})$$

 $= \delta(\text{great} \in \mathbf{w} \land y = \text{pos})$
 $f_{177}(\mathbf{x}, y) := x_{177} \times \delta(y = \text{pos})$
 $f_{10176}(\mathbf{x}, y) := x_{176} \times \delta(y = \text{neg}) \dots$

We usually add an "offset" feature at the end of each vector.

Linear classification

We can then define weights for each feature:

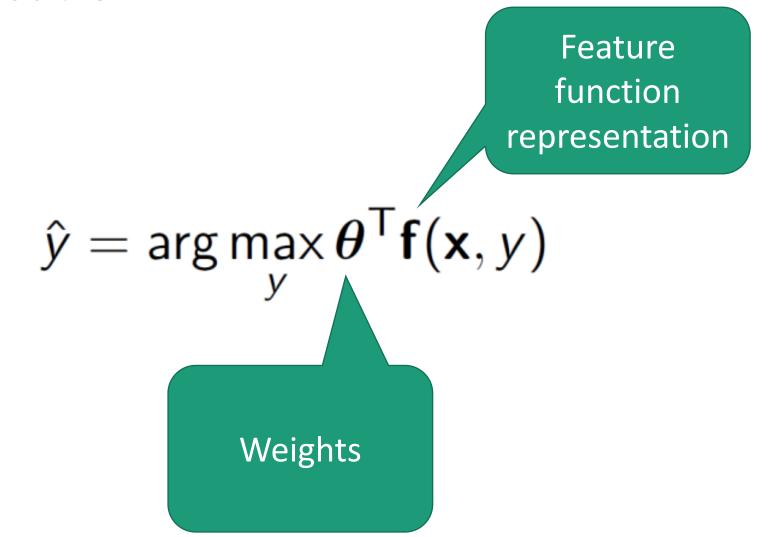
$$eta = \{\langle ext{great}, ext{pos} \rangle = 1, \langle ext{great}, ext{neg} \rangle = -1, \langle ext{great}, ext{neut} \rangle = 0, \langle ext{ugly}, ext{pos} \rangle = -1, \langle ext{ugly}, ext{neg} \rangle = 1, \langle ext{ugly}, ext{neut} \rangle = 0, \langle ext{buford}, ext{pos} \rangle = 0, \langle ext{buford}, ext{neg} \rangle = 0, \langle ext{buford}, ext{neut} \rangle = 0, \dots \do$$

We can arrange these weights into a vector.

The **score** for any instance and label is equal to the sum of the weights for all features in the instance:

$$\psi_{y,\mathbf{x}} = \sum_{n} \theta_{n} f_{n}(\mathbf{x}, y)$$
$$= \boldsymbol{\theta}^{\mathsf{T}} \mathbf{f}(\mathbf{x}, y)$$
$$\hat{y} = \arg \max_{y} \boldsymbol{\theta}^{\mathsf{T}} \mathbf{f}(\mathbf{x}, y)$$

Linear Models for Classification



How can we learn weights?

By hand

- Probability
 - e.g., Naïve Bayes

- Discriminative training
 - e.g., perceptron, support vector machines

Generative Story for Multinomial Naïve Bayes

 A hypothetical stochastic process describing how training examples are generated

For each document i,

- draw the label $y_i \sim \text{Categorical}(\mu)$
- draw the vector of counts $\boldsymbol{x}_i \sim \text{Multinomial}(\phi_{y_i})$

$$p_{\text{mult}}(\boldsymbol{x};\phi) = \frac{\left(\sum_{j} x_{j}\right)!}{\prod_{j} x_{j}!} \prod_{j} \phi_{j}^{x_{j}}$$

Prediction with Naïve Bayes

```
Score(x,y) := \log P(\mathbf{x}, y; \phi, \mu)

= \log P(\mathbf{x}|y; \phi)P(y; \mu)

= \log P(\mathbf{x}|y; \phi) + \log P(y; \mu)
```

Prediction with Naïve Bayes

Score(x,y)
$$:= \log P(\mathbf{x}, y; \phi, \mu)$$

 $= \log P(\mathbf{x}|y; \phi)P(y; \mu)$
 $= \log P(\mathbf{x}|y; \phi) + \log P(y; \mu)$
 $= \log \text{Multinomial}(\mathbf{x}; \phi_y) + \log \text{Cat}(y; \mu)$
 $= \log \frac{(\sum_n x_n)!}{\prod_n x_n!} + \log \prod_n \phi_{y,n}^{x_n} + \log \mu_y$
 $\propto \sum_n x_n \log \phi_{y,n} + \log \mu_y$
 $= \theta^T \mathbf{f}(\mathbf{x}, y)$
where
 $\theta = [\log \phi_1^T, \log \mu_1, \log \phi_2^T, \log \mu_2, \dots]^T$
 $\mathbf{f}(\mathbf{x}, y) = [\mathbf{0}, \dots, \mathbf{0}, \mathbf{x}^T, 1, \mathbf{0}, \dots, \mathbf{0}]^T$

Parameter Estimation

- "count and normalize"
- Parameters of a multinomial distribution

$$\phi_{y,j} = \frac{\sum_{i:Y_i = y} x_{i,j}}{\sum_{j'} \sum_{i:Y_i = y} x_{i,j'}} = \frac{\text{count}(y,j)}{\sum_{j'} \text{count}(y,j')}$$

- Relative frequency estimator
- Formally: this is the maximum likelihood estimate
 - See CIML for derivation

Smoothing (add alpha / Laplace)

$$\phi_{y,j} = \frac{\alpha + \sum_{i:Y_i = y} x_{i,j}}{\sum_{j'=1}^{V} \left(\alpha + \sum_{i:Y_i = y} x_{i,j'}\right)} = \frac{\alpha + \operatorname{count}(y,j)}{V\alpha + \sum_{j'=1}^{V} \operatorname{count}(y,j')}$$

Naïve Bayes recap

- Define p(x, y) via a generative model
- Prediction: $\hat{y} = \arg \max_{y} p(\boldsymbol{x}_i, y)$
- Learning:

$$\theta = \arg \max_{\theta} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

$$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i} p(\boldsymbol{x}_{i}, y_{i}; \boldsymbol{\theta}) = \prod_{i} p(\boldsymbol{x}_{i} | y_{i}) p(y_{i})$$

$$\phi_{y,j} = \frac{\sum_{i:Y_{i}=y} x_{ij}}{\sum_{i:Y_{i}=y} \sum_{j} x_{ij}}$$

$$\mu_{y} = \frac{\operatorname{count}(Y = y)}{N}$$

This gives the maximum likelihood estimator (MLE; same as relative frequency estimator)

Today

- Text classification problems
 - and their evaluation
- Linear classifiers
 - Features & Weights
 - Bag of words
 - Naïve Bayes