



COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Classification, Linear Models, Naïve Bayes

CMSC 470

Marine Carpuat

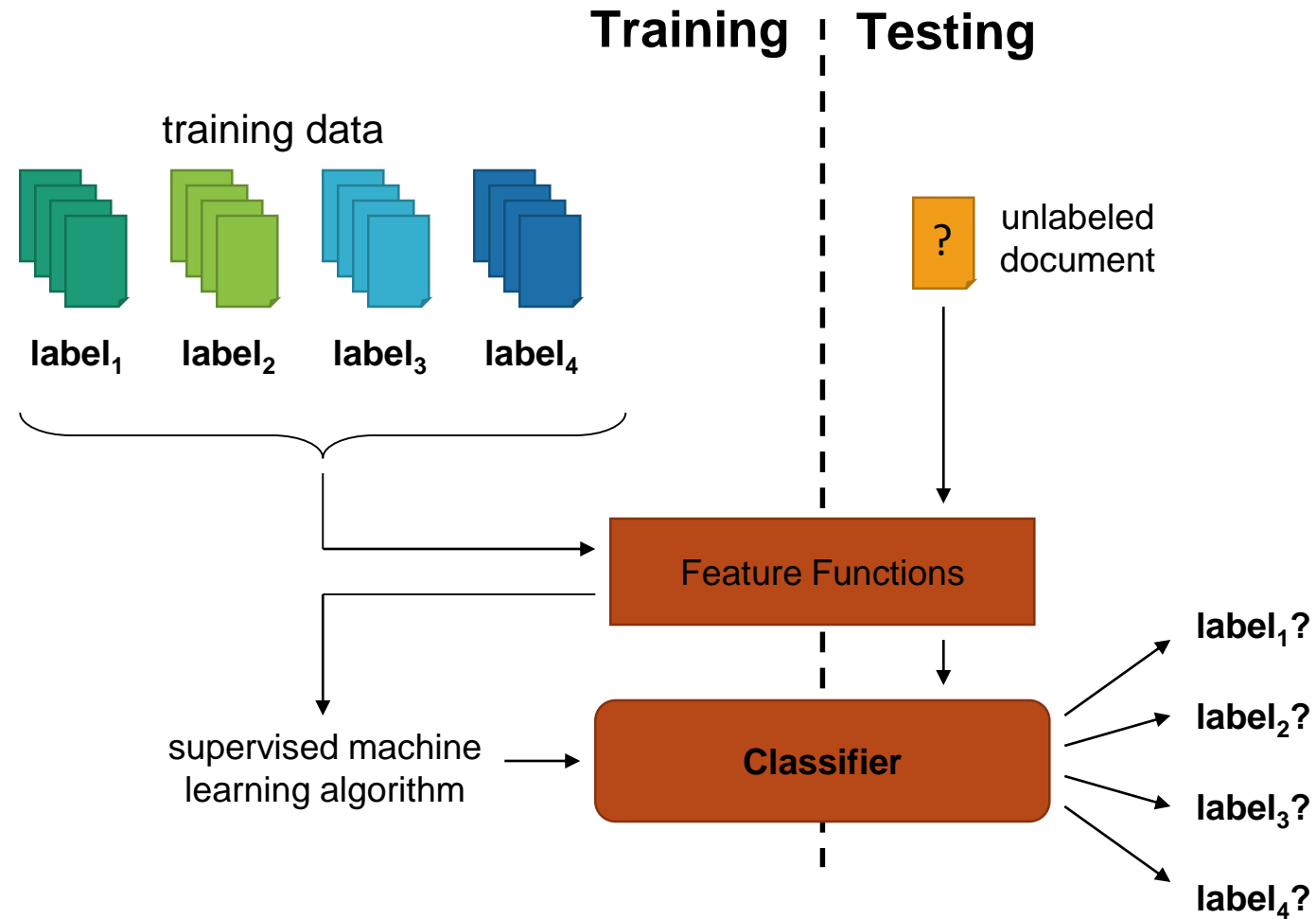
Slides credit: Dan Jurafsky & James Martin,
Jacob Eisenstein

Today

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

Classification problems

Multiclass Classification



Is this spam?

From: "Fabian Starr"
<Patrick_Freeman@pamietaniepeerelu.pl>
Subject: Hey! Software 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 software is instantly available to download - No
Need Wait!

What is the subject of this article?

MEDLINE Article



Available on the association's website

SCIENCE DIRECT

Brain & Cognition

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

Russina Gahl,^a Lisa Munn,^b Carl Rosenberger,^b David S. Jansky,^b Elizabeth Elder,^a Molly Kavage,^a and L. Hatfield Aulady^a

^a University of Maryland, Baltimore, MD, USA
^b University of Maryland, College Park, USA
Received 1 July 2012

Abstract

The study investigated how factors that have been argued to define 'grammatical form' in aphasia comprehension systems interact to influence verb and frequency of usage. We first examined the extent that comprehension ability was related to different analyses of form in passive sentences. Using a plausibility judgement task, we show that a broad group of aphasic comprehenders plausibly judge to be more likely than on passives. We then look at the characteristics of passives that are generally harder than active for aphasia. We show that this effect is modulated by word class, i.e., the likelihood that a verb appears in a given syntactic context. Results of word class verbs were significantly better than passives of the same class. Most generally, we show that active sentences receive the benefit of the match of morphological markers that are present in both active and passive that is not shown. These findings suggest that 'grammatical form' reflects frequency and lexical bias.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

The concept of 'grammatical form' or 'grammatical word order' for normal and aphasic comprehension has often been taken as synonymous to the analysis of morphosyntactic structure. However, as has been pointed out by Munn (2007), the morphological markers of 'grammatical form' have different functions. Different definitions of 'grammatical form' yield different predictions. One approach to the definition of grammatical form is that proposed by Bates, Franck, and Trulova (2007) (see also Bates et al., 2008) that associates with grammatical form the presence of English word order: present, the presence of English word order. A second approach is based on syntactic 'movement' analysis and defines as 'grammatical form' any word order that deviates from the (S)VO (or (S)O)V configuration assumed for the deep structure of English sentences. Based on the latter meaning of grammatical form, King (1988) argues that word order with grammatical verbs is crucial for aphasic comprehension for active sentences, in particular for passives with 'Agreement-marked' for reasons that are analogous to the reasons given for the greater difficulty of passives compared to actives. Although the precise definition of grammatical form is essential (see e.g., Bates & Rosenberger, 2009), grammatical forms are generally understood to be morphosyntactic structure. Examples of grammatical forms include tense, the infinitive form, clause and transformational analysis assumed in King (1988), the surface subjects of main clauses, and the subjects of embedded clauses in deep structure. Unambiguous examples include the very same difficulties as shown in sentences, according to King's analysis, and should be as hard as sentences for aphasics.

A different approach to grammatical form has been proposed by Bates et al. (2008) who suggest that grammatical form refers to the morphosyntactic structure for a given verb. Under this view, active sentences with preceding and understanding passives derive from the fact that, for these sentences, passives occur less frequently than actives. One restriction of this approach was advanced by Bates (2012), that grammatical difficulty should vary with the relative frequency of the active

MeSH Subject Category Hierarchy

- Antagonists 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 $Y = \{y_1, y_2, \dots, y_J\}$
- *Output:* a predicted class $y \in Y$

Classification Methods: Supervised Machine Learning

- *Input*

- a document d
- a fixed set of classes $Y = \{y_1, y_2, \dots, y_j\}$
- 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):

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

- People usually use balanced F1 measure
 - i.e., with $\beta = 1$ (that is, $\alpha = \frac{1}{2}$):

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

Linear Models for Multiclass Classification

Linear Models for Classification

$$\hat{y} = \arg \max_y \boldsymbol{\theta}^T \mathbf{f}(\mathbf{x}, y)$$

Feature
function
representation

Weights

Defining features: Bag of words



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

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

	<i>aardvark</i>	<i>abacus</i>	...	<i>behind</i>	...	<i>buford</i>	...	<i>clouds</i>	...	<i>great</i>	...	<i>ugly</i>	.
$\mathbf{x}_1^T =$	0	0	0...0	1	0...0	0	0...0	0	0...0	1	0...0	0	0
$\mathbf{x}_2^T =$	0	0	0...0	0	0...0	1	0...0	0	0...0	0	0...0	1	0

$$\mathbf{x}_1 = \{\text{great : 1, sunset : 1, tonight : 1, ...}\}$$

$$\mathbf{x}_2 = \{\text{ugly : 1, skies : 1, buford : 1, ...}\}$$

Defining features

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

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

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

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

Linear Classification

We can then define **weights** for each feature:

$$\boldsymbol{\theta} = \{ \langle \text{great}, \text{pos} \rangle = 1, \langle \text{great}, \text{neg} \rangle = -1, \langle \text{great}, \text{neut} \rangle = 0, \\ \langle \text{ugly}, \text{pos} \rangle = -1, \langle \text{ugly}, \text{neg} \rangle = 1, \langle \text{ugly}, \text{neut} \rangle = 0, \\ \langle \text{buford}, \text{pos} \rangle = 0, \langle \text{buford}, \text{neg} \rangle = 0, \langle \text{buford}, \text{neut} \rangle = 0, \\ \dots \}$$

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:

$$\begin{aligned} \psi_{y,\mathbf{x}} &= \sum_n \theta_n f_n(\mathbf{x}, y) \\ &= \boldsymbol{\theta}^\top \mathbf{f}(\mathbf{x}, y) \\ \hat{y} &= \arg \max_y \boldsymbol{\theta}^\top \mathbf{f}(\mathbf{x}, y) \end{aligned}$$

Linear Models for Classification

$$\hat{y} = \arg \max_y \boldsymbol{\theta}^T \mathbf{f}(\mathbf{x}, y)$$

Feature
function
representation

Weights

How can we learn weights?

- By hand
- Probability
 - e.g., Naïve Bayes
- Discriminative training
 - e.g., perceptron, support vector machines

Naïve Bayes Models for Text Classification

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 $\mathbf{x}_i \sim \text{Multinomial}(\phi_{y_i})$.

$$P_{\text{mult}}(\mathbf{x}; \phi) = \frac{(\sum_j x_j)!}{\prod_j x_j!} \prod_j \phi_j^{x_j}$$

Prediction with Naïve Bayes

$$\begin{aligned}\text{Score}(\mathbf{x}, y) &:= \log P(\mathbf{x}, y; \phi, \mu) \\ &= \log P(\mathbf{x}|y; \phi)P(y; \mu) \\ &= \log P(\mathbf{x}|y; \phi) + \log P(y; \mu)\end{aligned}$$

Definition of conditional probability

Generative story assumptions

This is a linear model!

Prediction with Naïve Bayes

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

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

Definition of conditional probability

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

$$= \log \text{Multinomial}(\mathbf{x}; \phi_y) + \log \text{Cat}(y; \mu)$$

Generative story assumptions

$$= \log \frac{(\sum_n x_n)!}{\prod_n x_n!} + \log \prod_n \phi_{y,n}^{x_n} + \log \mu_y$$

This is a linear model!

Prediction with Naïve Bayes

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

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

Definition of conditional probability

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

$$= \log \text{Multinomial}(\mathbf{x}; \phi_y) + \log \text{Cat}(y; \mu)$$

Generative story assumptions

$$= \log \frac{(\sum_n x_n)!}{\prod_n x_n!} + \log \prod_n \phi_{y,n}^{x_n} + \log \mu_y$$

$$\propto \sum_n x_n \log \phi_{y,n} + \log \mu_y$$

This is a linear model!

$$= \boldsymbol{\theta}^\top \mathbf{f}(\mathbf{x}, y)$$

where

$$\boldsymbol{\theta} = [\log \phi_1^\top, \log \mu_1, \log \phi_2^\top, \log \mu_2, \dots]^\top$$

$$\mathbf{f}(\mathbf{x}, y) = [\mathbf{0}, \dots, \mathbf{0}, \mathbf{x}^\top, 1, \mathbf{0}, \dots, \mathbf{0}]^\top$$

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)

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

Naïve Bayes recap

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

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

$$p(\mathbf{x}, \mathbf{y}; \theta) = \prod_i p(\mathbf{x}_i, y_i; \theta) = \prod_i p(\mathbf{x}_i | y_i) p(y_i)$$

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

$$\mu_y = \frac{\text{count}(Y = y)}{N}$$

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

Why is this model called “Naïve Bayes”?
Another view of the same model

$$\begin{aligned}\hat{y} &= \operatorname{argmax}_y P(Y = y | X = x) \\ &= \operatorname{argmax}_y P(Y = y) P(X = x | Y = y) \\ &= \operatorname{argmax}_y P(Y = y) \prod_{i=1}^d P(X_i = x_i | Y = y)\end{aligned}$$

Bayes rule

+ Conditional independence assumption

Today

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