Processamento de Linguagem Natural

Classificação de Texto e Naive Bayes

Prof.: Hansenclever Bassani (Hans) hfb@cin.ufpe.br
Site da disciplina: www.cin.ufpe.br/~hfb/pln/

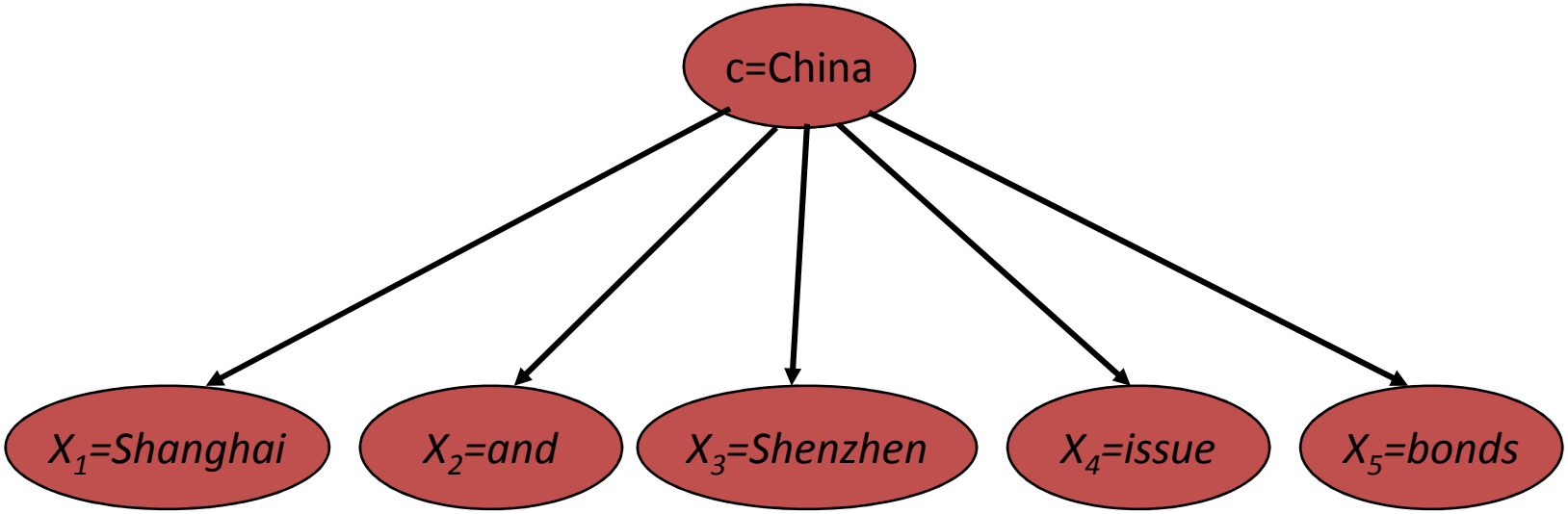
Baseado nos slides do curso de Stanford no Coursera
por Daniel Jurafsky e Christopher Manning.

Tradução: Ygor Sousa
Revisão: Hansenclever Bassani

Classificação de Texto e Naive Bayes

Naive Bayes: Relacionamento com Modelagem de Linguagens

Modelo Generativo para Naive Bayes Multinomial



Naive Bayes e Modelagem de Linguagens

- Classificadores Naive bayes podem usar qualquer tipo de característica
  - URL, endereço de mail, dicionários, características de rede
- Mas e se, como nos slides anteriores
  - Nós usarmos **apenas** características de palavras
  - Nós usamos **todas** as palavras em um texto (não um subgrupo)
- Então
  - Naive bayes tem uma similaridade importante com a modelagem de linguagens.

Cada classe = um modelo de linguagem unigram

- Atribui a cada palavra: P(word | c)
- Atribui a cada sentença: P(s|c)= Π P(word|c)

Class *pos*

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

I	love	this	fun	film
0.1	0.1	.05	0.01	0.1

P(s | pos) = 0.0000005

Naive Bayes como um Modelo de Linguagem

- Que classe atribui a maior probabilidade a *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

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$$P(s|pos) > P(s|neg)$$

Classificação de Texto e Naive Bayes

Naive Bayes: Relacionamento com Modelagem de Linguagens

Classificação de Texto e Naive Bayes

Naive Bayes Multinomial: Um Exemplo

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priori:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Probabilidades Condicionais:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

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

Escolha de uma classe:

P(c|d5) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14

≈ 0.0003

P(j|d5) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9

≈ 0.0001

Naive Bayes em Filtragem de Spam

- Características “matadoras” para identificar Spam:
  - Menções a Viagra genérico
  - Farmácia Online
  - Menções a milhões de (dólar) ((dólar) NN,NNN,NNN.NN)
  - Frase: impressionar ... garota
  - From: começa com muitos números
  - Assunto está todo em maiúsculo
  - HTML tem uma baixa proporção de texto por área de imagem
  - 100% garantido
  - Afirma que você pode ser removido de uma lista
  - http://spamassassin.apache.org/tests_3_3_x.html

Resumo: Naive Bayes não é tão "Naive"

- Muito rápido, baixos requisitos de armazenamento
- Robusto para características irrelevantes
  - Características irrelevantes cancelam uma a outra sem afetar muito os resultados
- Muito bom em domínios com muitas características igualmente importantes
  - Árvores de Decisão sofrem de *fragmentação* em alguns casos – especialmente com poucos dados
- Ótimo se as suposições de independência são válidas: Se a independência assumida estiver correta, então o classificador de Bayes é ótimo pro problema
- Um confiável ponto de partida para classificação de texto
  - **Mas veremos outros classificadores melhores, de maior precisão**

Classificação de Texto e Naive Bayes

Naive Bayes Multinomial : Um Exemplo

Classificação de Texto e Naive Bayes

Precisão, Recall e medida F

Tabela de contingência 2x2

	Correto	Incorreto
Selecionado	*tp*	*fp*
Não Selecionado	*fn*	*tn*

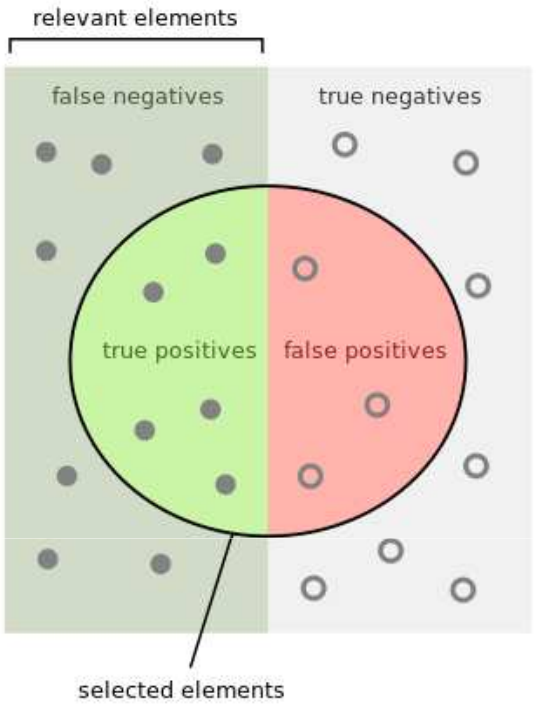
Precisão e Recall

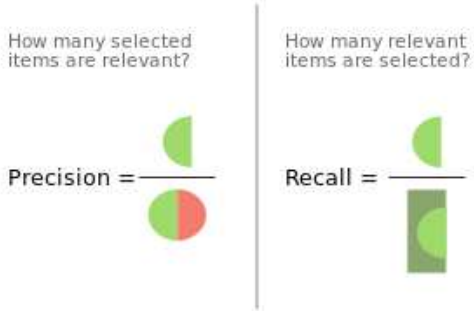
- **Precisão**: % de itens selecionados que estão corretos
  Recall: % de itens corretos que estão selecionados

	Correto	Incorreto
Selecionado	*tp*	*fp*
Não Selecionado	*fn*	*tn*

$$\text{Precision}: \quad P = \frac{tp}{(tp + fp)}$$

$$\text{Recall}: \quad R = \frac{tp}{(tp + fn)}$$





Uma medida combinada: F

- Uma medida combinada que avalia o "P/R tradeoff" é a medida F (média harmônica ponderada):

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

- A medida harmônica é uma média bastante conversadora; Ver em *"Introduction to Information Retrieval"* § 8.3
- As pessoas costumam usar a medida F1 balanceada
  - i.e., com $\beta = 1$ (isto é, $\alpha = ½$): $F = 2PR/(P+R)$

Classificação de Texto e Naive Bayes

Precisão, Recall e medida F

Classificação de Texto e Naive Bayes

Classificação de Texto: Avaliação

Mais que duas classes: Conjuntos de classificadores binários

- Lidando com classificação any-of ou multivalue
  - Um documento pode pertencer a 0, 1, ou >1 classes.
- Para cada classe $c \in C$
  - Construa um classificador γ_c para distinguir c de todas as outras classes $c' \in C$ (um reconhecedor para cada classe)
- Dado documento de teste d,
  - Avaliar sua adesão a cada classe usando cada γ_c
  - d pertence a qualquer classe em que γ_c retorne verdadeiro

Mais que duas classes: Conjuntos de classificadores binários

- Classificação One-of ou multinomial
  - Classes são mutuamente exclusivas: cada documento em exatamente uma classe
- Para cada classe $c \in C$
  - Construa um classificador γ_c para distinguir c de todas outras classes $c' \in C$
- Dado documento de teste d,
  - Avaliar sua adesão a cada classe usando cada γ_c
  - d pertence a classe com o score máximo

Avaliação: Conjunto de dados Classic Reuters-21578

- Conjunto de dados mais usado, 21.578 docs (cada um com 90 tipos, 200 tokens)
- 9603 treinamento, 3299 artigos de teste (ModApte/Lewis split)
- 118 categorias
  - Um artigo pode estar em mais de uma categoria
  - Aprender 118 distinções de categorias binárias
- Documento médio (com ao menos uma categoria) tem 1.24 classes
- Apenas cerca de 10 das 118 categorias são grandes

Categorias comuns (#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Categorização de Texto de Documento do Conjunto de dados Reuters (Reuters-21578)

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Matriz de Confusão C

- Para cada par de classes <c_1,c_2> quantos documentos de c_1 foram incorretamente atribuídos a c_2?
  - $c_{3,2}$: 90 documentos “wheat” atribuídos incorretamente a “poultry”

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Medidas de Avaliação por Classe

Recall:
Fração de documentos na classe *i* classificados corretamente:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precisão:
Fração de documentos atribuídos a classe *i* que são realmente da classe *i*:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Acurácia: (1 – taxa de erro)
Fração de documentos classificados corretamente:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Média: Micro- vs. Macro

- Se tivermos mais de uma classe, como podemos combinar várias medidas de desempenho em um valor único?
- **Média Macro**: Calcule o desempenho para cada classe, em seguida, média.
- **Média Micro**: Recolha decisões para todas as classes, calcule a tabela de contingência, avalie.

Micro- vs. Macro: Exemplo

Classe 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Classe 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

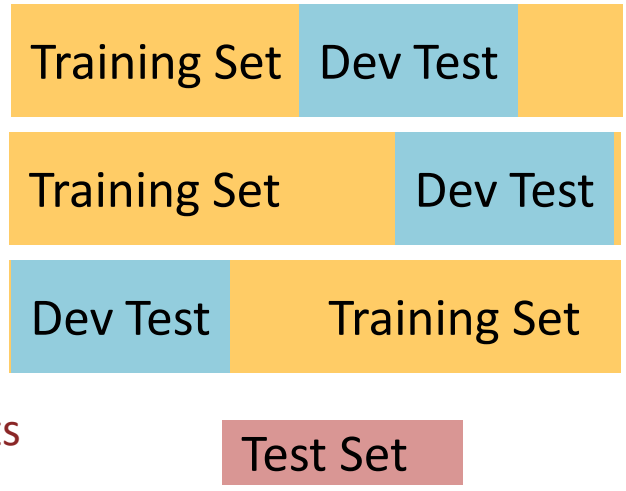
Tabela da Média Micro

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Precisão da Média Macro: (0.5 + 0.9)/2 = 0.7
- Precisão da Média Micro: 100/120 = .83
- Score da Média Micro é dominado por scores em classes comuns

Conjuntos de Teste de Desenvolvimento e Validação Cruzada

- Métrica: P/R/F1 ou Acurácia
- Conjunto de teste não visto
  - Evitar *overfitting* (‘ajuste ao conjunto visto’)
  - Estimativa mais conservadora de desempenho
- Validação cruzada em múltiplas divisões
  - Lida com erros de amostragem de diferentes datasets
  - Resultados calculados para cada divisão
  - Calcula a performance do conjunto de desenvolvimento pelo resultado das divisões



Classificação de Texto e Naive Bayes

Classificação de Texto: Avaliação

Classificação de Texto e Naive Bayes

Classificação de Texto: Questões Práticas

No Mundo Real

- Cara, quero criar um classificador de texto!
- O que eu deveria fazer?

Sem dados de treinamento? Regras escritas manualmente

If (wheat or grain) and not (whole or bread) then
    Categorize as grain

- Precisa de elaboração cuidadosa
  - Ajuste humano de dados de treinamento
  - Consumo de tempo: 2 dias por classe

Pouquíssimos dados?

- Use Naive Bayes
  - Naive Bayes é um algoritmo "high-bias" (Ng and Jordan 2002 NIPS)
- Obter mais dados rotulados
  - Encontre maneiras espertas para fazer humanos rotularem dados para você
- Tente métodos de treinamento semi-supervisionados:
  - Bootstrapping, EM em documentos não rotulados, …

Mais em: https://en.wikipedia.org/wiki/Bias-variance_tradeoff

Uma quantidade razoável de dados?

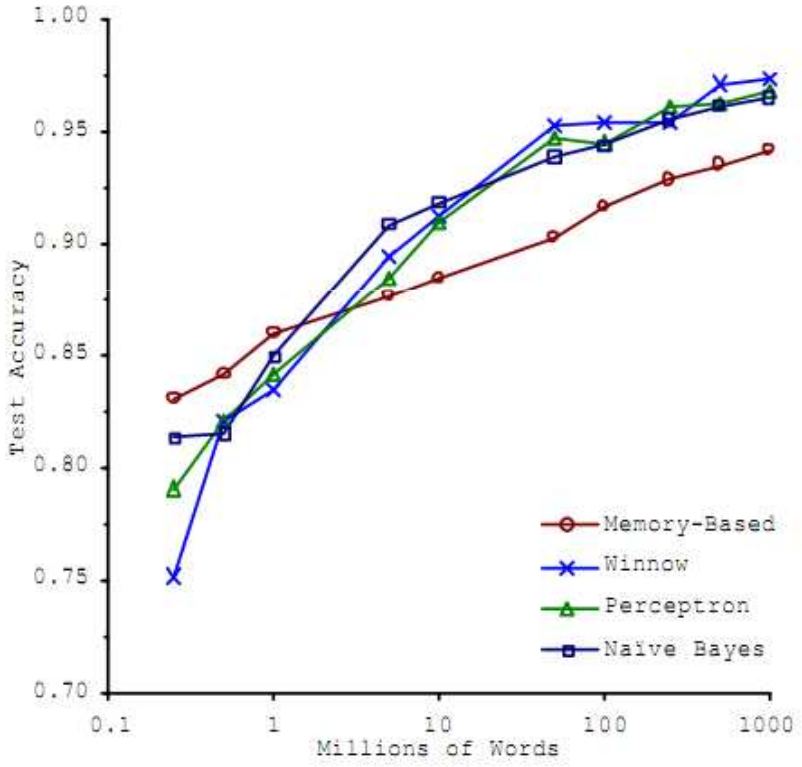
- Perfeito para todos os classificadores inteligentes
  - SVM, KNN, etc
  - Regularized Logistic Regression
- Você ainda pode usar árvores de decisão interpretáveis pelo usuário
  - Usuários gostam de “hackear”
  - Gestão gosta de correções rápidas

Uma grande quantidade de dados?

- Pode alcançar alta acurácia!
- Com um custo:
  - SVMs (tempo de treinamento) ou kNN (tempo de teste) pode ser bem lento
  - Regularized logistic regression pode ser um pouco melhor
- Então Naive Bayes pode ser utilizado novamente!

Acurácia como uma função do tamanho dos dados

- Com dados suficientes
  - Classificador pode não importar



Brill and Banko on spelling correction

Sistemas do mundo real geralmente combinam

- Classificação Automática
- Revisão manual de casos incertos/difíceis/"novos”

Prevenção de Underflow: log space

- Multiplicar várias probabilidades pode resultar em underflow de ponto flutuante.
- Dado que log(*xy*) = log(*x*) + log(*y*)
  - Melhor somar os logs das probabilidades do que as multiplicar.
- A classe com maior score de probabilidade de log não normalizado ainda é a mais provável.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i \mid c_j)$$

- Modelo é agora o máximo da soma dos pesos

Como refinar o desempenho

- Características de domínio específico e pesos: *muito* importante em desempenho real
- As vezes é necessário colapsar termos:
  - Números, formulas químicas, …
  - Mas geralmente não ajuda
- *Upweighting*: Contar uma palavra como se ela tivesse ocorrido duas vezes
  - Palavras de título (Cohen & Singer 1996)
  - Primeira sentença de cada parágrafo (Murata, 1999)
  - Em sentenças que contém palavras de título (Ko *et al,* 2002)

Classificação de Texto e Naive Bayes

Classificação de Texto: Questões Práticas