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

A Tarefa de Classificação de Texto

É spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

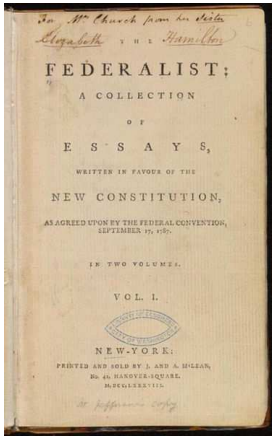
Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

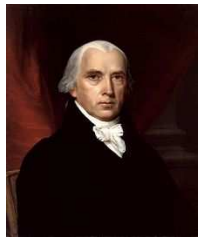
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Quem escreveu cada documento federal?


FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
IN TWO VOLUMES.
VOL. I.
NEW-YORK:

- 1787-8: trabalhos anônimos tentaram convencer Nova York a retificar a Constituição dos U.S: Jay, Madison, Hamilton.
- Autoria de 12 das cartas em disputa
- 1963: resolvido por Mosteller e Wallace utilizando métodos bayesianos


James Madison


Alexander Hamilton

Autor masculino ou feminino?

1. “Dizem que a vida é para quem sabe viver, mas ninguém nasce pronto. A vida é para quem é corajoso o suficiente para se arriscar e humilde o bastante para aprender.”

2. “Presente, passado e futuro? Tolice. Não existem. A vida é uma ponte interminável. Vai-se construindo e destruindo. O que vai ficando para trás com o passado é a morte. O que está vivo vai adiante.”

1 – Clarice Lispector; 2 – Darcy Ribeiro

Crítica de filme positiva ou negativa?

- Inacreditavelmente desapontador

- Cheio de personagens malucos em uma sátira ricamente aplicada e algumas ótimas reviravoltas na história

- Foi a melhor comédia excêntrica já filmada

- Foi patético. A pior parte foram as cenas de boxe.

Qual o assunto do artigo?

Artigo da MEDLINE

?

Hierarquia de Categorias de Assuntos MeSH

- Antagonistas e inibidores
- Fornecimento de Sangue
- Química
- Terapia Medicamentosa
- Embriologia
- Epidemiologia
- ...

Classificação de Texto

- Atribuição de categoria de assuntos, tópicos, ou gêneros
- Detecção de Spam
- Identificação de Autoria
- Identificação de Idade/Gênero
- Identificação de Linguagem
- Análise de Sentimento
- …

Classificação de Texto: Definição

- Entrada:
  - um documento d
  - um grupo fixo de classes $C = \{c_1, c_2, \ldots, c_J\}$

- Saída: uma classe prevista $c \in C$

Métodos de Classificação: Regras codificadas a mão

- Regras baseadas na combinação de palavras e outras características
  - spam: endereços-lista-negra OR (“dollars” AND “have been selected”)
- Acurácia pode ser alta
  - Se as regras forem cuidadosamente refinadas pelo expert
- Porém construir e manter essas regras é caro

Métodos de Classificação: Supervised Machine Learning

- Entrada:
  - Um documento *d*
  - Um grupo fixo de classes $C = \{c_1, c_2, \ldots, c_J\}$
  - Um conjunto de treinamento de *m* documentos rotulados manualmente $(d_1,c_1),\ldots,(d_m,c_m)$
- Saída:
  - Um classificador treinado $\gamma: d \rightarrow c$

Métodos de Classificação: Supervised Machine Learning

- Qualquer tipo de classificador
  - Naive Bayes
  - Regressão Logística
  - Support-vector machines
  - k-NN
  - …

Classificação de Texto e Naive Bayes

A Tarefa de Classificação de Texto

Classificação de Texto e Naive Bayes

Naive Bayes(I)

Naive Bayes

- Simples método de classificação (“naive”) baseado na regra de Bayes
- Conta com uma representação muito simples de documento
  - Bag-of-Words – BoW (bolsa de palavras)

A representação BoW
Exemplo em Análise de Sentimento

Y(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.)=C

A representação BoW

Y(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.)=C

A representação BoW:
utilizando um subconjunto de palavras

Y(x **love** xxxxxxxxxxxxxxx **sweet**
xxxxxxx **satirical** xxxxxxxxxx
xxxxxxxxxxx **great** xxxxxxx
xxxxxxxxxxxxxxxxxx **fun** xxxx
xxxxxxxxxxxxx **whimsical** xxxx
romantic xxxx **laughing**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx **recommend** xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx **several** xxxxxxxxxxxxxxxx
xxxxx **happy** xxxxxxxxx **again**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxx)=C

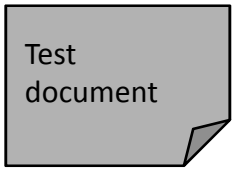
Representação BoW

Y(

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

)=C

BoW para classificação de documentos



parser
language
label
translation
…

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	…
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network…	language…	region…	language…	

Classificação de Texto e Naive Bayes

Naive Bayes(I)

Classificação de Texto e Naive Bayes

Formalização do Classificador Naive Bayes

Regra de Bayes Aplicada a Documentos e Classes

- Para um documento *d* e uma classe *c*

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

Classificador Naive Bayes (I)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP é "valor máximo a posteriori" = classe mais provável

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Regra de Bayes

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Descartando o denominador

Classificador Naive Bayes (II)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

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

Documento d representado como $x_1..x_n$ características

Classificador Naive Bayes (IV)

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

O($|X|^n \bullet |C|$) parâmetros

Só poderia ser estimado, se um número muito, muito grande de exemplos de treinamento estiver disponível

Com que frequência essa classe ocorre?

Nós podemos apenas contar as frequências relativas em uma coleção

Premissas de Independência de Naive Bayes Multinomiais

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

- **Premissa da BoW**: Assume que posição não importa
- **Independência Condicional**: Assume que as probabilidades das características $P(x_i \mid c_j)$ são independentes dada a classe *c.*

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

Classificador Naive Bayes Multinomial

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

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

Aplicação de Classificadores Naive Bayes Multinomiais à Classificação de Texto

positions ← todas as posições das palavras no documento de teste

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

Classificação de Texto e Naive Bayes

Formalização do Classificador Naive Bayes

Classificação de Texto e Naive Bayes

Naive Bayes: Aprendizagem

Aprendendo o Modelo Naive Bayes Multinomial

- Primeira tentativa: estimativa de máxima verossimilhança
  - Simplesmente use a frequência nos dados

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

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

Estimação de Parâmetro

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

Quantidade de vezes que a palavra w_i aparece entre todas as palavras nos documentos do tópico c_j

- Cria um mega documento para o tópico *j* concatenando todos os documentos neste tópico
  - Usa a frequência de *w* no mega documento

Problema com Máxima Verossimilhança

- E se nós não vimos nenhum documento de treinamento com a palavra ***fantastic*** classificado no tópico **positive?**

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

- Resultará em produto zero!

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

Suavização Laplace (add-1) para Naive Bayes

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

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

Naive Bayes Multinomial: Aprendizagem

- Do conjunto de treinamento, extrair *Vocabulary*
- Calcule os termos $P(c_j)$
  - Para cada c_j em C faça

    $docs_j \leftarrow$ todos documentos com classe $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calcule os termos $P(w_k \mid c_j)$
  - $Text_j \leftarrow$ documento único contendo todos $docs_j$
  - Para cada palavra w_k em *Vocabulary*

    $n_k \leftarrow$ # de ocorrências de w_k no $Text_j$

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

Suavização Laplace (add-1): palavras desconhecidas

- Adicionar uma palavra extra ao vocabulário, a w_u "unknown word"

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

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

Classificação de Texto e Naive Bayes

Naive Bayes: Aprendizagem