Natural Language Processing

Assist. Prof. Dr. Tuğba YILDIZ

İSTANBUL BİLGİ UNIVERSITY
Department of Computer Engineering

April 21, 2017

1 The Task of Text Classification

Task of Text Classification

- Is it 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.

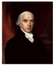
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

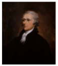
Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.

Task of Text Classification

- Who wrote Federalistpapers?
  - 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
  - Authorship of 12 of the letters in dispute
  - 1963: solved by Mosteller and Wallace using Bayesian methods


James Madison


Alexander Hamilton

Task of Text Classification

- Male or Female?

  1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
  2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

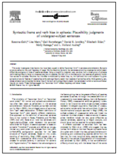
Task of Text Classification

- Positive or Negative Movie Review
  - unbelievably disappointing
  - Full of zany characters and richly applied satire, and some great plot twists
  - this is the greatest screwball comedy ever filmed
  - It was pathetic. The worst part about it was the boxing scenes.

Task of Text Classification

- What is the subject of paper?

MEDLINE Article



?

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: C={c1,c2,c3,...cj}
- Output: a predicated class c $\in$ C

Classification Methods: Supervised Machine Learning

- Input:
  - a document : d
  - a fixed set of classes: C={c1,c2,c3,...cj}
  - a training set of m hand-labeled documents (d1,c1)...(dm,cm)
- Output: a learned classifier y:d→ c

Classification Methods: Supervised Machine Learning

- Naive Bayes
- Linear Regression
- SVM
- k-nn

Naive Bayes Algorithm

- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature
- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- Naive Bayes model is easy to build and particularly useful for very large data sets.
- Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Naive Bayes Algorithm

- Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).
- Look at the equation below:

Likelihood   Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability   Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Naive Bayes Algorithm

- Naive Bayes Classifier

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

Naive Bayes Algorithm

- Naive Bayes Classifier

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

MAP is "maximum a posteriori" = most likely class

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

Bayes Rule

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

Dropping the denominator

Naive Bayes Algorithm

- Naive Bayes Classifier

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

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

Document d represented as features x1..xn

Naive Bayes Algorithm

- Naive Bayes Classifier

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

O(|*X*|^n•|*C*|) parameters

How often does this class occur?

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

Naive Bayes Algorithm

- Multinominal Naive Bayes Independence Assumption

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

  - **Bag of Words assumption**: Assume position doesn't matter
  - **Conditional Independence**: Assume the feature probabilities $P(x_i \mid c_j)$ are independent given the class c.

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

Naive Bayes Algorithm

- Multinominal Naive Bayes Independence Assumption

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

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

Naive Bayes Algorithm

- Multinominal Naive Bayes Independence Assumption
  - First attempt: maximum likelihood estimates
    - simply use the frequencies in the data

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

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

Naive Bayes Algorithm

- Problem on MLE
  - What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive** (***thumbs-up***)?

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

  - Zero probabilities cannot be conditioned away, no matter the other evidence!

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

Naive Bayes Algorithm

- Laplace (Add-1) Smooting

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

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

Naive Bayes Algorithm

- Laplace (Add-1) Smooting
  - From training corpus, extract *Vocabulary*
  - Calculate $P(c_j)$ terms
    - For each c_j in C do

      $docs_j \leftarrow$ all docs with class $= c_j$

      $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$
  - Calculate $P(w_k \mid c_j)$ terms
    - $Text_j \leftarrow$ single doc containing all $docs_j$
    - For each word w_k in *Vocabulary*

      $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

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

Naive Bayes Algorithm

- Naive Bayes Algorithm - example

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Naive Bayes Algorithm

- Naive Bayes Algorithm - example
  1. Convert the data set into a frequency table
  2. Create Likelihood table by finding the probabilities
  3. Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Naive Bayes Algorithm

- Naive Bayes Algorithm - example
  1. Convert the data set into a frequency table

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Naive Bayes Algorithm

- Naive Bayes Algorithm - example
  1. Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Naive Bayes Algorithm

- Naive Bayes Algorithm - example
- Problem: Players will play if weather is sunny. Is this statement is correct?

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

- P(Yes | Sunny) = (P(Sunny | Yes) * P(Yes)) / P (Sunny)
- P (Sunny | Yes) = 3/9 = 0.33
- P(Sunny) = 5/14 = 0.36
- P(Yes)= 9/14 = 0.64
- P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes Algorithm

- Naive Bayes Algorithm - example

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Naive Bayes Algorithm

- Naive Bayes Algorithm - example

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Yeni veri

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$P(C_i \mid X) = P(X \mid C_i) \times P(C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \times P(C_i)$$

İki Sınıf için olasılık:

P("yes"|X) = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

P("no"|X) = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Normalize edilmiş olasılıklar:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Naive Bayes Algorithm

- Bayes rule applied to document and class
- For a document and a class

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Yeni veri

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$P(C_i \mid X) = P(X \mid C_i) \times P(C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \times P(C_i)$$

İki Sınıf için olasılık:

P("yes"|X) = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

P("no"|X) = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Normalize edilmiş olasılıklar:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Naive Bayes Algorithm

- Example:

Dan Jurafsky

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

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

	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	?

Priors:

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

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

Conditional Probabilities:

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

Choosing a class:

P(c|d5) $\propto$ 3/4 * (3/7)^3 * 1/14 * 1/14 $\approx$ 0.0003

P(j|d5) $\propto$ 1/4 * (2/9)^3 * 2/9 * 2/9 $\approx$ 0.0001

45

References

- Speech and Language Processing (3rd ed. draft) by D. Jurafsky & J. H. Martin (web.stanford.edu)