# Методы классификации

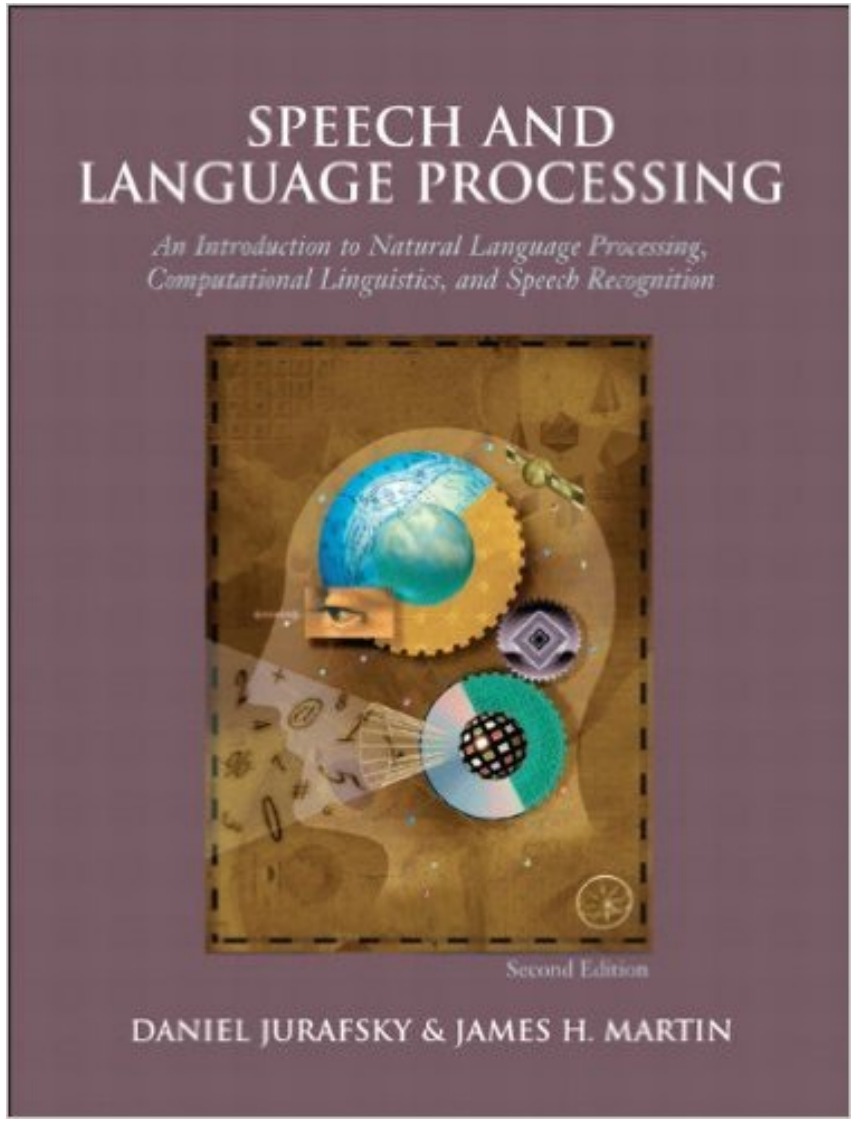
Елена Тутубалина

Казанский федеральный университет, Россия

1 марта 2018

# Литература

Speech and Language Processing,
by Daniel Jurafsky and
James H. Martin, 2008
https://web.stanford.edu/~jurafsky/slp3/6.pdf



# Содержание

# Примеры

Unable to process your most recent Payment

HSBC.co.uk (service@hsbc.co.uk) 2014/6/18 10:08


HSBC_Payme
nt_
9854711.pdf

HSBC The world's local bank

You have a new e-Message from HSBC.co.uk

This e-mail has been sent to you to inform you that we were unable to process your most recent payment.

Please check attached file for more detailed information on this transaction.

Pay To Account Number: **********12
Due Date: 18/06/2014
Amount Due: £ 077.05

IMPORTANT: The actual delivery date may vary from the Delivery by date estimate. Please make sure that there are sufficient available funds in your account to cover your payment

**Antipin Yury** @1antipin · 2 окт.

Иногда нужно зайти в сбер оплатить квитанцию налом. Но с 1 числа я этого сделать не могу. Нет карты сбера не можешь оплатить. Дно #сбербанк

2 3 2

**Oleg Pleshkov** @olegpleshkov · 2 окт.

Как дозвониться до службы поддержки корпоративных клиентов #сбербанк ? Это просто издевательство! Пора менять банк, похоже...

1 2 2

**Patrick Lancaster** @PLnewstoday · 28 сент.

Пожалуйста поддержите мою независимую журналистику пожертвованиями через #Сбербанк #России,4276520695551214 #Украина

12

**MaxD** @Xd17Ma · 2 окт.

Очень долго идет смс-сообщение от сбер-онлайн. Вы не цените время своих клиентов. Придется отказываться от Сбербанка #Сбербанк

2 2 1

**Мятный Ким|AJ Soul** @tohiro_twt · 19 мин.

В ответ @SSSOOOUUULLL

У меня тоже так как я **болею**. Это я Вам просто желаю доброго дня. А у меня далеко не доброе. Что случилось зайка?! 💕

**Наташа Семенова** @mxgzpvwxllabpz1 · 21 мин.

С 1998 года я **болею** за "Манчестер Юнайтед". Черно-белый телевизор и проволока вместо антенны. Так и смотрел, когда был студентом.

**Romanovna** ◆ @mashchenko192 · 24 мин.

боже ,как я "люблю" осень за то что **болею** и температура выше 38 ,спасибо)

3

**Арина Афонина** @afoninaarina201 · 24 мин.

Я **болею** 💻😭💜

1

**kindness_evil** @kimsoekjinlove · 28 мин.

я тут **болею**

темприч и все дела

смотрю дорамы всякие

# Зачем нам классификация?

Задачи

- Присваивание категории, темы или жанра
- Определение спама
- Определение авторства
- Определение возраста/пола автора
- Определение языка
- Анализ тональности
- ...

# Типы классификации

Типы классификации

- Binary classification (true, false)
- Multi-class classification (politics, sports, gossip)
- Multi-label classification (#party #FRIDAY #fail)
- Clustering (labels unknown)

# Методы

- Вручную (By hand)
- На правилах (Rule-based)
- Статистические (Statistical)

# Методы

- Вручную (By hand)
  - E.g. Yahoo in the old days
  - + Very accurate and consistent assuming experts
  - - Super slow, expensive, does not scale
- На правилах (Rule-based)
  - E.g. Advanced search criteria (”site:ox.ac.uk”)
  - + Accuracy high if rule is suitable
  - - Need to manually build and maintain rule-based system.
- Статистические (Statistical)
  - Лекция
  - + Scales well, can be very accurate, automatic
  - - Requires classified training data. Sometimes a lot!

# Классификация текста – формально

Формально:

- Вход:
  - документ $d$
  - фиксированный набор классов $C = \{c_1, c_2, \ldots, c_J\}$
- Выход: предсказанный класс $c \in C$
- want to learn the probability of d being of class c!

Key questions:

- How to represent d.
- How to calculate P(c|d).

# Классификация в 2 частях

Think of text classification as a two stage process

## Representation

Process text into some (fixed) representation.
How to learn d

## Classification

Classify document given that representation.
How to learn P(c|d)

# Possible Representations for Text

- Bag of Words (BOW)
  - Easy, no effort required.
  - Variable size, ignores sentential structure.
- Hand-crafted features
  - Full control, can use of NLP pipeline, class-specific features
  - Over-specific, incomplete, makes use of NLP pipeline.
- Learned feature representation
  - Can learn to contain all relevant information.
  - Needs to be learned.

# Подход к классификации: Обучение с учителем

англ. Supervised

- Вход:
  - Фиксированный набор классов $C = \{c_1, c_2, \ldots, c_J\}$
  - Обучающее множество, состоящее из $m$ вручную размеченных документов $(d_1, c_1), \ldots, (d_m, c_m)$
- Выход: функция предсказания $\gamma : d \rightarrow c$

# Алгоритмы supervised-классификации

- (Наивный байес) Naïve Bayes
- (Логистическая регрессия) Logistic regression
- (Метод опорных векторов) Support-vector machines
- (K ближайших соседей) k-Nearest Neighbors

- ...

# Генеративная модель

Классификатор:

$$\hat{y} = \operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y P(x|y)P(y)$$

- оценивается генерация $x$ из класса $y$
- правдоподобие признаков $x$ при заданном $y$
- E.g. n-gram models, hidden Markov models, probabilistic context-free grammars, IBM machine translation models, Naive Bayes, ...

# Дискриминативная модель

- вычисляем $P(y|x)$ напрямую
- проводим различия между значениями $y$
- извлекаем признаки, комбинируем их, применяем функцию
- E.g. logistic regression, maximum entropy models, conditional random fields, support-vector machines, ...

# Содержание

# Правило Байеса прим. к документам и классам

- Для документа $d$ и класса $c$

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

# Наивный байесовский классификатор (I)

The best class is the maximum a posteriori (MAP) class:
$C_{MAP} = \text{argmax}_{c \in C} P(c|d) =$

- MAP – "maximum a posteriori" – оценка апостериорного максимума (MAP)
- применяем правило Байеса
- отбрасываем знаменатель (не зависит от c)

# Наивный байесовский классификатор (II)

$$C_{MAP} = \operatorname*{argmax}_{c \in C} P(d|c)P(c) =$$

$$= \operatorname*{argmax}_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c)$$

Документ $d$ представлен признаками $x_1 \ldots x_n$

# Наивный байесовский классификатор (III)

Два типа параметров – can be estimated from labelled training data

"правдоподобие" : $P(x_1, x_2, \ldots, x_n|c)$

априорная вероятность : $P(c)$

# Наивный байесовский классификатор (IV)

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

- Может быть подсчитано, только если доступно огромное число обучающих примеров.
- Мы можем посчитать только относительные частоты в обучающем множестве.

# Предположения в модели наивного Байеса

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

Мешок слов : Позиция слова не имеет значения

Условная независимость : вероятности признаков $P(x_i|c_j)$ взаимно независимы при заданном классе.

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

# Полиномиальный наивный байесовский классификатор

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

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

# Применение к задаче классификации текста

positions $\leftarrow$ все позиции слов во входном документе

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

# Содержание

# Feature Representations

A feature representation (of text) can be viewed as a vector where each element indicates the presence or absence of a given feature in a document.

## Note

Features can be

- binary (presence/absence)
- multinomial count)
- continuous (eg. TF-IDF weighted)

# Обучение полиномиального НБК

Первая попытка: метод максимального правдоподобия

- просто считаем частоты в обучающем корпусе

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

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

# Подробнее про параметры “правдоподобия”

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

доля слова $w_i$ относительно всех слов в документах класса $c_j$

- Создаём агрегирующий (“мега”-)документ для класса $j$ конкатенацией всех документов этой темы
- Используем частоту $w$ в этом мега-документе.

# Проблема метода максимального правдоподобия

Представим задачу анализа тональности.

- Что если нет обучающего документа со словом “удивительный” и скласифицированного как позитивный?

$$\hat{P}(\text{“удивительный”} \mid \text{поз}) = \frac{count(\text{“удивительный”} \mid \text{поз})}{\sum_{w \in V} count(w, \text{поз})} = 0$$

- Подобные нулевые параметры приведут к обнулению итога, вне зависимости от других признаков!

$$C_{MAP} = \underset{c}{\operatorname{argmax}}\, \hat{P}(c) \prod_{i} \hat{P}(x_i|c)$$

# Сглаживание по Лапласу (add-1) для НБК

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum\limits_{w \in V} (count(w,c)+1)} =$$

$$= \frac{count(w_i,c)+1}{(\sum\limits_{w \in V} count(w,c))+|V|}$$

# Обучение полиномиального НБК: алгоритм

- Из обучающего корпуса извлечь список слов ("лексикон") $V$

- Вычислить параметры $P(c_j)$
  - For each $c_j$ in $C$ do $docs_j \leftarrow$ all docs with class $= c_j$

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

- Вычислить параметры $P(w_k|c_j)$
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - Foreach word $w_k$ in Vocabulary
    $n_k \leftarrow \#$ of occurrences of $w_k$ in $Text_j$
    $P(w_k|c_j) \leftarrow \frac{n_k+\alpha}{n+\alpha|Vocabulary|}$

# Summary

- Advantages
  - Simple
  - Interpretable
  - Fast (linear in size of training set and test document)
  - Text representation trivial (bag of words)
- Drawbacks
  - Independence assumptions often too strong
  - Sentence/document structure not taken into account
  - Naıve classifier has zero probabilities; smoothing is awkward

# Пример

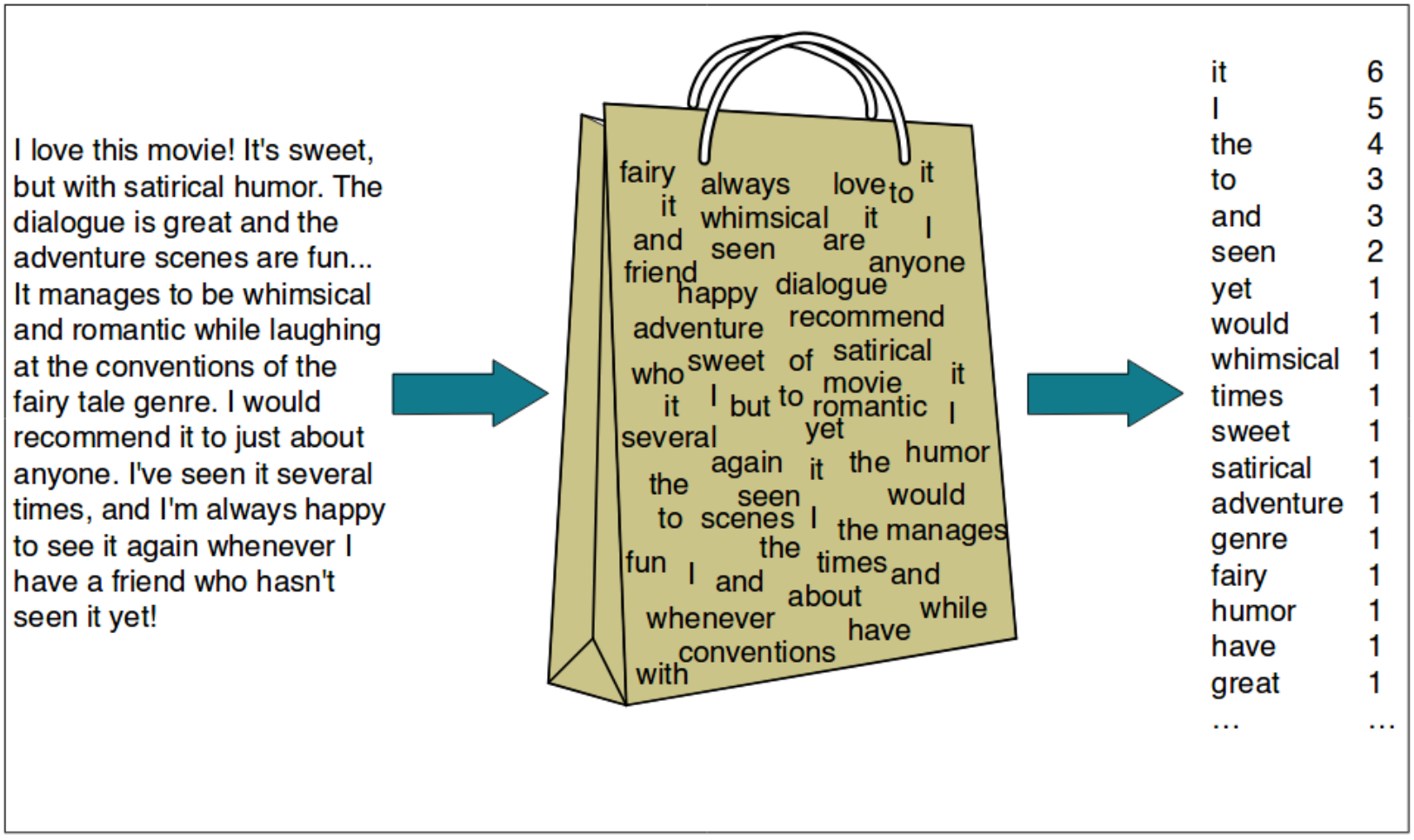
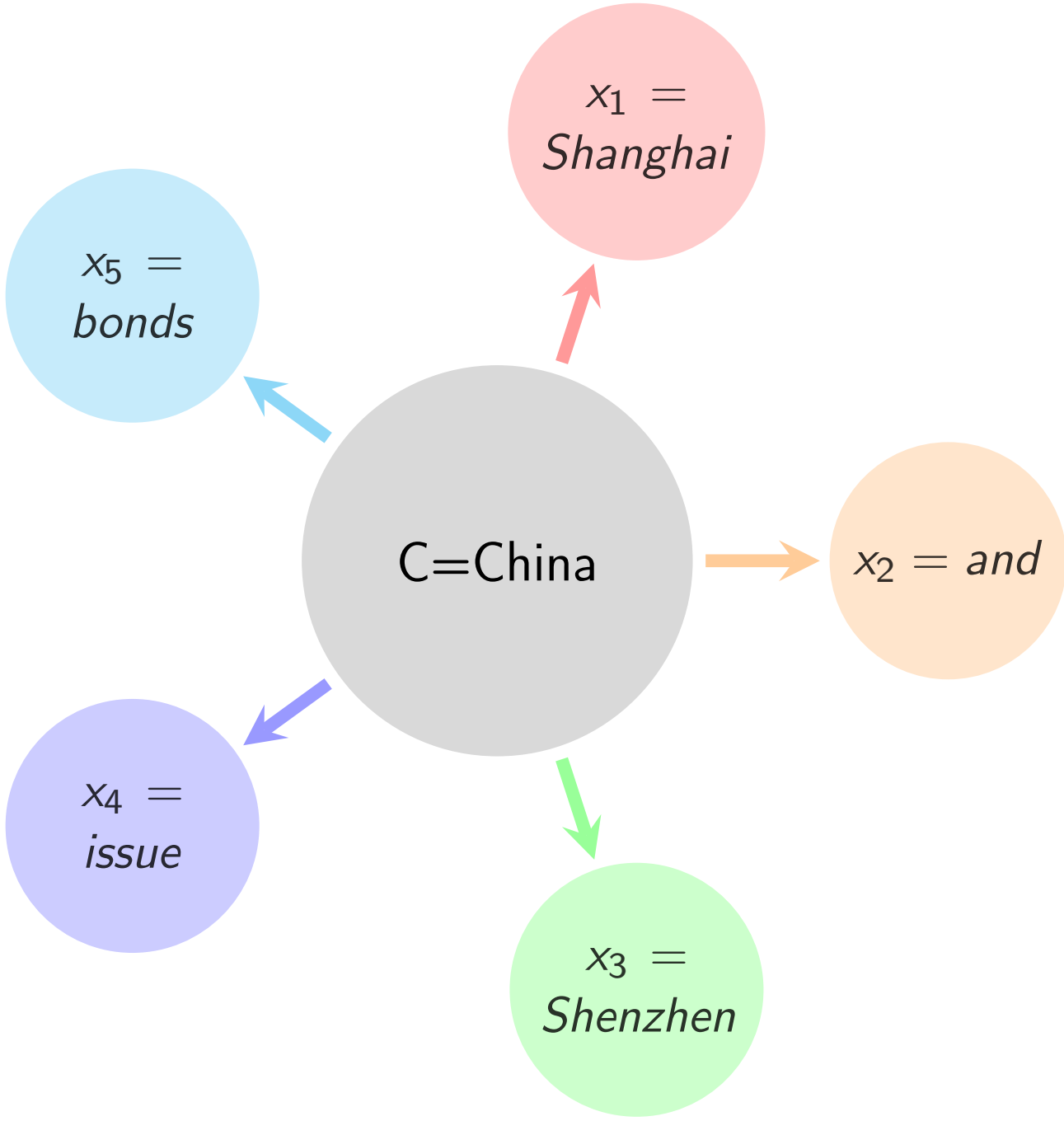



**Figure 6.1** Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.


$x_1$ = *Shanghai*
$x_5$ = *bonds*
C=China
$x_2$ = *and*
$x_4$ = *issue*
$x_3$ = *Shenzhen*

# Each class = a unigram language model

- Assigning each word: $P(word|c)$
- Assigning each sentence: $P(s|c) = \prod P(word|c)$

| 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 | 0.05 | 0.01 | 0.1 |

$$P(s|pos) = 0.0000005$$

# Naïve Bayes as a Language Model

- Which class assigns the higher probability to $s$?

| Class 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 |

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

# Содержание

# Таблица сопряженности 2-на-2

| | gold correct | gold not correct |
|---|---|---|
| system sel. | true positive | false positive |
| system n.s. | false negative | true negative |

# Точность и полнота

**Точность** : % правильных среди обнаруженных системой

$$P = \frac{TP}{TP + FP}$$

**Полнота** : % правильно обнаруженных системой среди всех обнаружаемых

$$R = \frac{TP}{TP + FN}$$

# Комбинированная оценка: $F$

- Баланс между $P/R$ – $F$-мера (взвешенное гармоническое среднее)

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

- консервативная мера, т.е., тяготеет к более низкому значению
- На практике чаще всего применяется F1, т.е. $\beta = 1$

$$F = \frac{2PR}{(P+R)}$$

# Содержание

# Заключение

- термины, введённые на лекции: обучение с учителем, генеративная модель, дискриминативная модель, наивный байесовский классификатор, точность, полнота, F-мера
- следующая лекция будет про логистическую регрессию (или метод максимальной энтропии)

# Задание 4

- Write a text classification pipeline to classify reviews as either positive or negative
- Find a good set of parameters using grid search, see http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.ht
- Evaluate the performance on a held out test set

# Tutorials

- https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f
- https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9
- http://ataspinar.com/2016/01/21/sentiment-analysis-with-bag-of-words/