Основы обработки текстов

Лекция 4

Статистические методы в обработки текстов. Поиск словосочетаний

Словосочетания/коллокации

- Для данной лекции Словосочетания = Коллокации = Фразеологические обороты - цепочки слов состоящие из двух или более элементов, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого
- Примеры:
  - Крепкий чай (не “сильный чай”)
  - Схема Бернулли (сравнить значения со значениями “Схема” и “Бернулли”)

Приложения

- Сравнения корпусов текстов
  - кластеризация документов в информационном поиске
  - Поиск плагиата
- Синтаксический разбор
- Компьютерная лексикография
- Генерация естественного языка
- Машинный перевод
- Выделение ключевых слов (терминов)

Выделение словосочетаний

Поиск кандидатов

- Основная предпосылка
  - Если два (или более) слова встречаются вместе часто, то, вероятно, это словосочетание
- Инструменты
  - Частота
  - Частота и фильтрация по тэгам
  - Математическое ожидание и дисперсия

Частота

- Подсчет частоты n-грам
- Выбрать наиболее встречающиеся
- Результат
  - –Корпус: New York Times
    - August-November, 1990
  - –Результат не интересен

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Частота с фильтрацией по тэгам

- Подсчет частоты n-грам
- определить части речи
- фильтрация кандидатов по шаблонам для частей речи
- выбрать наиболее встречающиеся

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Частота с фильтрацией по тэгам

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Мат. ожидание и дисперсия

- Часто устойчивые пары слов находятся не рядом
  - Пример
    - She knoked on his door
    - They knoked on the door
    - a man knocked on the metal front door
  - Важно это понимать, например при генерации текстов

Мат. ожидание и дисперсия

- Техника
  - Рассмотрим все пары слов в некотором окне
  - Посчитаем расстояние между словами
- Меры
  - Мат. ожидание (возможно отрицательное)
    - Показывает на сколько часто два слова встречаются вместе
  - Дисперсия (среднеквадратичное отклонение)
    - Вариабельность позиции

Мат. ожидание и дисперсия

She knocked on his door
Пары в окне длиной 3:
She knocked She on She his
knocked on knocked his knocked door
on his on door
his door

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \overline{d})^2}{n-1}$$

n - число раз, когда два слова встретились
d_i - смещение между словами
$\overline{d}$ - выборочное среднее смещений

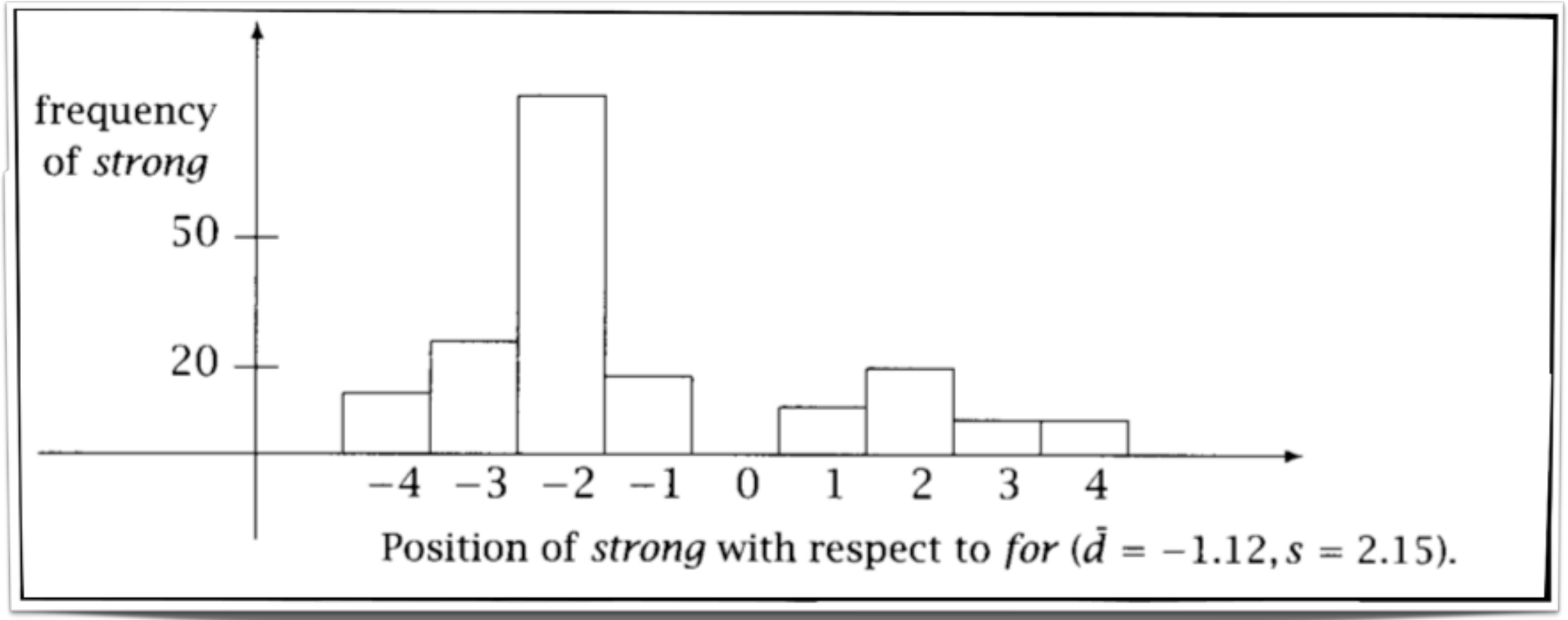
Пример: knocked ... door

$$\overline{d} = \frac{1}{3}(3+3+5) \approx 3.67$$

$$s = \sqrt{\frac{1}{2}((3-3.67)^2 + (3-3.67)^2 + (5-3.67)^2)}$$
$$\approx 1.15$$

Гисторамма

- Пример: strong ... for
  - “strong [business] support for”



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Пример

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

- Большое среднеквадратичное отклонение показывает, что сочетание не очень интересное

Проверка статистических гипотез

Проверка статистических гипотез

- Основная идея: слова словосочетания встречаются вместе значительно чаще чем просто случайно
- Инструменты:
  - t-критерий Стьюдента (t-test)
  - Критерий согласия Пирсона (Хи-квадрат)
  - Критерий отношения правдоподобия (Likelihood ratio test)

Нулевая гипотеза

- H_0-слова встречаются независимо
  - $P(w_1,w_2)=P(w_1)P(w_2)$
- Какова вероятность получить словосочетание w_1w_2, при условии что гипотеза верна?
  - $p=P(w_1w_2|H_0)$

Уровень значимости. p-value

- Уровень значимости (significance level)
  - Допустимая вероятность отвергнуть гипотезу H_0 при условие что она верна. (совершить ошибку 1-го рода)
  - Обозначают α
  - Часто в качестве уровня значимости выбирают $\alpha = 0.05$
- p-value (достигаемый уровень значимости)
  - минимальный уровень значимости (вероятность отвергнуть верную гипотезу H_0) при котором отвергается H_0
  - Чем меньше p–value тем лучше

Ошибки 1-го и 2-рода

- Возможны 4 ситуации

	Принимается H_0	Принимается H_1
Верна гипотеза H_0	OK	Ошибка 1-го рода
Верна альтернатива H_1	Ошибка 2-го рода	OK

- Цена ошибок 1-го и 2-го рода различна и ее сложно оценить. В области анализа данных обычно фиксируют вероятность ошибки 1-го рода и минимизируют вероятность ошибки 2-го рода.

T-критерий Стьюдента

- T-статистика

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

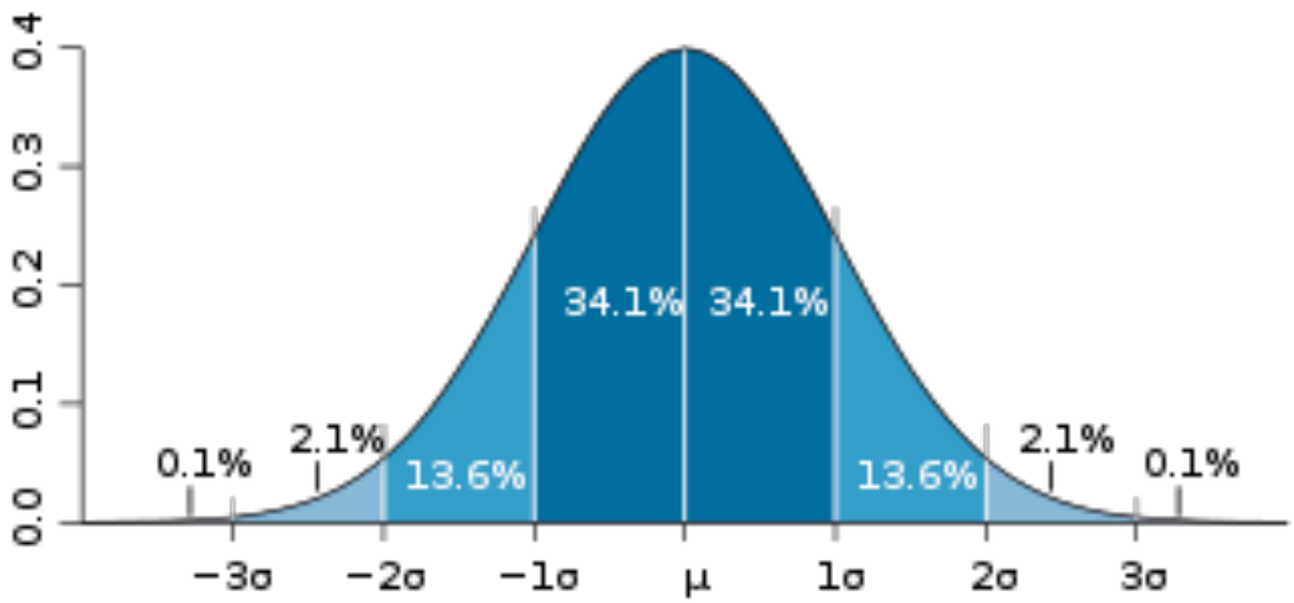
μ -ожидаемое мат. ожидание
$\overline{x}$ -выборочное среднее
s^2-выборочная дисперсия
N-размер выборки

- Распределение Стьюдента (стремится к нормальному при больших N)

	P	0.05	0.025	0.01	0.005	0.001	0.0005
	C	90%	95%	98%	99%	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(z)	∞	1.645	1.960	2.326	2.576	3.091	3.291

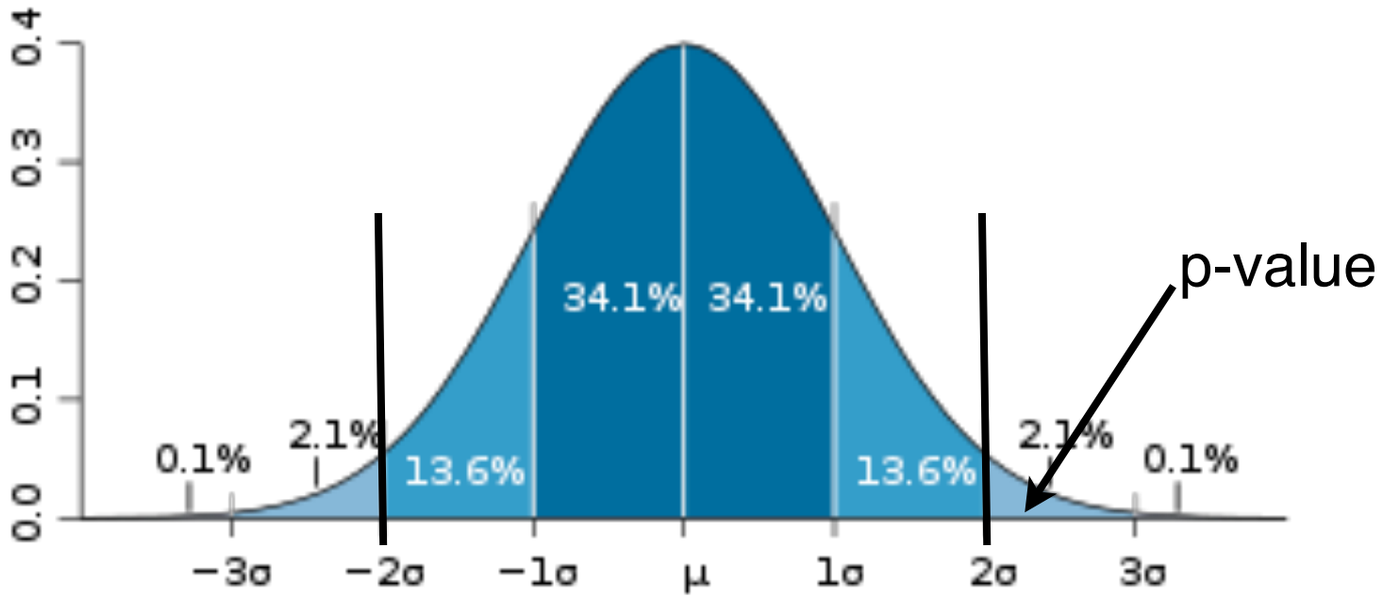
T-критерий Стьюдента

- Разработан Уильямом Госсетом для оценки качества пива Гиннесс
- Рассмотрим распределение выборочного среднего у всевозможных выборок длины n
- По ЦПТ, при больших n:



T-критерий Стьюдента

- Если для наших данных наблюдаемое выборочное среднее сильно отклоняется от ожидаемого при нулевой гипотезе, то с вероятностью p гипотеза не верна
- α - ошибка первого рода
- $p < \alpha$ - отвергаем гипотезу



T-критерий. Пример

- Предположим, что средний рост мужчин в популяции равен 158 см
- Для выборки из 200 мужчин $\overline{x} = 169, s^2 = 2600$
- Тогда $t = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} = 3.05$

	P	0.05	0.025	0.01	0.005	0.001	0.0005
	C	90%	95%	98%	99%	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(z)	∞	1.645	1.960	2.326	2.576	3.091	3.291

- Для α=0.005:
- 3.05>2.576
- отвергаем гипотезу

T-критерий для словосочетаний

- Пусть нулевая гипотеза верна
- Рассмотрим процесс случайной генерации биграмм, если встретили биграмму w_1w_2 (с вероятностью p) генерируем 1, в противном случае 0 (схема Бернулли) ⇒ биномиальное распределение
- мат. ожидание = p
- дисперсия = p(1-p)≈p при малых p

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

μ =H_0=P(w_1)P(w_2)
$\overline{x}$ -отношение w_1w_2 к общему кол-ву биргамм
s^2-отношение w_1w_2 к общему кол-ву биргамм
N-общее количество биграмм

Пример

- new companies (встретилась 8 раз)

$$P(new) = \frac{15828}{14307668} \quad P(companies) = \frac{4675}{14307668}$$

$$H_0 : P(new\ companies) = P(new)P(companies) = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

$$\overline{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$$

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 0.999932$$

- не можем отвергнуть гипотезу

Для корпуса

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Хи-квадрат

- Сравнить наблюдаемые частоты в корпусе с ожидаемыми частотами при верной гипотезе о независимости
- Если различие большое - отвергаем гипотезу
- (Выборка должна быть большая)

χ^2 - общая формула

- Меры:
  - E_{ij} = ожидаемое кол-во коллокаций
  - O_{ij} = наблюдаемое кол-во коллокаций

$$\chi^2 = \sum_{i,j} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

- Результат
  - Смотрим число в таблице для распределения χ^2
    - если число в таблице меньше, то отвергаем гипотезу

χ^2 - для биграмм

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

	p	0.99	0.95	0.10	0.05	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	3.84	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

Критерий отношения правдоподобия

- На сколько более правдоподобна одна гипотеза, чем другая
- H_1: $P(w_2|w_1) = p = P(w_2|\neg w_1)$
- H_2: $P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$

$$(p_1 >> p_2)$$

Критерий отношения правдоподобия

	H_1	H_2
$P(w_2\|w_1)$	$p = \frac{c_2}{N}$	$p_1 = \frac{c_{12}}{c_1}$
$P(w_2\|\neg w_1)$	$p = \frac{c_2}{N}$	$p_2 = \frac{c_2 - c_{12}}{N - c_1}$

- Так же как в t-критерии предполагаем схему Бернулли и биномиальное распределение

$$b(k; n, x) = C_n^k x^k (1-x)^{n-k}$$

	H_1	H_2
c_{12} из c_1 биграм-это w_1w_2	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
c_2-c_{12} из N-c_1 биграм-это не w_1w_2	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}; N - c_1, p_2)$

$$L(H_1) = b(c_{12}; c_1, p)b(c2 - c_{12}; N - c_1, p)$$
$$L(H_2) = b(c_{12}; c_1, p_1)b(c2 - c_{12}; N - c_1, p_2)$$

Отношение правдоподобия

$$
\begin{aligned}
\log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\
&= \log \frac{\mathrm{b}(c_{12}, c_1, p)\mathrm{b}(c_2 - c_{12}, N - c_1, p)}{\mathrm{b}(c_{12}, c_1, p_1)\mathrm{b}(c_2 - c_{12}, N - c_1, p_2)} \\
&= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\
&\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)
\end{aligned}
$$

где $L(k, n, x) = x^k(1 - x)^{n-k}$

Результат для корпуса

$-2\log\lambda$	$C(w^1)$	$C(w^2)$	$C(w^1w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

- $-2log\lambda$ имеет распределение χ^2

Проверка статистических гипотез для сравнения классификаторов

Сравнение классификаторов

- Пусть у нас есть два классификатора
  - Разработанный нами
  - State-of-the-art решение

	F1
Наш	**0,853**
SotA	0,845

- Действительно ли наше решение лучше?

Сравнение классификаторов

- Посчитаем F1-меру на нескольких наборах данных

```
y1 = [0.81922220, 0.80076079, 0.89213162, 0.85617380, 0.84878412,
      0.87187914, 0.88404899, 0.83593810, 0.89170907, 0.82974126]
y2 = [0.80006565, 0.81164693, 0.83667603, 0.83123125, 0.88698,
      0.89964834, 0.80711508, 0.86260427, 0.84889595, 0.86098152]
```

- Какую нулевую гипотезу выбрать?
- Что можно сказать о распределениях?

Непараметрические тесты

- Когда нельзя сделать предположение о распределении случайной величины
  - Скошенные распределения
  - Мало данных для ЦПТ
- Используют непараметрические тесты
  - Знаковые критерии
  - Ранговые критерии
  - Перестановочные критерии

Вариационный ряд

- Пусть у нас есть выборка $X_1, \ldots, X_n$
- Упорядочим ее по не убыванию

$$X_{(1)} \leq \ldots < X_{(k_1)} = \ldots = X_{(k_2)} < \cdots \leq X_{(}n)$$

Связка размера k2-k1+1

- Ранг наблюдения :
  - Если X_i не в связке, то $rank(X_i) = r, X_i = X_{(r)}$
  - Если X_i в связке, $X_{(k_1)} = \ldots = X_{(k_2)}$ то

$$rank(X_i) = \frac{k_1 + k_2}{2}$$

Критерий Уилкоксона

- Ранговый критерий
- Нулевая гипотеза $H_0 : med(X_1 - X_2) = 0$
- Альтернатива $H_1 : med(X_1 - X_2) \neq 0$
- Считаем статистику

$$W(X_1^n, X_2^n) = \sum_{i=1}^{n} rank(|X_{1i} - X_{2i}|) * sign(X_{1i} - X_{2i})$$

- Нулевое распределение табличное
  - при справедливости нулевой гипотезы, каждый из рангов в нашей выборке мог с одинаковой вероятностью реализоваться с любым знаком: и с «+», и с «–». Таким образом, мы получаем 2 в степени n вариантов распределения знаков по рангам. Переберём все эти варианты, и на каждом из этих вариантов знаков посчитаем значение статистики.

Пример

```
from scipy.stats import wilcoxon, rankdata
import numpy as np
```

```
y1 = [0.81922220, 0.80076079, 0.89213162, 0.85617380, 0.84878412,
      0.87187914, 0.88404899, 0.83593810, 0.89170907, 0.82974126]

y2 = [0.80006565, 0.81164693, 0.83667603, 0.836123125, 0.88698,
      0.89964834, 0.80711508, 0.86260427, 0.84889595, 0.86098152]
```

```
np.mean(y1)
```

```
0.853038909
```

```
np.mean(y2)
```

```
0.8450736895000001
```

```
rankdata(np.concatenate((y1,y2)))
```

```
array([ 5.,  2., 19., 12., 10., 15., 16.,  7., 18.,  6.,  1.,  4.,  9.,
        8., 17., 20.,  3., 14., 11., 13.])
```

```
wilcoxon(y1,y2,alternative='greater')
```

```
WilcoxonResult(statistic=32.0, pvalue=0.3232311013320848)
```

Проблемы использования статистической проверки гипотез

- Заблуждения о p-value
- Множественная проверка гипотез

Заблуждения о p-value

- p-value это не вероятность того, что верна основная гипотеза H_0
  - p-value так же не является вероятностью того, что неверна альтернатива H_1
  - Статистические критерии вообще не умеют оценивать такие вероятности.
- Большое значение p-value не означает что H_0 верна
  - Возможны различные объяснения большого значение p-value:
    - Гипотеза H_0 верна
    - Гипотеза H_0 не верна, но размер выборки недостаточно большой. Большее число экспериментов позволят отвергнуть H_0
    - Гипотеза H_0 не верна, но мы применяем неправильный критерий

Множественная проверка гипотез

```
y1 = 0.8 + 0.1*np.random.rand(10)

p = 1
while p > 0.05:
    y2 = 0.8 + 0.1*np.random.rand(10)
    w = wilcoxon(y1,y2,alternative='greater')
    u, p = w
    print(w)
```

```
WilcoxonResult(statistic=42.0, pvalue=0.06970698666076602)
WilcoxonResult(statistic=39.0, pvalue=0.12056068413870913)
WilcoxonResult(statistic=41.0, pvalue=0.08440347677825405)
WilcoxonResult(statistic=40.0, pvalue=0.10131080385615598)
WilcoxonResult(statistic=36.0, pvalue=0.1931353601832413 4)
WilcoxonResult(statistic=29.0, pvalue=0.4392408717164356)
WilcoxonResult(statistic=46.0, pvalue=0.029668059940454 31)
```

Ура! Мой классификатор значимо лучше!

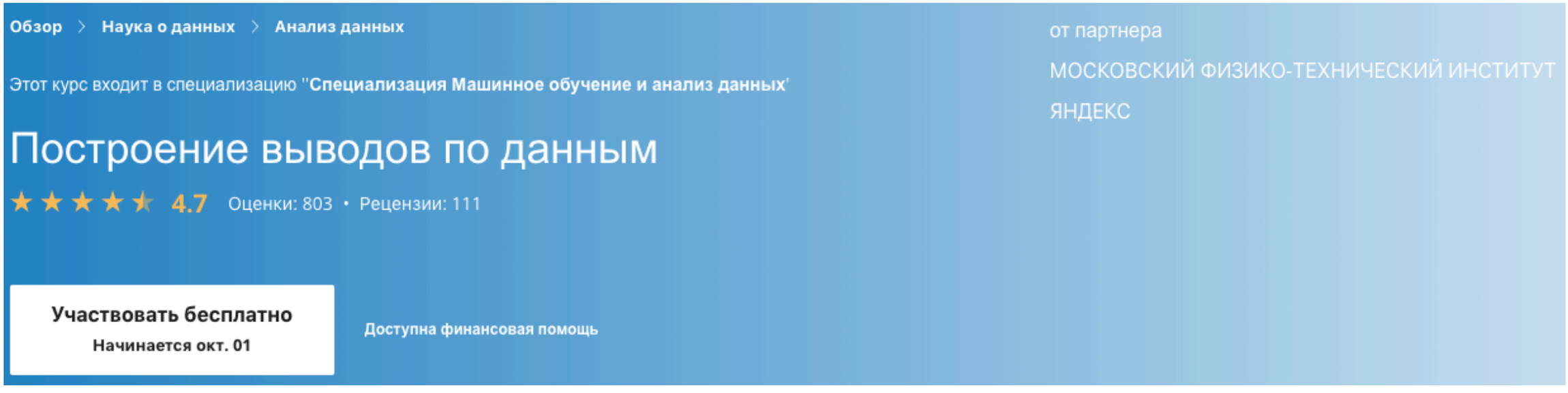
Для самостоятельного изучения

- Решение проблемы множественной проверки гипотез
- Доверительные интервалы
  - сравнение классификаторов с помощью доверительных интервалов

Полезная инфомрация

Илья Козлов: Проверка статистических гипотез для сравнения алгоритмов классификации
https://www.youtube.com/watch?v=6cnF8lQRYN0

https://www.coursera.org/learn/stats-for-data-analysis



Заключение

- Аппарат проверки статистических гипотез должен знать любой человек, занимающийся анализом данных
- В обработке текстов методы проверки статистических гипотез можно применять для поиска словосочетаний применяются
- Одним из наиболее важных применений является сравнение моделей машинного обучения

Следующая лекция

- Векторные представления слов