

Две задачи машинного обучения

Воронцов Константин Вячеславович

Машинное обучение • МФТИ • 16 мая 2017

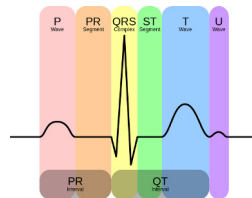
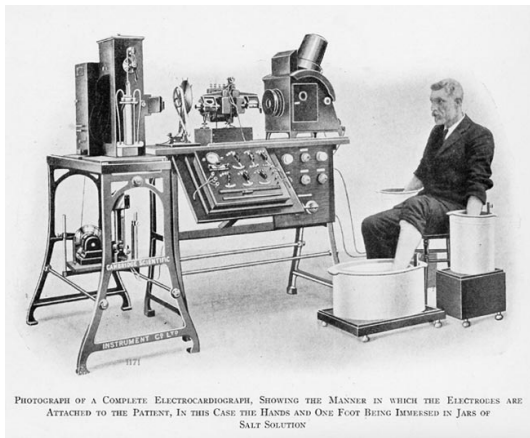
1 Диагностика заболеваний по ЭКГ

- Информационный анализ электрокардиосигналов
- Статистические обоснования
- Измерение качества диагностики

2 Разведочный информационный поиск

- Тематическое моделирование текстовых коллекций
- Разведочный тематический поиск
- Оптимизация параметров модели

Электрокардиография

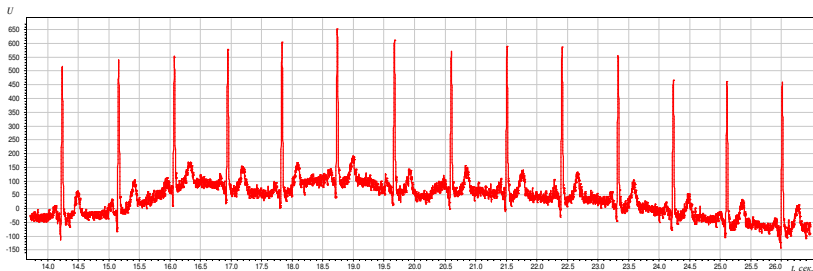
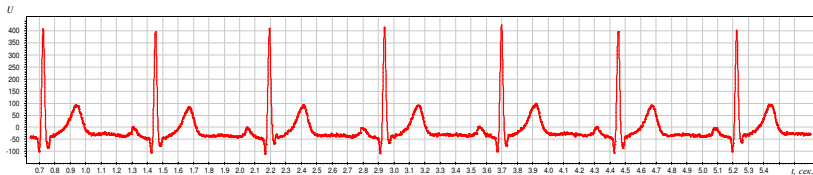


1872 — первые записи электрической активности сердца

1911 — коммерческий электрокардиограф (фото)

1924 — нобелевская премия по медицине, Виллем Эйнтховен

Примеры электрокардиограмм



В основе диагностики заболеваний сердца — многочисленные наблюдения за особенностями PQRST-комплекса

Теория информационной функции сердца [В.М.Успенский]

Возможна ли диагностика несердечных заболеваний по ЭКГ?

Предпосылки:

- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Цифровая электрокардиография высокого разрешения

Предположения:

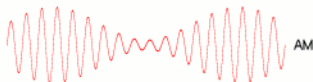
- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика
- Каждое заболевание по-своему «модулирует» ЭКГ-сигнал

Аналогии в теорию передачи сигналов

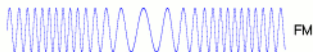
Модуляция — процесс, при котором высокочастотная волна используется для переноса низкочастотного сигнала.



— низкочастотный сигнал



— амплитудная модуляция



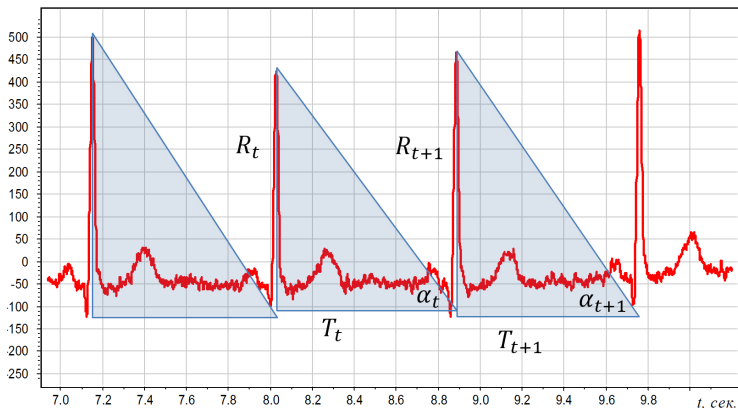
— частотная модуляция

Демодуляция — процесс, обратный модуляции, преобразование модулированных колебаний высокой (несущей) частоты в исходный низкочастотный сигнал.

В случае ЭКГ несущая частота — биения сердца, ~ 1 Гц
А что будет аналогом модуляции и демодуляции?

Вариабельность интервалов и амплитуд кардиоциклов

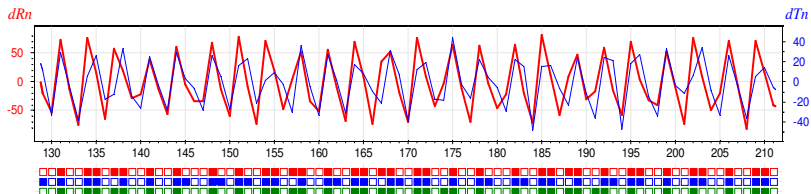
приращение амплитуд: $dR_t = R_{t+1} - R_t$
приращение интервалов: $dT_t = T_{t+1} - T_t$
приращение углов: $d\alpha_t = \alpha_{t+1} - \alpha_t$, $\alpha_t = \arctg \frac{R_t}{T_t}$



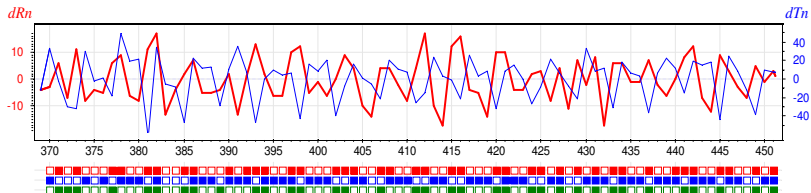
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t , dT_t , $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:



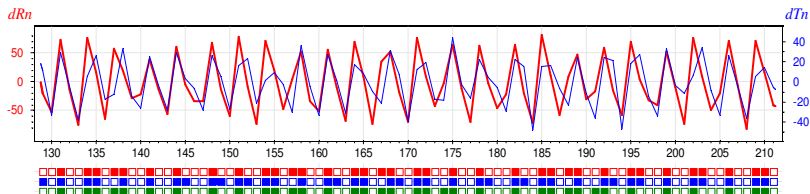
Больной (язвенная болезнь):



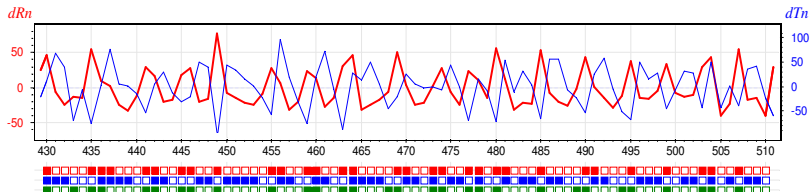
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t , dT_t , $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:



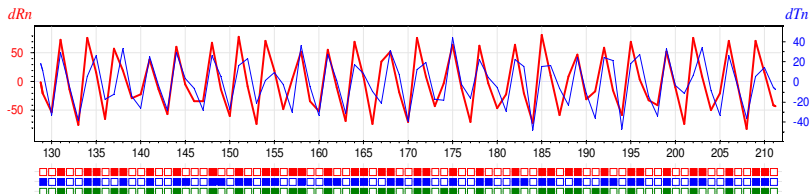
Больной (гипертония):



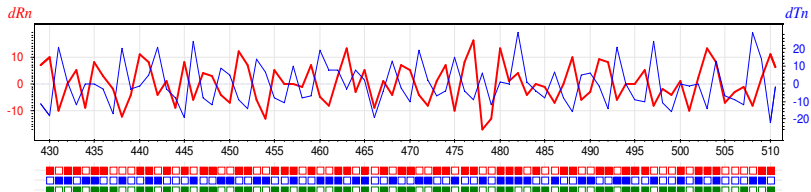
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t , dT_t , $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:



Больной (рак):



Технология информационного анализа ЭКГ-сигналов

Этап I. Методы символьной динамики

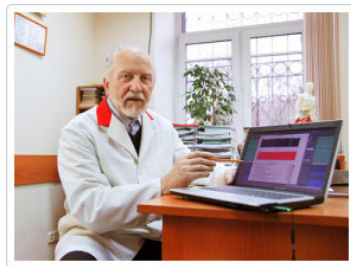
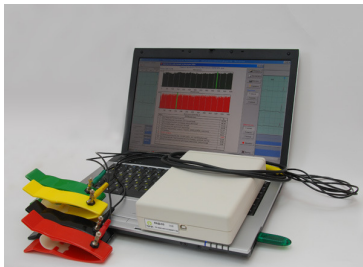
- 1 Демодуляция — вычисление амплитуд, интервалов и углов по кардиограмме длиной 600 кардиоциклов
- 2 Дискретизация — перевод в кодограмму — 599-символьную строку в 6-буквенном алфавите
- 3 Векторизация — перевод в вектор $6^3=216$ частот триграмм

Этап II. Методы машинного обучения

- 1 Формирование обучающих выборок здоровых и больных
- 2 Формировании модели классификации
- 3 Оптимизация модели классификации
- 4 Оценивание качества диагностики

Диагностическая система «Скринфакс»

Цифровой электрокардиограф с улучшенной помехозащищённостью и расширенной полосой пропускания.



- более 15 лет исследований и накопления данных
- более 20 тысяч прецедентов (кардиограмма + диагнозы)
- более 40 заболеваний

Объём исходных данных (по заболеваниям)

«абсолютно здоровые»	АЗ	193
гипертоническая болезнь	ГБ	1894
ишемическая болезнь сердца	ИБС	1265
сахарный диабет (СД1 и СД2)	СД	871
язвенная болезнь	ЯБ	785
миома матки	ММ	781
узловой (диффузный) зоб щитовидной железы	УЩ	748
дискинезия желчевыводящих путей	ДЖВП	717
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
вегетососудистая дистония	ВСД	694
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
холецистит хронический	ХХ	340
асептический некроз головки бедренной кости	НГБК	324
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
желчнокаменная болезнь	ЖКБ	278
аднексит хронический	АХ	276
аденома простаты	ДГПЖ	260
анемия железододефицитная	ЖДА	260

Вход: последовательность интервалов и амплитуд $(T_t, R_t)_{t=1}^{N+1}$

$dR_t = R_{t+1} - R_t$	+	-	+	-	+	-
$dT_t = T_{t+1} - T_t$	+	-	-	+	+	-
$d\alpha_t = \alpha_{t+1} - \alpha_t$	+	+	+	-	-	-
s_t	A	B	C	D	E	F

Выход: кодограмма $x = (s_t)_{t=1}^N$ — последовательность символов алфавита $\{A, B, C, D, E, F\}$:

[illegible]

Векторизация ЭКГ-сигнала

x — кодограмма, последовательность символов $\{A, B, C, D, E, F\}$:

DBEEACFDAAFBABDDAADF AAFEEACFEACFBREFFAABFFAFAFFAFAFFAFAEBAEFBEAFAFCAFFAAD
FCAFFAADFCADFCCDFDACDFACDFAEFFACFEADFCAFBACADFFECFFAFAFFAFAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAABBAADFADFDAAFCFCFCEDFCEEFCAEFBECBBBAADBAACFFAFAFFA
CFFCECFDAABDAEFFAFAFFCEDBFAFFAEFFAEFBACFBADFEAFAFFCAFFDAFFAEBDAADBBADFADFF
EABFCCAFDEEBDECFFACFFAABFAADFBAFFACFFFAEFFACFFACFFCECFBAFFFAFAFFFAFAFFAADFBA
AABFACDFAEFFAABBAEFFEAFBCECFDECCFBAFFAADFACDFAFFAADFCAADFAEFBFAFFCADFE
AFFCECFCECFFAFFABCFDAFAFFADBFCAEFFAABFACBFAEBFAEBFAFFBAFFAFAFFDADFADABFB
CAFFAECCFACFFACDFCADFADABFAEDDABBFACDBAFAFFAFAFFCADFAADFACFFAEDFCACFAEBCE

$f_j(x)$ — частота триграммы $j = 1, \dots, n$ ($n=216$) в кодограмме x :

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Линейная модель классификации с двумя классами

$\{x_i\}_{i=1}^{\ell}$ — обучающая выборка кодограмм

y_i — класс объекта x_i : больной $y_i = 1$, здоровый $y_i = 0$

Основная эмпирическая гипотеза:

- у больных одни триграммы частые, у здоровых — другие

Линейная модель классификации:

$$\langle x, w \rangle = \sum_{j=1}^n w_j f_j(x), \quad a(x) = \begin{cases} 1, & \langle x, w \rangle \geq w_0 \\ 0, & \langle x, w \rangle < w_0 \end{cases}$$

где w_j — вес j -й триграммы:

- $w_j > 0$, если триграмма более характерна для больных
- $w_j < 0$, если триграмма более характерна для здоровых
- $w_j = 0$, если триграмма не информативна для этой болезни

Наивный байесовский классификатор и его модификации

Число объектов класса y , для которых триграмма j частая

$$S_y^j = \sum_{i=1}^{\ell} [y_i = y] [f_j(x_i) \geq \theta]$$

Число объектов класса y , для которых триграмма j редкая

$$s_y^j = \sum_{i=1}^{\ell} [y_i = y] [f_j(x_i) < \theta]$$

Вес триграммы j больше, если S_1^j , s_0^j больше, S_0^j , s_1^j меньше:

$$w_j = S_1^j / S_0^j$$

$$w_j = S_1^j s_0^j / S_0^j s_1^j$$

$$w_j = \log(S_1^j / S_0^j)$$

$$w_j = \log(S_1^j s_0^j / S_0^j s_1^j)$$

$$w_j = \sqrt{S_1^j} - \sqrt{S_0^j}$$

$$w_j = \sqrt{S_1^j s_0^j} - \sqrt{S_0^j s_1^j}$$

Методы машинного обучения

- **Наивный байесовский классификатор**
 - 😊 простой интерпретируемый линейный классификатор
 - 😞 качество классификации невысокое
- **Наивный байесовский классификатор + отбор признаков**
 - 😊 качество классификации лучше
 - 😊 находит один диагностический эталон каждой болезни
- **Метод главных компонент + логистическая регрессия**
 - 😊 качество классификации высокое
 - 😞 не определяет диагностические эталоны болезней
- **SVM, нейронные сети, случайный лес**
 - 😊 качество классификации высокое
 - 😞 неоправданно сложное, неинтерпретируемое решение
- **Тематические модели классификации**
 - 😊 автоматически находит все диагностические эталоны
 - 😊 качество классификации среднее

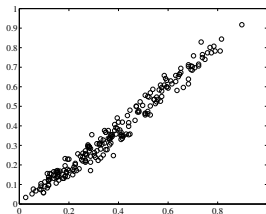
Существуют сочетания триграмм, специфичные для болезней

Точки на графиках соответствуют триграммам, $j = 1, \dots, 216$

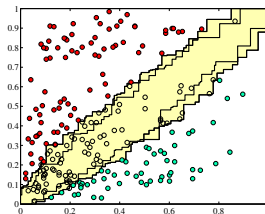
— ось X: доля здоровых с частой триграммой $f_j(x_i) \geq 2$

— ось Y: доля больных с частой триграммой $f_j(x_i) \geq 2$

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i



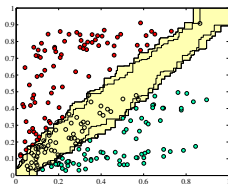
наблюдаемые y_i

Слева: как распределятся точки, если объектам x_i назначить случайно переставленные метки классов y_i .

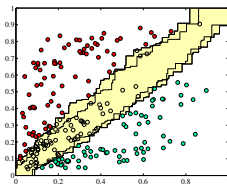
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

Существуют сочетания триграмм, специфичные для болезней

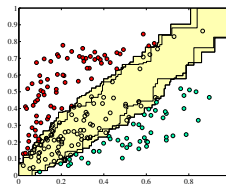
Для каждой болезни есть свои неслучайно частые триграммы



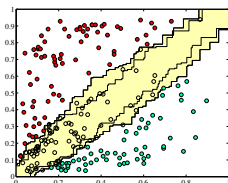
ишемия сердца



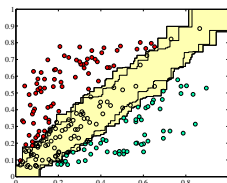
гипертония



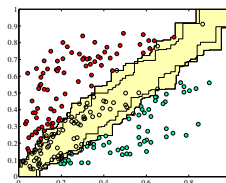
рак



желчнокаменная болезнь



миома матки

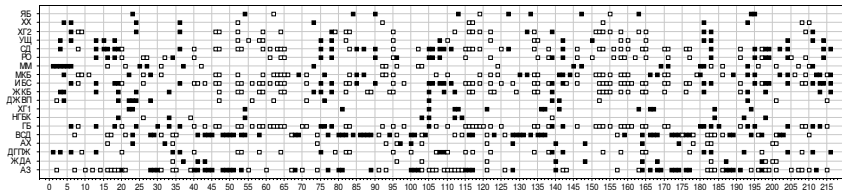


язвенная болезнь

Болезни отличаются наборами информативных триграмм

ось X: номера триграмм $j = 1, \dots, n$, $n = 216$

ось Y: болезни (АЗ — абсолютно здоровые)



□ — неслучайно низкая частота триграммы

■ — неслучайно высокая частота триграммы

Вывод 1. Для каждой болезни есть триграммы с неслучайно высокой и неслучайно низкой частотой

Вывод 2. Болезни отличаются *диагностическими эталонами* — наборами специфичных триграмм с неслучайно высокой частотой

Терминология диагностики

Положительный диагноз — алгоритм предсказывает болезнь

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{\sum_{i=1}^{\ell} [y_i = 1][a(x_i) = 1]}{\sum_{i=1}^{\ell} [y_i = 1]}$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{\sum_{i=1}^{\ell} [y_i = 0][a(x_i) = 0]}{\sum_{i=1}^{\ell} [y_i = 0]}$$

Максимизируем чувствительность и специфичность

- они не зависят от соотношения мощностей классов
- они подходят для несбалансированных выборок

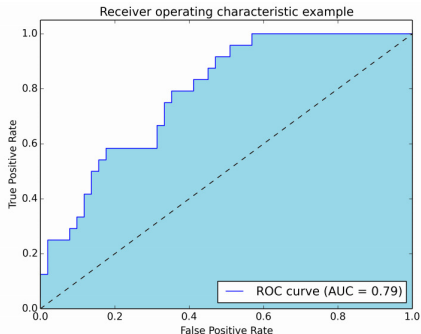
ROC-кривой и AUC — площадь под ROC-кривой

Модель классификации: $a(x) = [\langle x, w \rangle > w_0]$

по оси X: 1 – специфичность = FPR, False Positive Rate,

по оси Y: чувствительность = TPR, True Positive Rate

Каждая точка ROC-кривой соответствует значению порога w_0
(ROC — «receiver operating characteristic»),



Результаты кросс-валидации

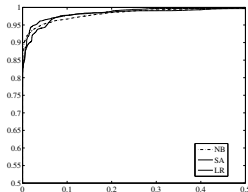
Обучающая выборка: оптимизация параметров модели

Тестовая выборка: Чувствительность, Специфичность, AUC

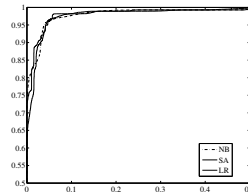
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	C% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

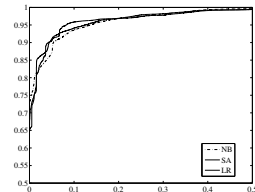
ROC-кривые в осях X:(1—специфичность), Y:чувствительность



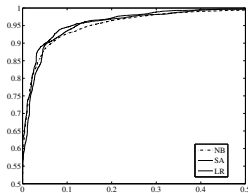
асептический некроз ГБК



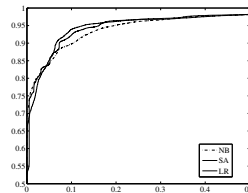
желчнокаменная болезнь



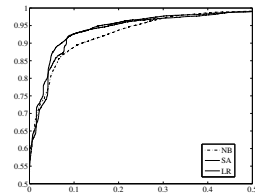
ишемическая болезнь



хронический гастрит 1



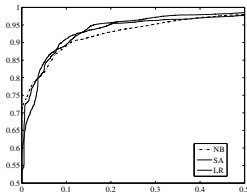
сахарный диабет



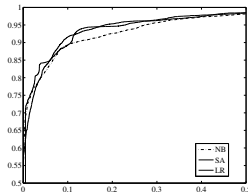
гипертония

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

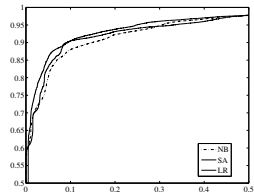
ROC-кривые в осях X:(1—специфичность), Y:чувствительность



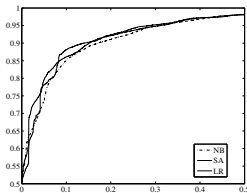
рак общий



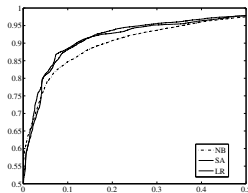
аденома простаты



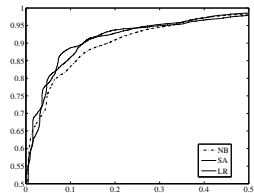
зоб щитовидной железы



хронический гастрит 2



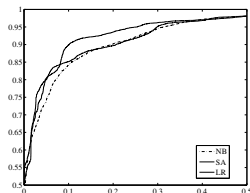
дискинезия ЖВП



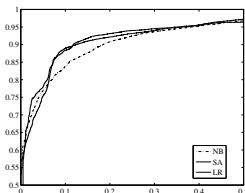
мочекаменная болезнь

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

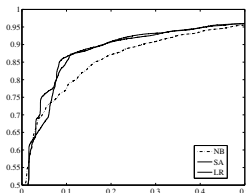
ROC-кривые в осях X:(1—специфичность), Y:чувствительность



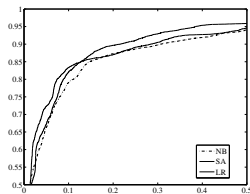
хронический холецистит



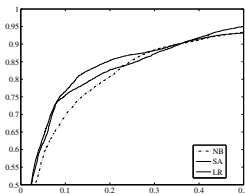
язвенная болезнь



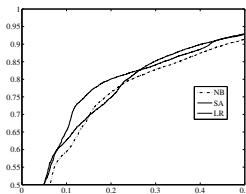
миома матки



хронический аднексит



анемия

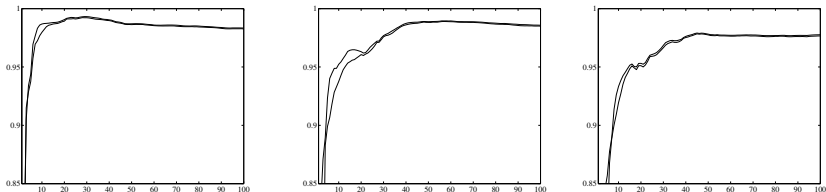


вегетососудистая дистония

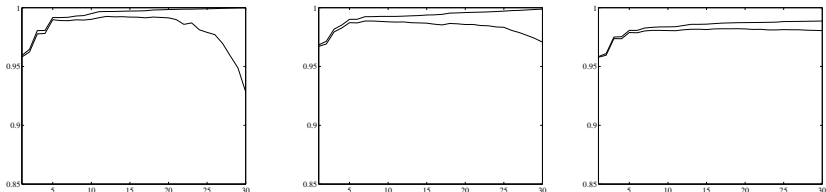
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (наивный Байес на K признаках):



Логистическая регрессия (K — число главных компонент):



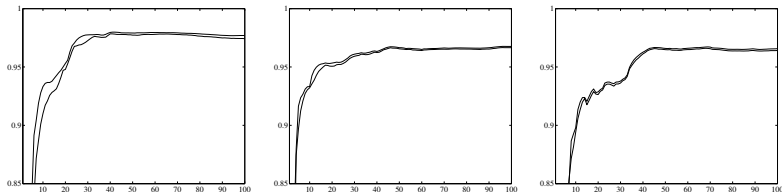
асептический некроз ГБК желчнокаменная болезнь ишемическая болезнь

Тонкая (верхняя) линия — на обучающей выборке

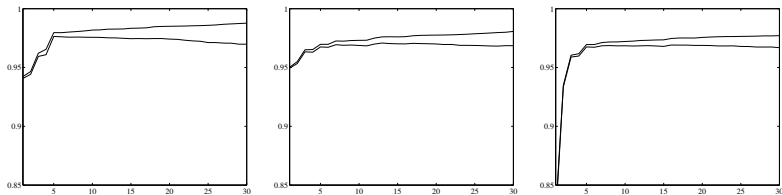
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):



Логистическая регрессия (K — число главных компонент):



хронический гастрит 1

сахарный диабет

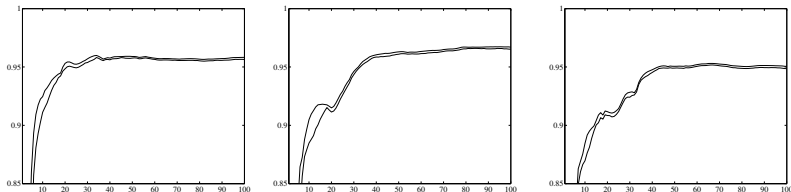
гипертония

Тонкая (верхняя) линия — на обучающей выборке

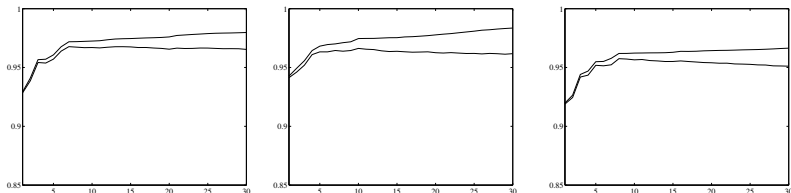
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):



Логистическая регрессия (K — число главных компонент):



рак общий

аденома простаты

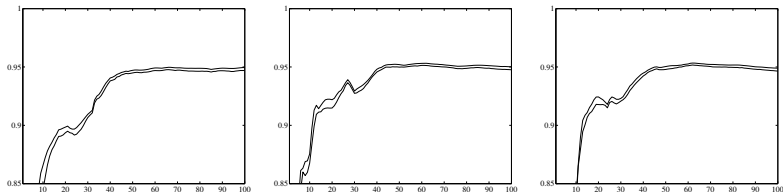
зоб щитовидной железы

Тонкая (верхняя) линия — на обучающей выборке

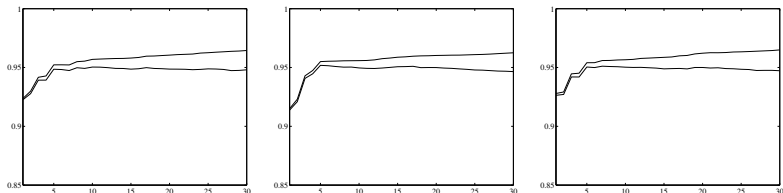
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):



Логистическая регрессия (K — число главных компонент):



хронический гастрит 2

дискинезия ЖВП

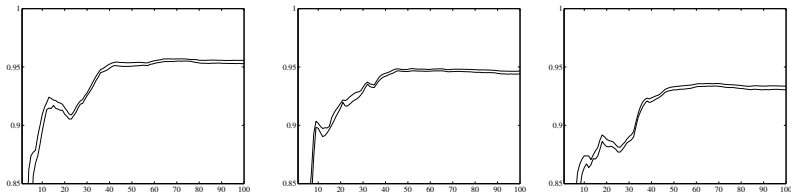
мочекаменная болезнь

Тонкая (верхняя) линия — на обучающей выборке

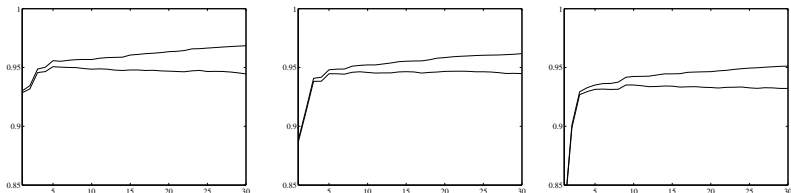
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):



Логистическая регрессия (K — число главных компонент):



хронический холецистит

язвенная болезнь

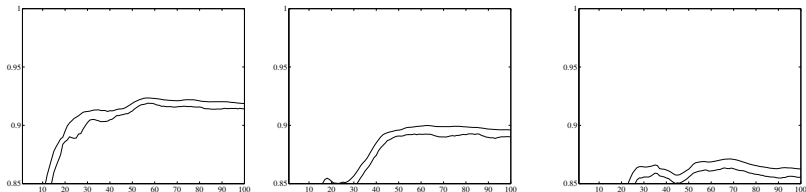
миома матки

Тонкая (верхняя) линия — на обучающей выборке

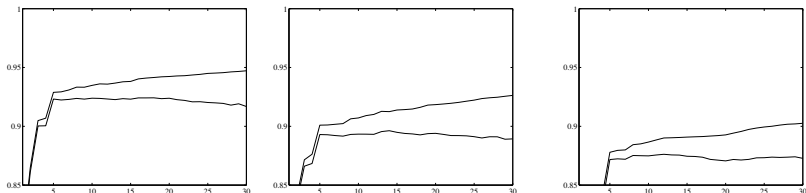
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):



Логистическая регрессия (K — число главных компонент):



хронический аднексит

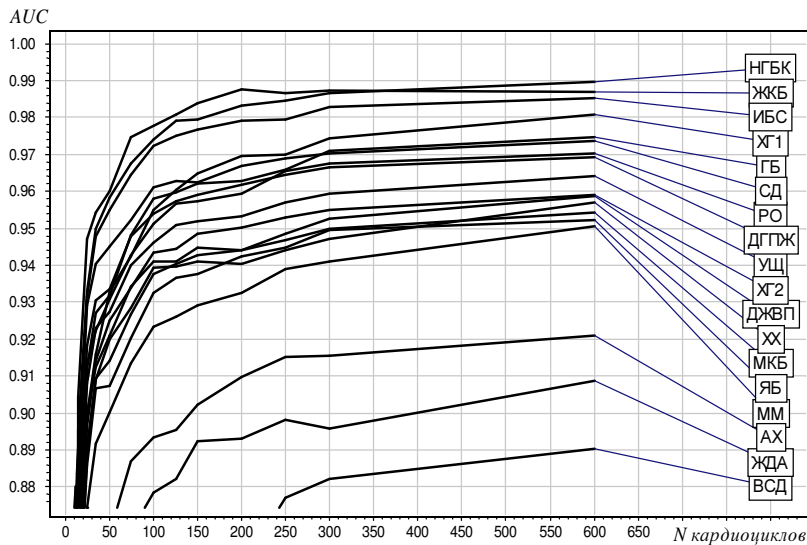
анемия

вегетососудистая дистония

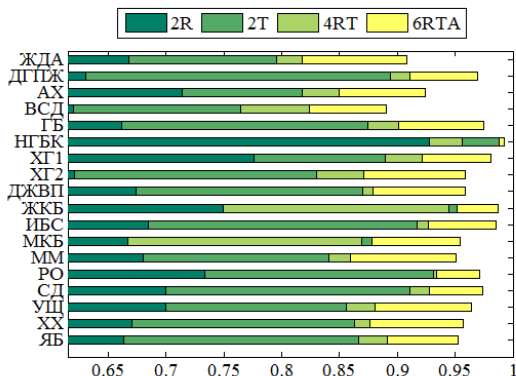
Тонкая (верхняя) линия — на обучающей выборке

Толстая (нижняя) линия — на тестовой выборке

Зависимость AUC от длительности регистрации ЭКГ



Зависимость AUC от типа символьного кодирования



2R: 2-символьная, только приращения амплитуд

2T: 2-символьная, только приращения интервалов

4RT: 4-символьная, приращения интервалов и амплитуд

6RTA: 6-символьная, приращения интервалов, амплитуд и их отношений

Открытые данные по инфарктам миокарда: база данных PTB

Данные национального метрологического института Германии.

Число записей ЭКГ-сигналов: 320 больных, 74 здоровых.

Длительность регистрации ЭКГ: 100–200 кардиоциклов.

AUC при 6-символьном кодировании (6RTA) для трёх методов:

LR — логистическая регрессия,

RF — случайный лес,

SA — наивный Байес с отбором признаков

	LR	RF	SA
2-граммы	87.7	87.9	86.1
3-граммы	89.4	89.6	87.1
4-граммы	88.6	87.7	86.9

Bousseljot R., Kreiseler D., Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik. 1995.

Выводы

О данной задаче:

- многие болезни можно диагностировать по ЭКГ
- с очень высокой чувствительностью и специфичностью
- достаточно 300–600 кардиоциклов (5–10 минут)

О прикладных исследованиях в машинном обучении:

- в основе — идеи специалиста о прикладной области
- изобретение удачных признаков — 90% успеха
- начинать с разведочного анализа и визуализации
- пробовать стандартные методы «из коробки»
- пробовать простые методы
- баланс сложности между недо- и пере-переобучением
- выбрать критерий качества и делать кросс-валидацию

Что такое «тема» в коллекции текстовых документов?

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов

Более формально,

- *тема* — условное распределение на множестве терминов,
 $p(w|t)$ — вероятность (частота) термина w в теме t ;
- *тематика* документа — условное распределение
 $p(t|d)$ — вероятность (частота) темы t в документе d .

Когда автор писал термин w в документе d , он думал о теме t , и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Задача тематического моделирования

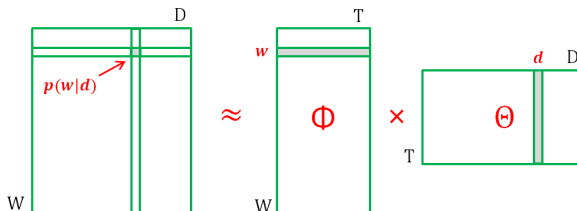
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $p(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$

- $\varphi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*,
если её решение

- существует,
- единственно,
- устойчиво.



Жак Саломон Адамар
(1865–1963),

Задача стохастического матричного разложения является
некорректно поставленной — её решение не единственно:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для невырожденных $S_{T \times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — дополнительные ограничения на Φ, Θ .

ARTM: Аддитивная Регуляризация Тематических Моделей

Максимизация \log правдоподобия с регуляризатором R :

$$\sum_{d,w} n_{dw} \ln \sum_t \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} \text{Е-шаг:} & \quad p_{tdw} \equiv p(t|d, w) = \text{norm}_{t \in T}(\varphi_{wt} \theta_{td}) \\ \text{М-шаг:} & \quad \begin{cases} \varphi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in D} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{aligned}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

Некоторые полезные регуляризаторы

PLSA, вероятностный латентный семантический анализ
— модель без регуляризации:

$$R(\Phi, \Theta) = 0$$

LDA, латентное размещение Дирихле — сглаживание или разреживание распределений $p(w|t)$ и $p(t|d)$:

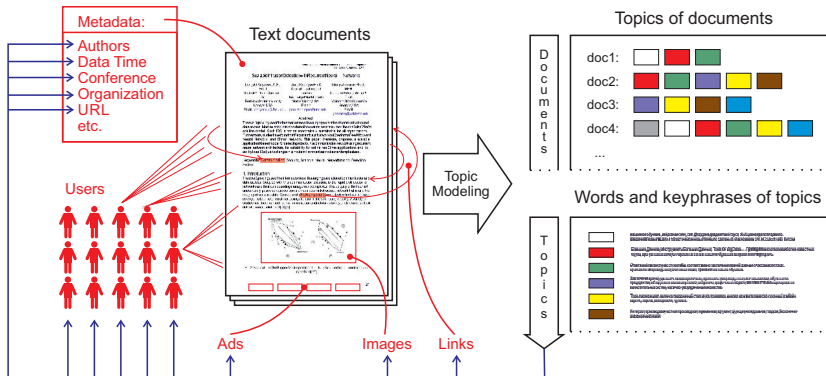
$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \varphi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$$

Декоррелирование тем:

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \varphi_{wt} \varphi_{ws}$$

Обобщение ARTM на мультимодальные задачи

Выявление тематики документов $p(t|d)$ и терминов $p(t|w)$, а также модальностей: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{изображение})$, $p(t|\text{пользователь})$, ...



Мультимодальная ARTM [Vorontsov et al, 2015]

W^m — словарь токенов m -й модальности, $m \in M$

$W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log-правдоподобий с регуляризацией:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} \text{Е-шаг:} & \quad \begin{cases} p_{tdw} = \text{norm}_{t \in T}(\varphi_{wt} \theta_{td}) \end{cases} \\ \text{М-шаг:} & \quad \begin{cases} \varphi_{wt} = \text{norm}_{w \in W^m} \left(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{aligned}$$

Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов,
- запросом может быть текст произвольной длины,
- информационной потребностью — систематизация знаний



Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

Разведочный тематический поиск

$q = (w_1, \dots, w_{n_q})$ — текст запроса произвольной длины n_q

$\theta_{tq} = p(t|q)$ — тематический профиль запроса q

$\theta_{td} = p(t|d)$ — тематические профили документов $d \in D$

Косинусная мера близости документа d и запроса q :

$$\text{sim}(q, d) = \frac{\sum_t \theta_{tq} \theta_{td}}{(\sum_t \theta_{tq}^2)^{1/2} (\sum_t \theta_{td}^2)^{1/2}}.$$

Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q, d)$

Выдача тематического поиска — k первых документов.

Реализация: *инвертированный индекс* для быстрого поиска документов d по каждой из тем t запроса

Данные коллективного блога Хабрахабр.ру

Данные

- 132 157 статей
- Модальности:
 - 52 354 терминов (слов)
 - 524 авторов статей
 - 10 000 комментаторов (авторов комментариев к статьям)
 - 2546 тегов
 - 123 хаба (категории)

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация rymorphy2

Методика оценивания качества разведочного поиска

Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

документы d с распределением $p(t|d)$,
близким к распределению $p(t|q)$ запроса

Два задания асессорам

- 1 найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- 2 оценить релевантность поисковой выдачи на том же запросе

Набор MapReduce

Набор MapReduce – программа поиска (**индексатор**) выполняет распределенное вычислений для больших объемов данных и раздает параллельные запросы, представляющие собой набор логических и исполнимых единиц для создания и обработки данных на параллельную обработку.

Основные компоненты Набор MapReduce можно сформулировать как:

- обработка вычислений больших объемов данных;
- масштабируемость;
- автоматическое распределение заданий;
- работа на неведомых оборудовании;
- автоматическая обработка отказов вычислений заданий.

Набор – популярная программа платформа (**обработка больших данных**) построена распределенных приложений для массово-параллельной обработки (**задачи**, **задачи**, **распределенная**, **МР**) данных.

Набор включает в себе следующие компоненты:

1. HDFS – распределенная файловая система.

2. **Набор MapReduce** – программа поиска (**индексатор**) выполняет распределенное вычислений для больших объемов данных и раздает параллельные запросы.

Клиенты, взаимодействующие в архитектуре **Набор MapReduce** в структуру HDFS, стали привычной роль, у них есть в своем компьютере, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничение платформ **Набор** в целом. К последним можно отнести:

Ограничение масштабируемости кластера **Набор** – АК вычислительных узлов, «АК параллельных заданий».

Сильная связность **Фреймворк** распределенных вычислений и клиентских вычислений, реализованных распределенных алгоритмов. Как следствие:

Отсутствие поддержки альтернативной программы поиска вычислений распределенных вычислений в **Набор v1.0** поддерживается только только вычислений параллельно.

Многие клиенты, такие, как клиенты, неспособность исполнения в среде с вычислениями требующими в надежности;

Проблемы **вычислений** совместности: требование по клиентскому обслуживанию всех вычислительных узлов кластера при обслуживании платформ **Набор** (установка новых версий или замена оборудования).

Пример запроса для
разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ..

Релевантные тексты: примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

Нерелевантные тексты: общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

Тематика запросов разведочного поиска

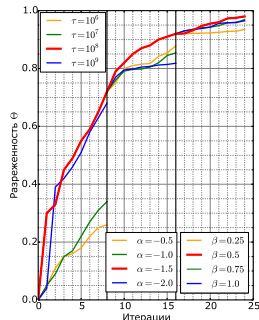
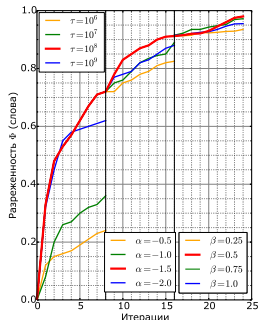
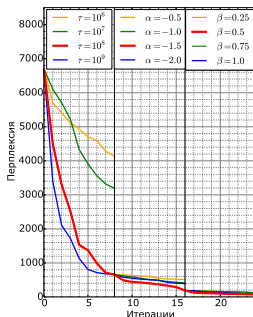
Примеры заголовков разведочных запросов к Хабру
(объём каждого запроса — около одной страницы A4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	АВ-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

Подбор коэффициентов регуляризации

Последовательное добавление регуляризаторов:

- декоррелирование распределений терминов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений терминов в темах (β).



Оценки качества поиска

Precision — доля релевантных среди найденных

Recall — доля найденных среди релевантных

$$P = \frac{TP}{TP + FP} \text{ — точность (precision)}$$

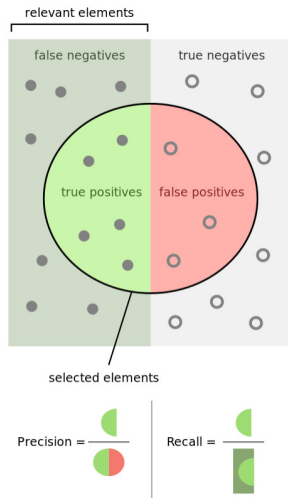
$$R = \frac{TP}{TP + FN} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2} \text{ — F1-мера}$$

TP (true positive) — найденные релевантные

FP (false positive) — найденные нерелевантные

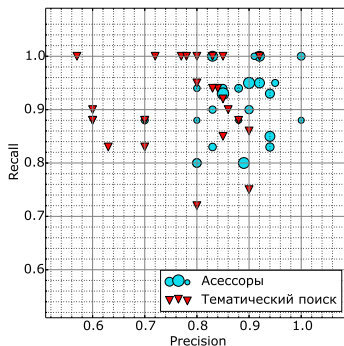
FN (false negative) — ненайденные релевантные



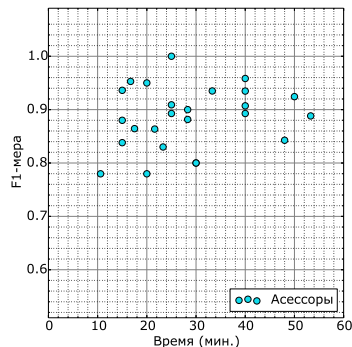
Результаты измерения точности и полноты по запросам

25 запросов, 3 ассессора на запрос

точность и полнота поиска



время и F_1 -мера (ассессоры)



- среднее время обработки запроса ассессором — 30 минут
- точность выше у ассессоров, полнота — у поисковика

Выбор модальностей по критериям точности и полноты

Habrahabr. Число тем $|T| = 200$.

Модальности: Слова, Авторы, Комментаторы, Теги, Хабы.

	асессоры	С	К	ТХ	СТ	СХ	СТХ	все
Pr@5	0.82	0.63	0.54	0.59	0.74	0.73	0.73	0.74
Pr@10	0.87	0.67	0.56	0.58	0.77	0.74	0.75	0.77
Pr@15	0.86	0.65	0.53	0.55	0.67	0.67	0.68	0.68
Pr@20	0.85	0.64	0.53	0.54	0.66	0.67	0.68	0.68
Re@5	0.78	0.77	0.63	0.69	0.82	0.81	0.82	0.82
Re@10	0.84	0.79	0.64	0.71	0.88	0.82	0.87	0.88
Re@15	0.88	0.82	0.67	0.73	0.90	0.84	0.89	0.90
Re@20	0.88	0.85	0.68	0.74	0.91	0.85	0.89	0.91

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и теги

Выбор модальностей по критериям точности и полноты

TechCrunch. Число тем $|T| = 450$.

Модальности: Слова, Биграмммы, Категории, Авторы.

	асессоры	С	К	СБ	СБК	СБКА
Pr@5	0.83	0.71	0.55	0.77	0.80	0.80
Pr@10	0.88	0.72	0.58	0.78	0.81	0.81
Pr@15	0.87	0.73	0.59	0.79	0.83	0.83
Pr@20	0.86	0.72	0.56	0.77	0.82	0.82
Re@5	0.81	0.75	0.65	0.77	0.82	0.83
Re@10	0.85	0.77	0.66	0.80	0.85	0.86
Re@15	0.89	0.78	0.68	0.82	0.87	0.91
Re@20	0.90	0.82	0.69	0.83	0.89	0.93

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и категории

Выбор числа тем по критериям точности и полноты

Habrahabr. Используем все 5 модальностей, меняем $|T|$

	асессоры	100	200	300	400	500
Pr@5	0.82	0.61	0.74	0.71	0.69	0.59
Pr@10	0.87	0.65	0.77	0.72	0.67	0.61
Pr@15	0.86	0.67	0.68	0.67	0.65	0.62
Pr@20	0.85	0.64	0.68	0.67	0.64	0.60
Re@5	0.78	0.62	0.82	0.80	0.72	0.63
Re@10	0.84	0.63	0.88	0.81	0.75	0.64
Re@15	0.88	0.67	0.90	0.82	0.77	0.67
Re@20	0.88	0.69	0.91	0.85	0.77	0.68

- Наилучшее качество поиска — при 200 темах
- Тематический поиск превосходит асессоров по полноте

Выбор числа тем по критериям точности и полноты

TechCrunch. Используем все 4 модальности, меняем $|T|$

	асессоры	350	400	450	475	500
Pr@5	0.83	0.65	0.72	0.75	0.80	0.68
Pr@10	0.88	0.66	0.73	0.76	0.81	0.69
Pr@15	0.87	0.68	0.74	0.78	0.82	0.68
Pr@20	0.86	0.65	0.74	0.77	0.81	0.67
R@5	0.81	0.65	0.75	0.78	0.83	0.79
R@10	0.85	0.66	0.78	0.79	0.86	0.80
R@15	0.89	0.68	0.79	0.79	0.91	0.83
R@20	0.90	0.69	0.79	0.80	0.93	0.85

- Наилучшее качество поиска — при 475 темах
- Тематический поиск превосходит асессоров по полноте

Выбор меры близости документа и запроса

Меры близости распределений:

Euclidean, Cosine, Manhattan, Kullback-Leibler

	<i>Habrahabr</i> , $ T = 200$				<i>TechCrunch</i> , $ T = 450$			
	E	C	M	KL	E	C	M	KL
Pr@5	0.61	0.74	0.68	0.72	0.63	0.80	0.67	0.71
Pr@10	0.65	0.77	0.69	0.75	0.66	0.81	0.68	0.73
Pr@15	0.62	0.68	0.63	0.70	0.64	0.82	0.64	0.72
Pr@20	0.62	0.68	0.62	0.70	0.64	0.81	0.63	0.71
Re@5	0.67	0.82	0.69	0.80	0.66	0.83	0.67	0.77
Re@10	0.68	0.88	0.70	0.85	0.67	0.86	0.68	0.78
Re@15	0.70	0.90	0.72	0.87	0.71	0.91	0.70	0.80
Re@20	0.70	0.91	0.73	0.88	0.71	0.93	0.71	0.81

- Наилучшее качество поиска — при косинусной мере

Все ли регуляризаторы были нужны?

Декоррелирование, Разреживание

	<i>Habrahabr</i>					<i>TechCrunch</i>				
	асессоры	все	ДР	Д	нет	асессоры	все	ДР	Д	нет
Pr@5	0.82	0.74	0.69	0.58	0.52	0.83	0.80	0.71	0.57	0.54
Pr@10	0.87	0.77	0.70	0.59	0.55	0.88	0.81	0.72	0.59	0.55
Pr@15	0.86	0.68	0.65	0.56	0.53	0.87	0.82	0.68	0.58	0.54
Pr@20	0.85	0.68	0.65	0.55	0.52	0.86	0.81	0.68	0.58	0.54
Re@5	0.78	0.82	0.75	0.63	0.59	0.81	0.81	0.76	0.65	0.60
Re@10	0.84	0.88	0.76	0.65	0.60	0.85	0.86	0.78	0.66	0.62
Re@15	0.88	0.90	0.77	0.66	0.61	0.89	0.89	0.81	0.64	0.63
Re@20	0.88	0.91	0.77	0.66	0.61	0.90	0.92	0.82	0.64	0.63

- Все регуляризаторы необходимы

Сравнение с поиском по векторам TF-IDF слов

Поиск по векторам TF-IDF($w|d$) = $\frac{n_{dw}}{\ln N_w}$

	<i>Habrahabr</i>			<i>TechCrunch</i>		
	ассесоры	topic	tf-idf	assessors	topic	tf-idf
Pr@5	0.82	0.74	0.76	0.83	0.80	0.78
Pr@10	0.87	0.77	0.77	0.88	0.81	0.79
Pr@15	0.86	0.68	0.72	0.87	0.82	0.76
Pr@20	0.85	0.68	0.71	0.86	0.81	0.75
Re@5	0.78	0.82	0.76	0.81	0.81	0.77
Re@10	0.84	0.88	0.77	0.85	0.86	0.78
Re@15	0.88	0.90	0.80	0.89	0.89	0.80
Re@20	0.88	0.91	0.81	0.90	0.92	0.83

- Тематический поиск немного лучше TF-IDF
- При этом поисковый индекс на 2–3 порядка компактнее

Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA, 2016.

Выводы

О данной задаче:

- для построения моделей использовался BigARTM
- аккуратная настройка модели даёт хороший поиск
- но для прорывного результата качество недостаточно

О прикладных исследованиях в машинном обучении:

- любое исследование — это сравнение вариантов
- комбинирование регуляризаторов — общий приём для учёта большого числа требований в одной модели
- многокритериальное обучение:
 - 1) выбрать поэтапную стратегию регуляризации
 - 2) подобрать коэффициент регуляризации на каждом этапе
- ассессоры оценивали данные, а не модель, поэтому удалось сравнить много моделей