МАШИННОЕ ОБУЧЕНИЕ
• распознавание языка текста •
• диагностика по электрокардиограмме •

Воронцов Константин Вячеславович
(лаборатория Машинного интеллекта МФТИ)

• Прикладной анализ данных •
(кружок для старшеклассников)
11 марта 2018 • МФТИ

Содержание

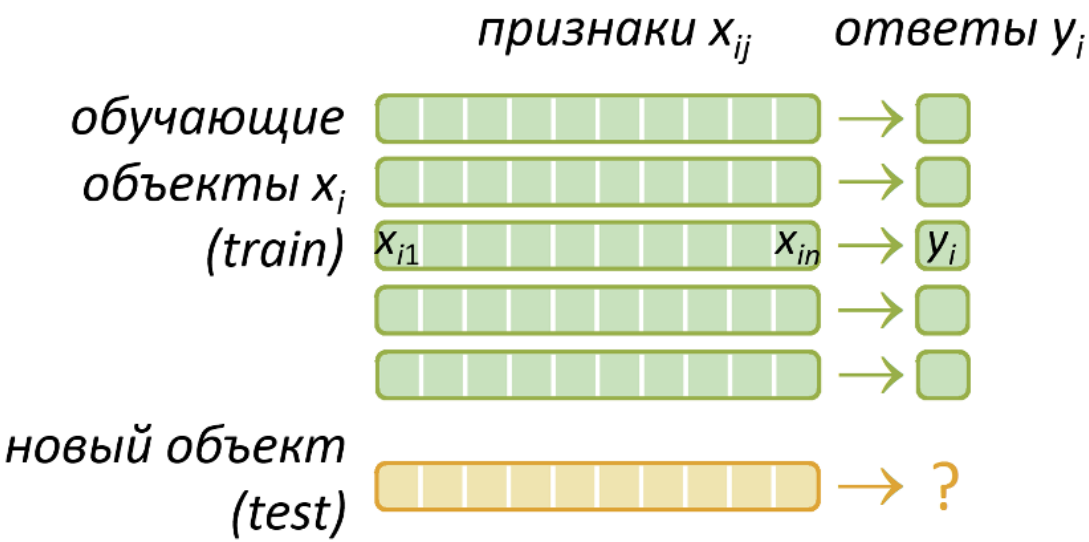
Напоминание. Задача классификации

Обозначения:

x_i — объекты обучающей выборки, $i = 1, \ldots, \ell$

y_i — ответ на объекте x_i

x_{ij} — значение j-го признака объекта x_i



Задача: восстановить зависимость $y(x)$

Декларация прав человека. На каких языках?

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

rus: Russian

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

ukr: Ukrainian

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

eng: English

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

frn: French

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

ger: German

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

afk: Afrikaans

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

spn: Spanish

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

por: Portuguese

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Декларация прав человека. На каких языках?

fin: Finnish

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

est: Estonian

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

swd: Swedish

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

nrn: Norwegian

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Задача распознавания языка текста (Language Identification)

Почему нам удаётся распознавать язык,
даже когда мы его почти не знаем?

Как обучить машину определять язык текста автоматически?

Зачем это нужно:

- ~~Это просто прикольно~~
- Поисковые системы
- Системы агрегации контента
- Системы автоматического перевода

Постановка задачи

Дано:
обучающая выборка текстов с известными *классификациями*:

$$\langle \text{текст}_1, \text{язык}_1 \rangle, \langle \text{текст}_2, \text{язык}_2 \rangle, \ldots, \langle \text{текст}_\ell, \text{язык}_\ell \rangle$$

Найти:
Правило (функцию, алгоритм) классификации любого текста

$$\text{текст} \xrightarrow{?} \text{язык}$$

Критерий качества решения:
Алгоритм должен как можно реже ошибаться.

Задача поставлена, когда у неё есть Д.Н.К.

Математическая модель классификации текстов

Каждый язык имеет уникальное распределение частот n-грамм.

Все люди рождаются свободными и равными в своем достоинстве и...

л, ю, д, и — униграммы

лю, юд, ди — биграммы

люд, юди — триграммы

Линейная модель классификации:

Оценка принадлежности текста x языку y по n-граммам j:

$$a_y(x) = \sum_j w_{jy} x_j,$$

признак x_j — частота n-граммы j,

параметры модели w_{jy} — важность n-граммы j для языка y.

Правило классификации:

отнести текст x к тому языку y, для которого $a_y(x)$ максимально.

Эксперимент с текстами Декларации прав человека

Цели эксперимента — проверить:

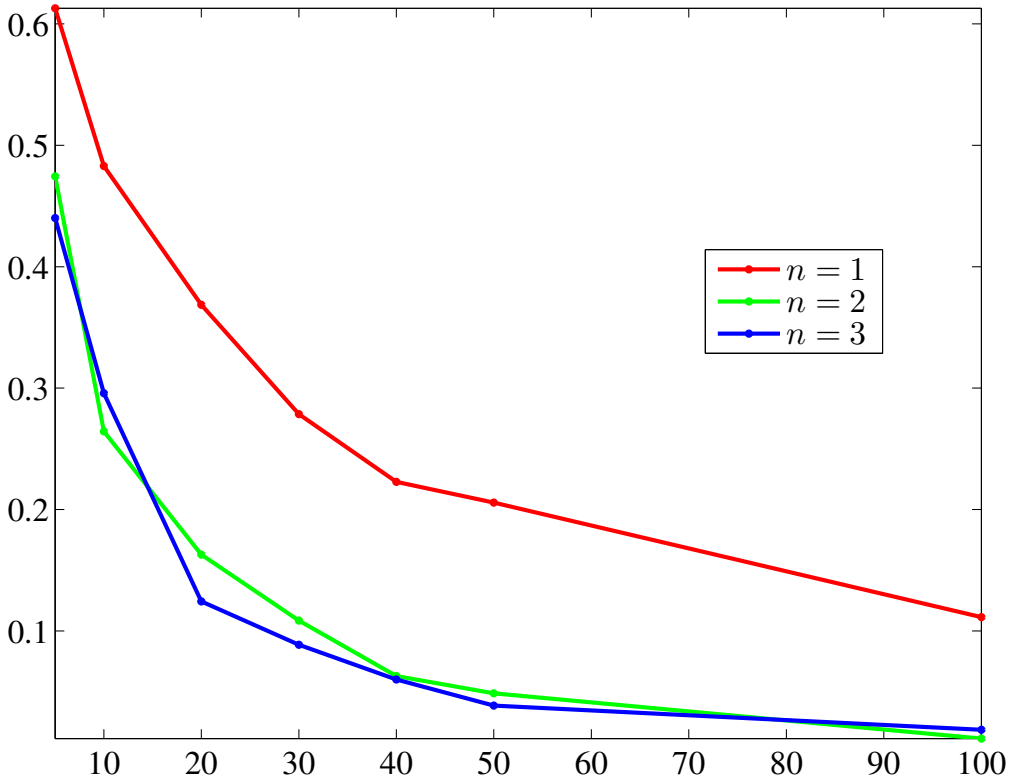
- действительно ли частоты триграмм распознают язык?
- как точность распознавания зависит от длины текста?

Методика эксперимента:

- используем тексты Декларации на 7 языках ($\sim 10^4$ букв)
- в каждом тексте случайно выбираем обучающий фрагмент длины ℓ и контрольный фрагмент длины k
- веса n-грамм определяем по обучающему фрагменту текста длины ℓ для каждого языка
- точность измеряем как долю ошибочных распознаваний языка по контрольным фрагментам

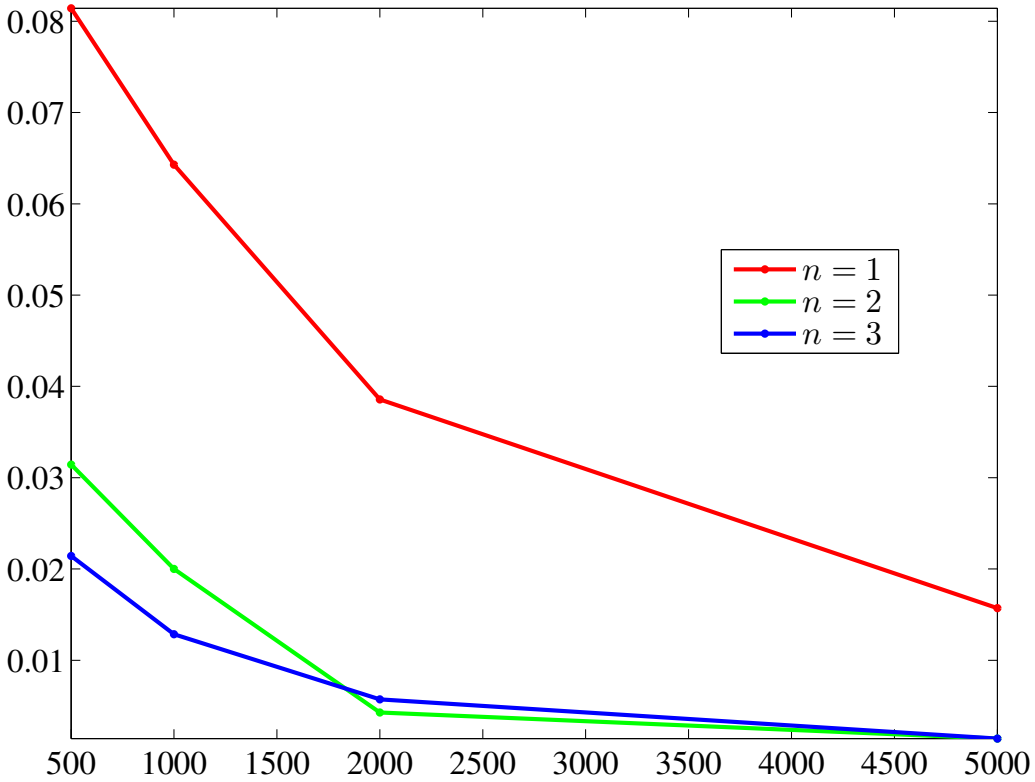
Результаты эксперимента

Зависимость доли ошибок на контроле от длины контрольных текстов для 1-,2-,3-грамм
(длина обучающих текстов 2000 символов)



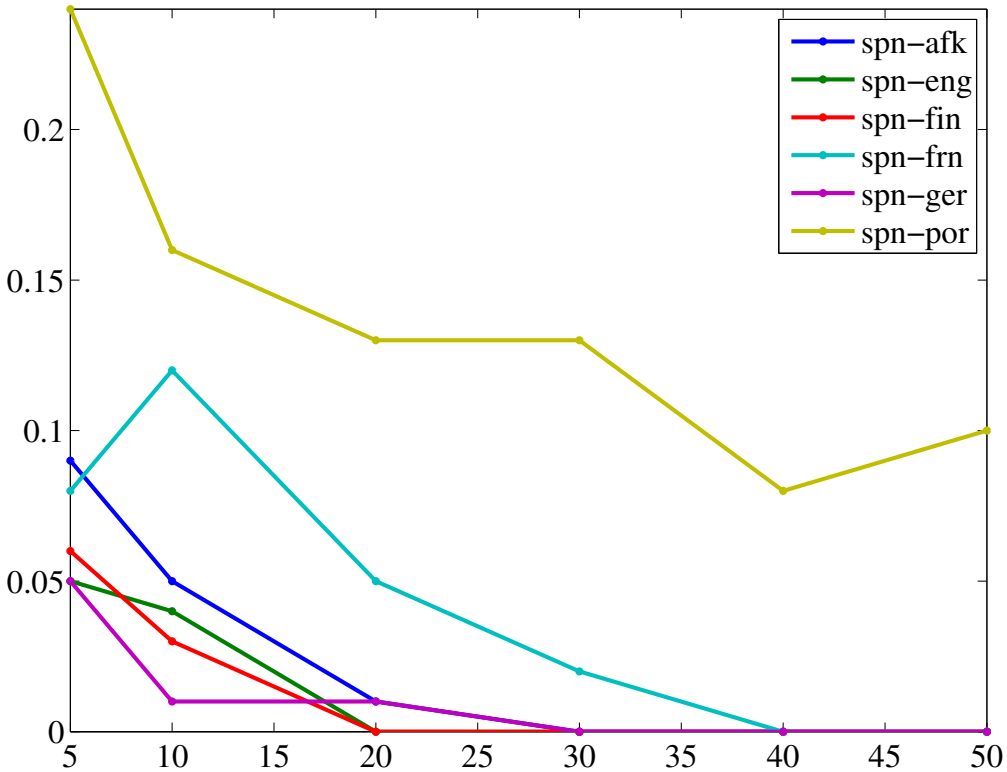
Результаты эксперимента

Зависимость доли ошибок на контроле от длины обучающих текстов для 1-,2-,3-грамм
(длина контрольной выборки 200 символов)



Результаты эксперимента

По оси X — длина контрольной выборки
По оси Y — доля случаев, когда испанский язык был перепутан с другим языком
(3-граммы, длина обучающих текстов 2000 символов)



Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 10)

	afk	eng	fin	frn	ger	por	spn
afk	69	7	9	6	14	1	5
eng	6	75	3	4	2	1	4
fin	4	1	82	3	0	1	3
frn	3	4	1	66	1	5	12
ger	15	4	1	4	80	1	1
por	1	6	2	5	1	62	16
spn	2	3	2	12	2	29	59

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 50)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	98	0	1	1	0	0
fin	0	0	100	0	0	0	0
frn	0	1	0	98	1	1	0
ger	0	1	0	0	98	0	0
por	0	0	0	0	0	89	10
spn	0	0	0	1	0	10	90

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 100)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	100	0	0	0	0	0
fin	0	0	100	0	0	0	0
frn	0	0	0	100	0	0	0
ger	0	0	0	0	100	0	0
por	0	0	0	0	0	92	5
spn	0	0	0	0	0	8	95

Результаты эксперимента

В столбцах — истинные классы, в строках — предсказанные
В ячейках — число случаев (из общего числа $100 \cdot 7 = 700$)
(3-граммы, длина обучения 2000, длина контроля 1000)

	afk	eng	fin	frn	ger	por	spn
afk	100	0	0	0	0	0	0
eng	0	100	0	0	0	0	0
fin	0	0	100	0	0	0	0
frn	0	0	0	100	0	0	0
ger	0	0	0	0	100	0	0
por	0	0	0	0	0	99	0
spn	0	0	0	0	0	1	100

20 самых частых триграмм в 7 языках

В этом эксперименте:

- использовались только языки на основе латиницы,
- все диакритические знаки и пробел были заменены на «-»

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
afk	ie-	die	-di	en-	ing	ng-	an-	et-	-re	reg	eg-	e-r	-en	nie	van	-ni	een	el-	e-o	n-h
eng	-an	and	nd-	the	-th	he-	ion	of-	-of	tio	al-	to-	-to	on-	ent	ati	-in	e-e	ll-	t-t
fin	ise	sta	an-	en-	ta-	ais	aan	la-	ell	ist	ike	kai	keu	oik	-ta	lla	on-	tai	-oi	ast
frn	-de	es-	de-	le-	et-	ion	nt-	tio	-et	te-	ent	e-d	e-p	ne-	on-	ati	a-d	e-s	la-	oit
ger	en-	ein	er-	der	ine	nd-	cht	ung	-un	ich	und	ech	gen	ht-	ng-	sei	ver	-ei	-ha	-se
por	de-	-de	os-	-e-	em-	o-d	to-	-a-	-di	dir	-co	-pe	ire	as-	ito	o-e	-se	eit	ess	e-d
spn	os-	-de	-la	de-	la-	-y-	es-	-a-	ent	ien	en-	al-	as-	ere	e-l	-el	-lo	cia	el-	los

Выводы

- язык текста можно распознавать автоматически,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- точность распознавания быстро увеличивается с ростом длины текста — сотни символов хватает для распознавания даже близких языков
- для профессионального решения дополнительно используют словари слов и аббревиатур

Линейная модель классификации

Оценка принадлежности объекта x классу y по признакам j:

$$a_y(x) = \sum_j w_{jy} x_j = \langle w_j, x \rangle,$$

где w_{jy} — вес (важность) признака (n-граммы) j для класса y.

Правило классификации:
отнести x к тому классу y, для которого $a_y(x)$ максимально.

Эвристика: чем выше в среднем признак j у объектов класса y и ниже у остальных объектов, тем он важнее для класса y:

$$w_{jy} = S_{jy} = \frac{\sum_i [y_i = y] x_{ij}}{\sum_i [y_i = y]}$$

Несколько полезных эвристик

1. Бинаризация признаков: важно, что триграмма часто встречается, но не так важно, *настолько* часто.
Отсюда эвристика — использовать $[x_{ij} \geqslant A]$ вместо x_{ij},
где A — параметр, который придётся подбирать.

2. Варианты формулы весов w_{jy}:

$$w_{jy} = S_{jy} \qquad w_{jy} = \log S_{jy} \qquad w_{jy} = \sqrt{S_{jy}}$$
$$w_{jy} = S_{jy} - S_{j\bar{y}} \qquad w_{jy} = \log S_{jy} - \log S_{j\bar{y}} \qquad w_{jy} = \sqrt{S_{jy}} - \sqrt{S_{j\bar{y}}}$$

где $S_{j\bar{y}}$ — среднее x_{ij} по объектам НЕ из класса y.

3. Отбор признаков (синдромный алгоритм):
отсортировать признаки по убыванию весов $\max\limits_{y} |w_{jy}|$;
взять первые K признаков; для остальных положить $w_{jy} = 0$.

Терминология диагностики (два класса: 1–больной, 0–здоровый)

Положительный диагноз — алгоритм предсказывает болезнь
(хотя, казалось бы, что тут положительного...)

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:

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

Чувствительность и специфичность хотим максимизировать.

⊕ Они не зависят от соотношения мощностей классов.

⊕ Хорошо подходят для несбалансированных выборок.

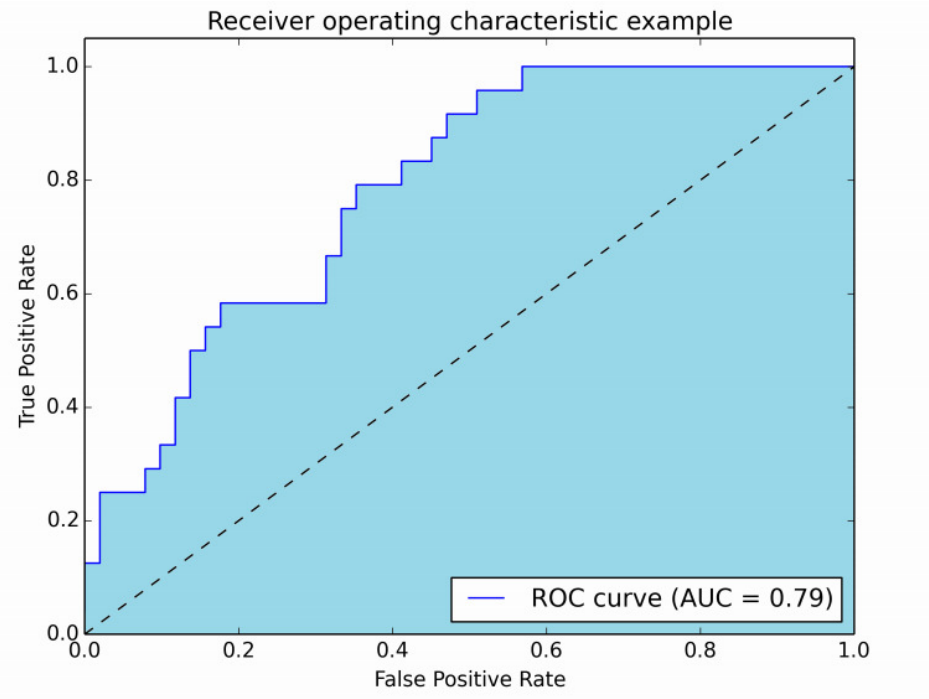
Определение ROC-кривой

Классификатор: $a(x) = \big[\langle w_1 - w_0, x\rangle > \theta\big] = \big[\langle w, x\rangle > \theta\big]$,

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

по оси Y: чувствительность = TPR, True Positive Rate

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



AUC — площадь под ROC-кривой

Классификатор: $a(x) = \big[\langle w_1 - w_0, x\rangle > \theta\big] = \big[\langle w, x\rangle > \theta\big]$,

AUC равна доле пар объектов (x_i, x_j) из разных классов $(y_i = 0,\ y_j = 1)$ с правильным порядком ответов (докажите):

$$\text{AUC} = \frac{\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\big[y_i < y_j\big]\big[\langle x_i, w\rangle < \langle x_j, w\rangle\big]}{\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\big[y_i < y_j\big]}$$

Преимущества AUC:

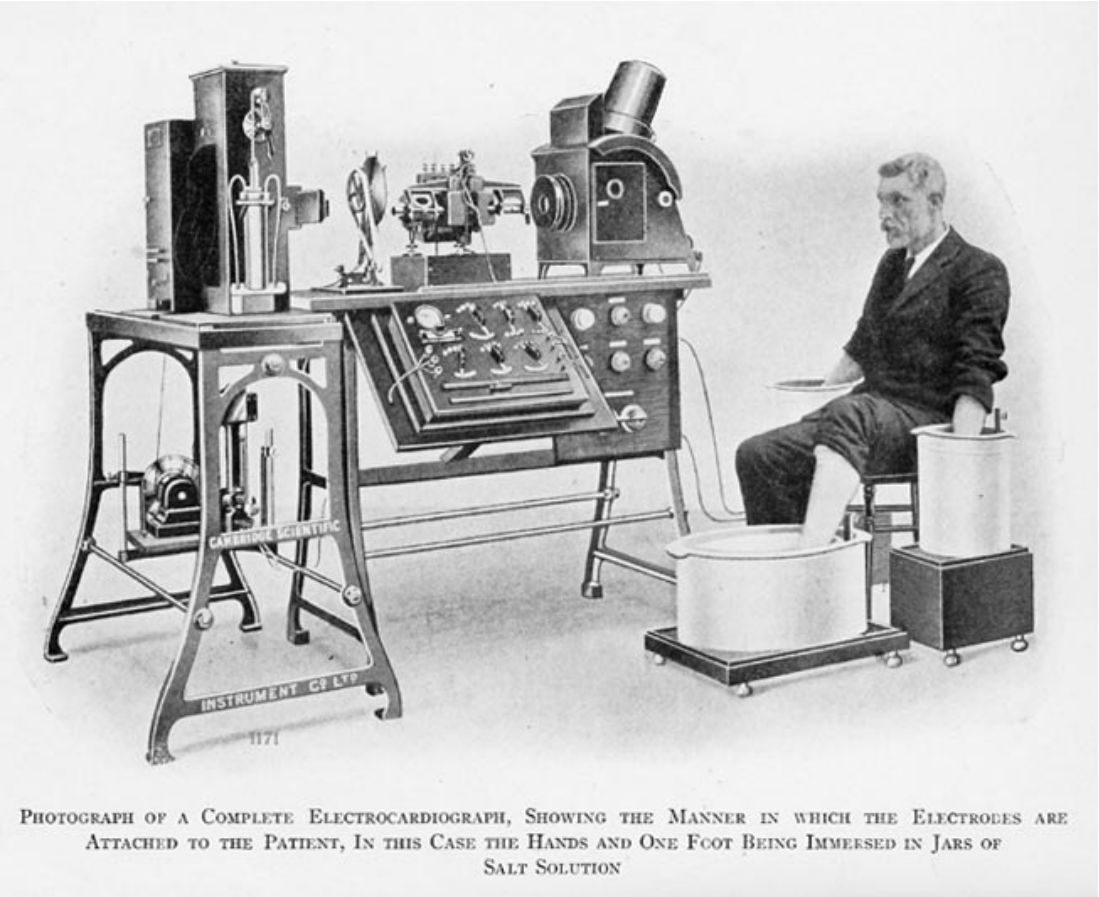
⊕ не зависит от порога θ, оценивает только качество w;

⊕ не зависит от численности классов;

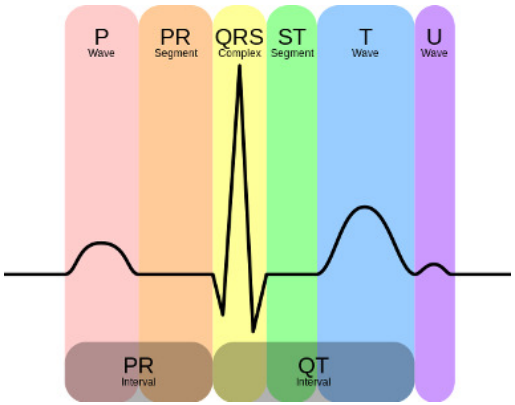
⊕ это общепринятая мера качества классификации;

Чтобы измерить предсказательную способностью μ, AUC вычисляют на контрольной выборке.

Электрокардиография



Photograph of a Complete Electrocardiograph, Showing the Manner in which the Electrodes are Attached to the Patient, In this Case the Hands and One Foot Being Immersed in Jars of Salt Solution

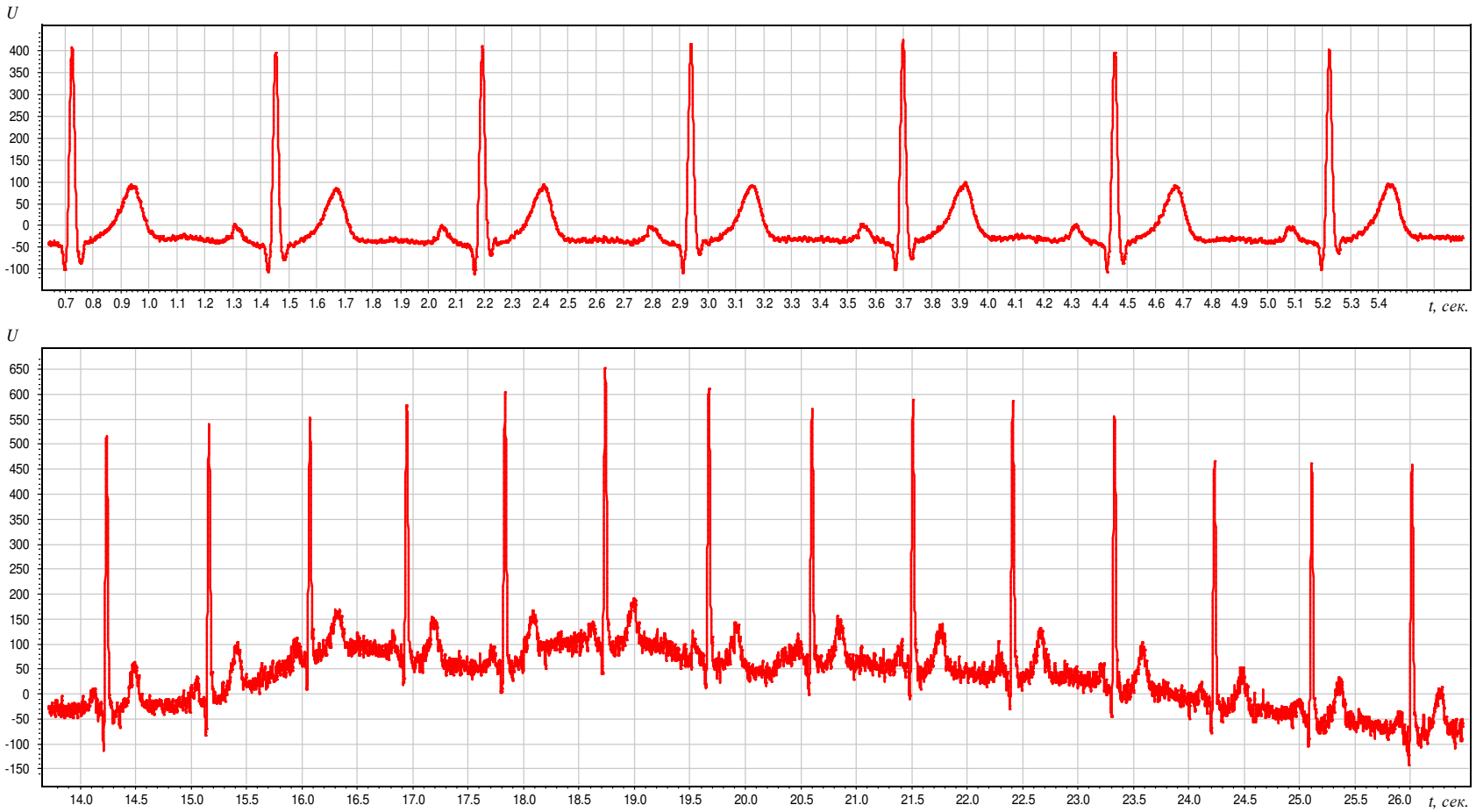


1872 — первые записи электрической активности сердца

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]

Предпосылки:

- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Появление цифровой электрокардиографии

Предположения:

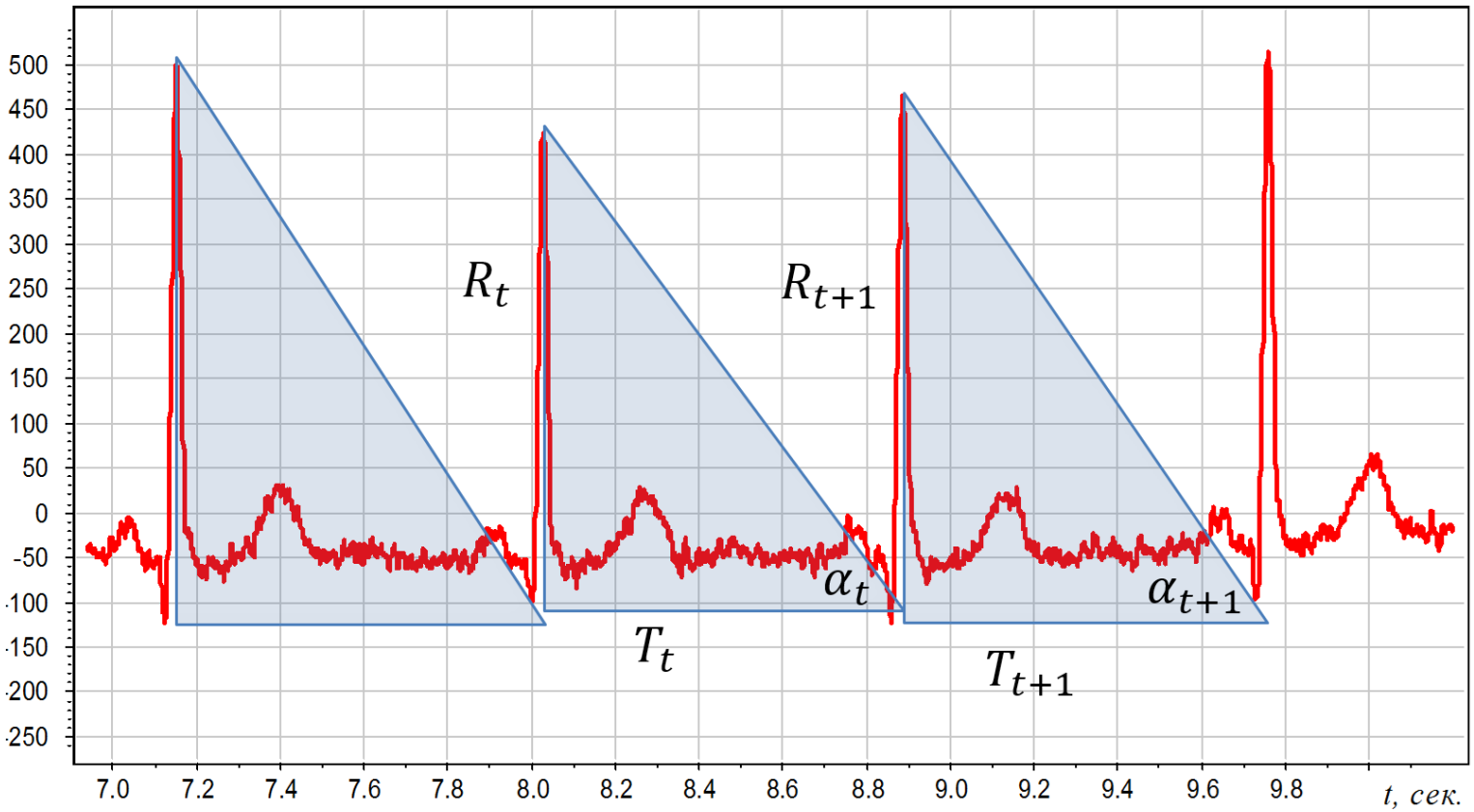
- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Каждое заболевание по-своему изменяет ЭКГ-сигнал
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика

Вариабельность интервалов и амплитуд кардиоциклов

приращение амплитуд: $dR_t = R_{t+1} - R_t$

приращение интервалов: $dT_t = T_{t+1} - T_t$

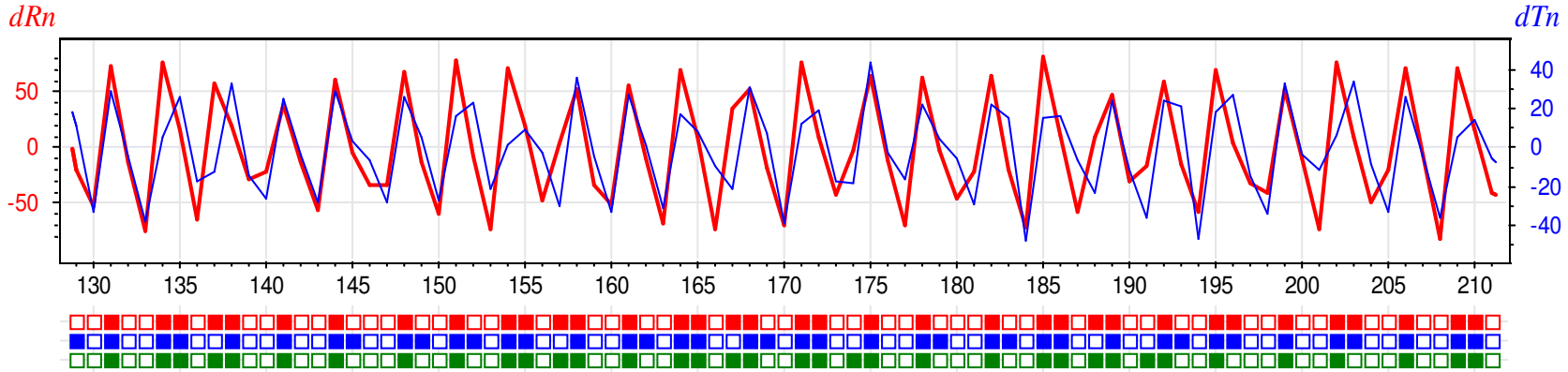
приращение углов: $d\alpha_t = \alpha_{t+1} - \alpha_t, \quad \alpha_t = \operatorname{arctg} \frac{R_t}{T_t}$



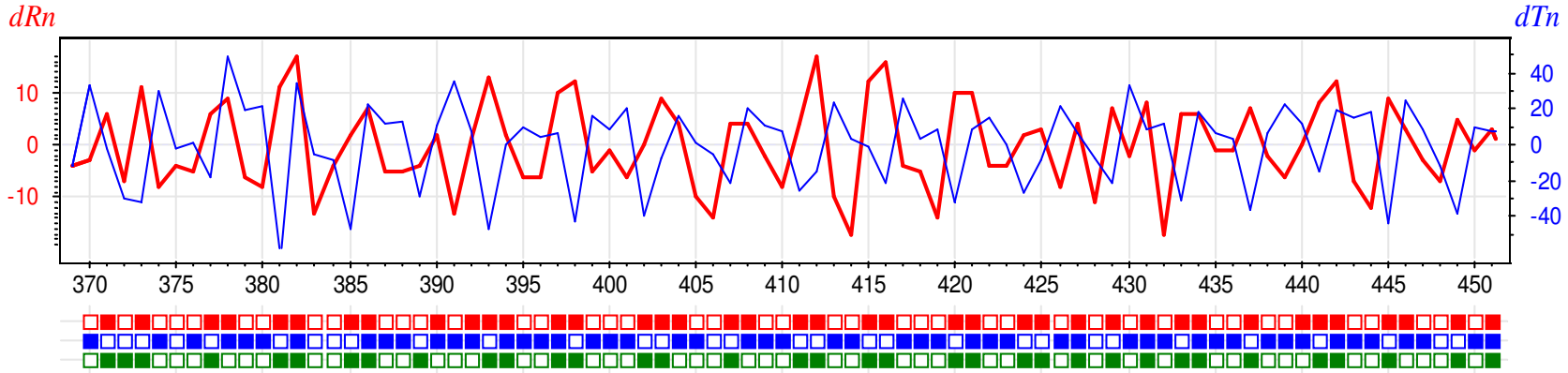
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:



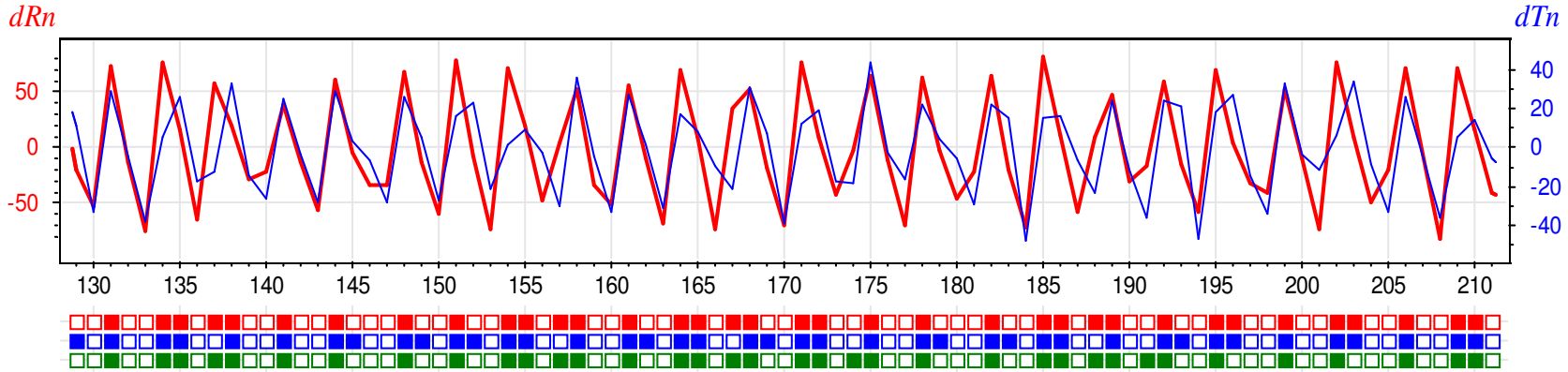
Больной (язвенная болезнь):



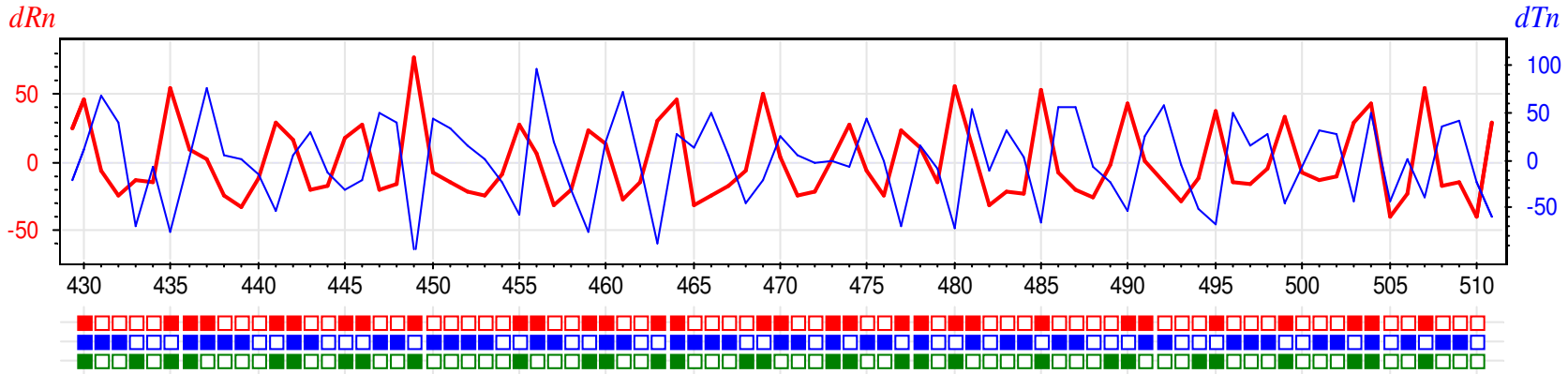
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:



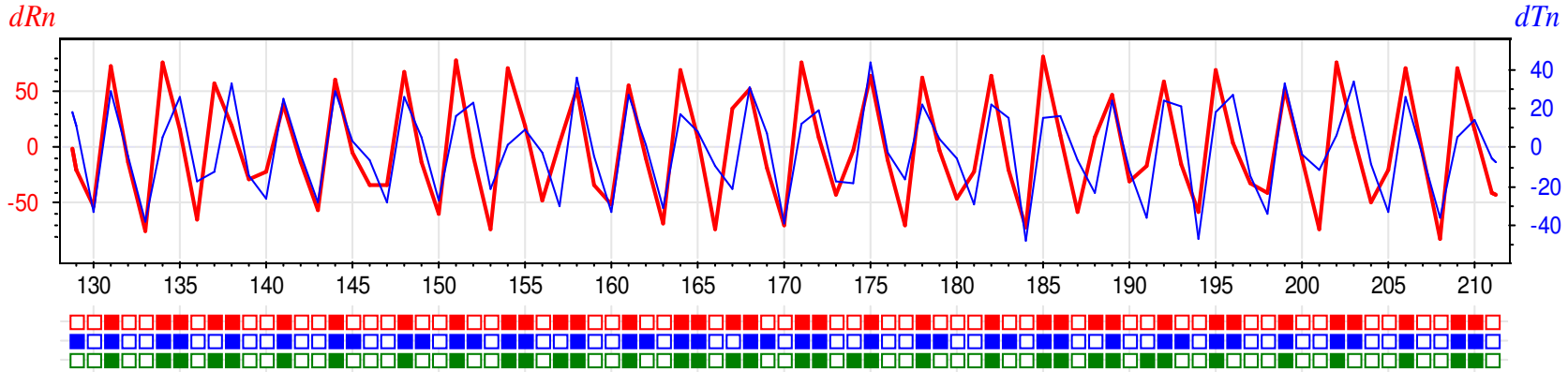
Больной (гипертония):



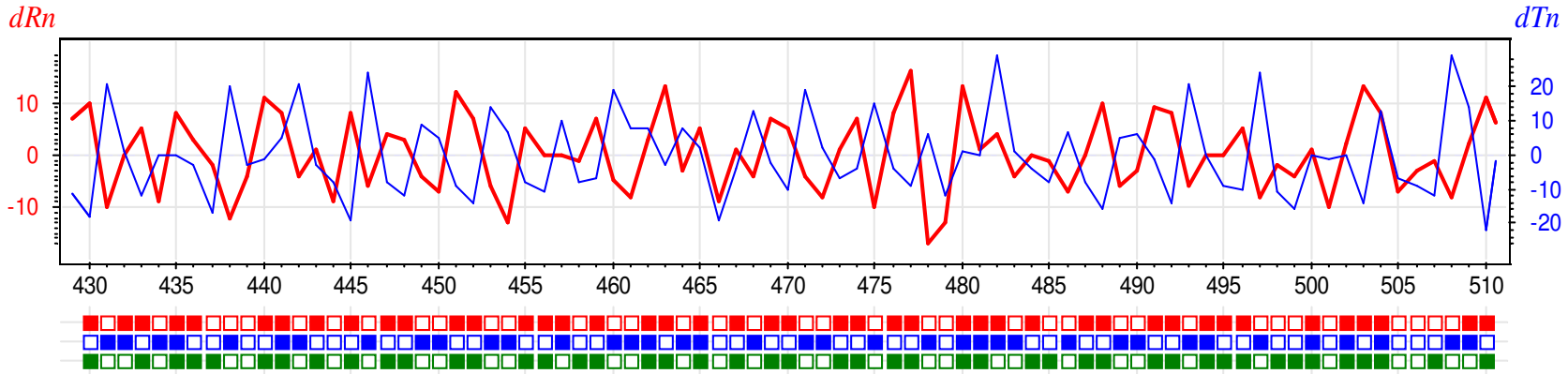
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:

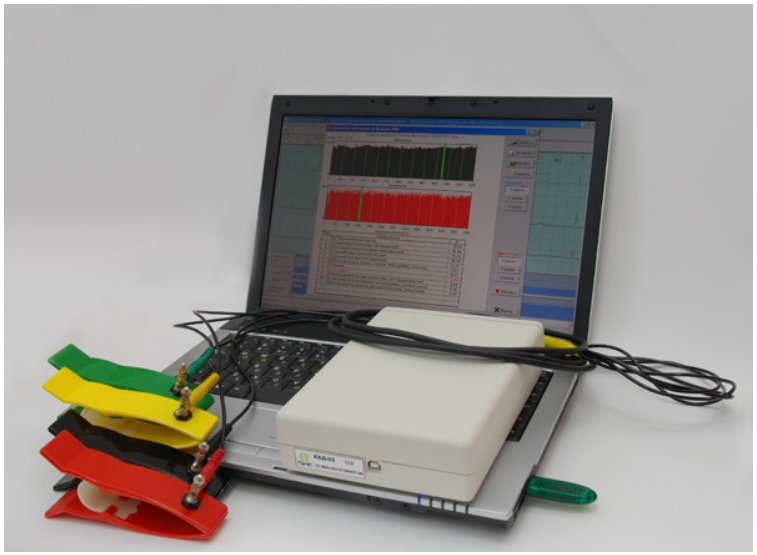


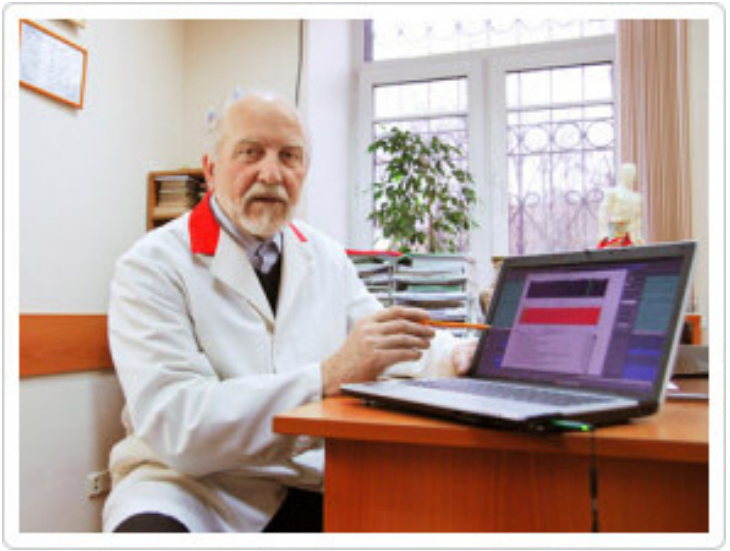
Больной (рак):



Диагностическая система «Скринфакс»

Цифровой электрокардиограф с улучшенной помехозащищённостью и расширенной полосой пропускания.





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

Объём исходных данных (по заболеваниям)

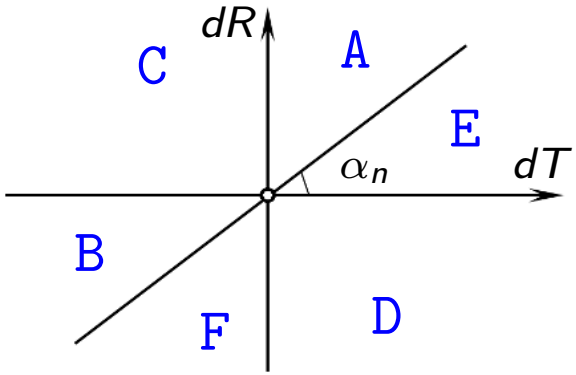
абсолютно здоровые	АЗ	193
гипертоническая болезнь	ГБ	1894
ишемическая болезнь сердца	ИБС	1265
сахарный диабет (СД1 и СД2)	СД	871
язвенная болезнь	ЯБ	785
миома матки	ММ	781
узловой (диффузный) зоб щитовидный железы	УЩ	748
дискинезия желчевыводящих путей	ДЖВП	717
хронический гастрит (гастродуоденит) гипоацидный	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 железодефицитная	ЖДА	260

Дискретизация ЭКГ-сигнала

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

Правила кодирования:

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



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

DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE

Векторизация ЭКГ-сигнала

По ЭКГ строится текстовая строка — *кодограмма*:

Частоты триграмм — число вхождений триграммы в кодограмму:

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

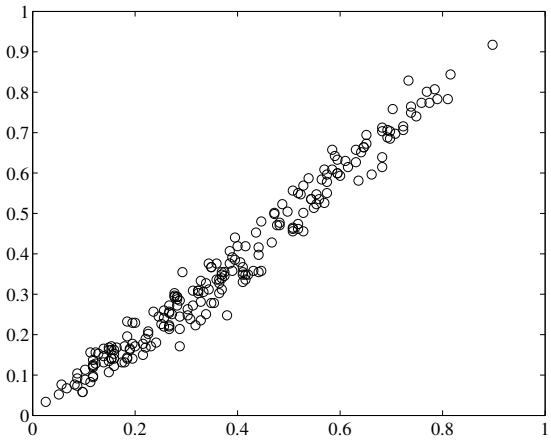
Существуют сочетания триграмм, специфичные для болезней

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

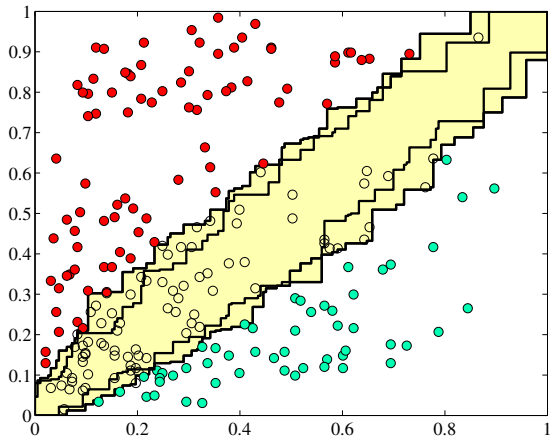
— ось X: доля здоровых x_i с частотой триграммы $x_i^j \geqslant 2$ из 600

— ось Y: доля больных x_i с частотой триграммы $x_i^j \geqslant 2$ из 600

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i



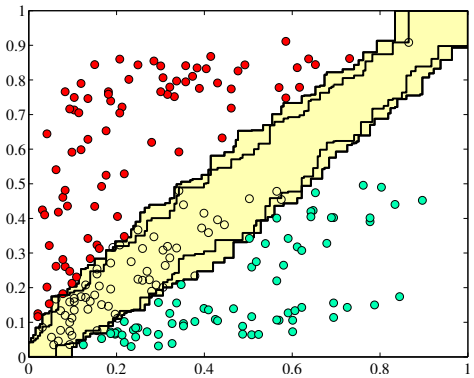
наблюдаемые y_i

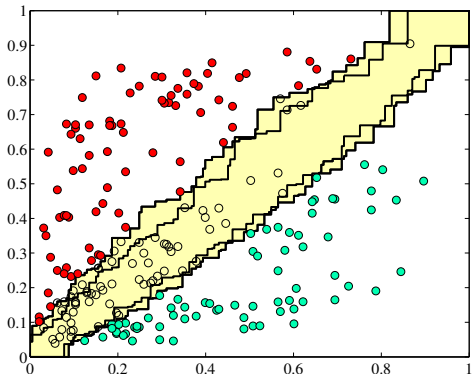
Слева: как распределятся точки, если объектам x_i назначить случайные (случайно перемешанные) метки классов y_i.

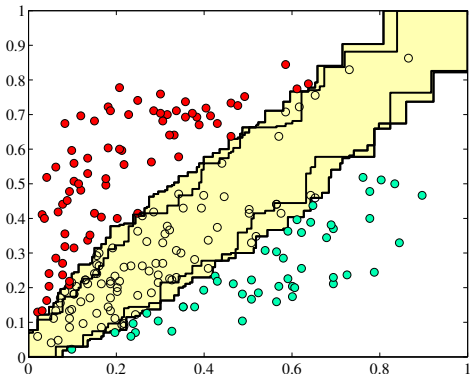
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

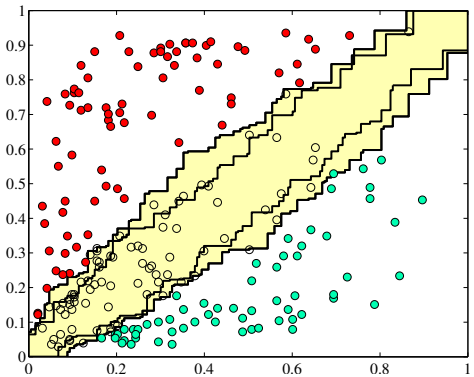
Существуют сочетания триграмм, специфичные для болезней

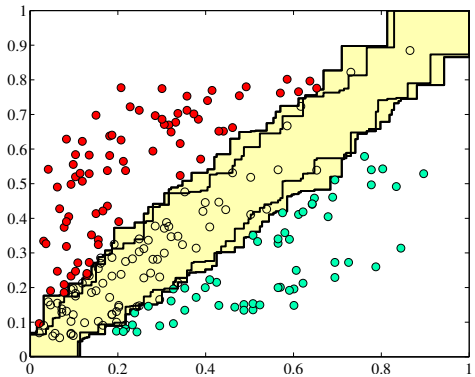
Для каждой болезни есть свои неслучайно частые триграммы

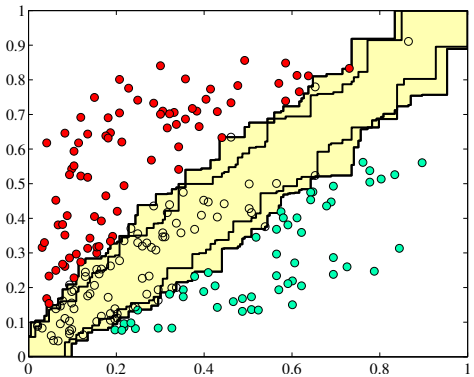

ишемия сердца


гипертония


рак


желчнокаменная болезнь

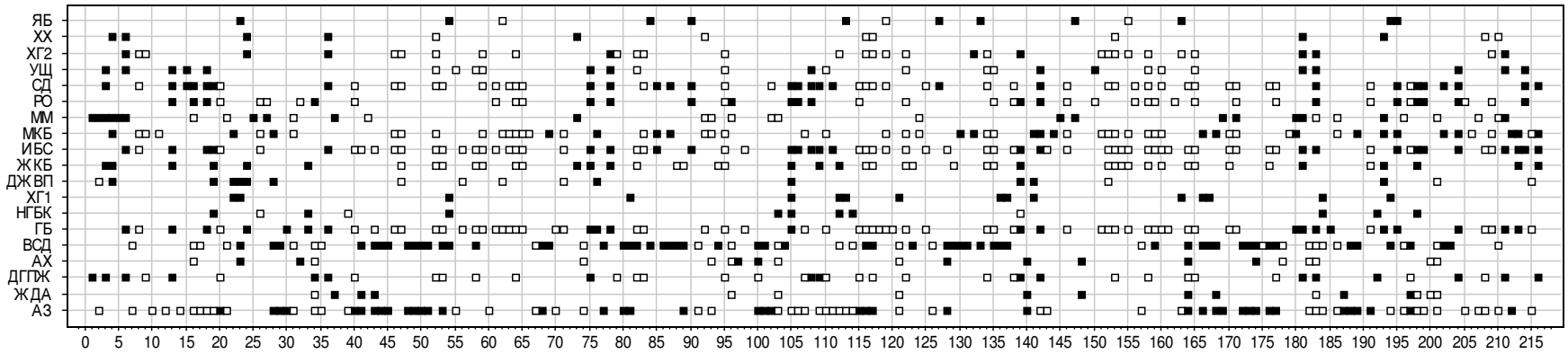

миома матки


язвенная болезнь

Болезни отличаются наборами информативных триграмм

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

ось Y: болезни (АЗ — абсолютно здоровые)



□ — неслучайно низкая частота триграммы

■ — неслучайно высокая частота триграммы

Вывод: для каждой болезни есть триграммы с неслучайно высокой и неслучайно низкой частотой

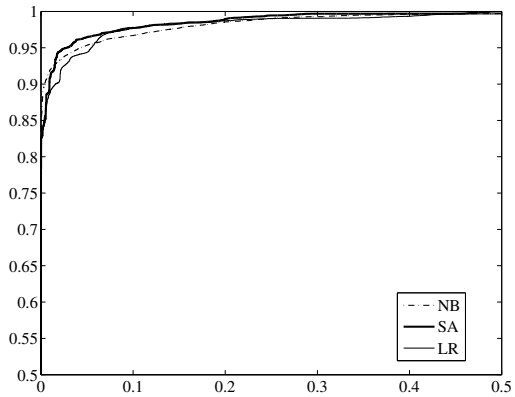
Диагностический эталон болезни — специфичное подмножество триграмм с неслучайно высокой частотой

Результаты кросс-валидации

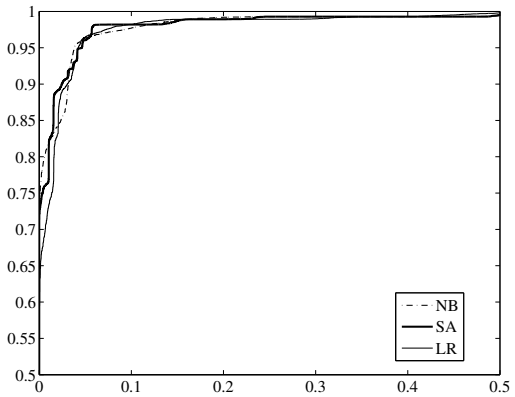
Обучающая выборка — для оптимизации параметров модели
Тестовая выборка — для оценивания чувс., спец., AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	С% 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

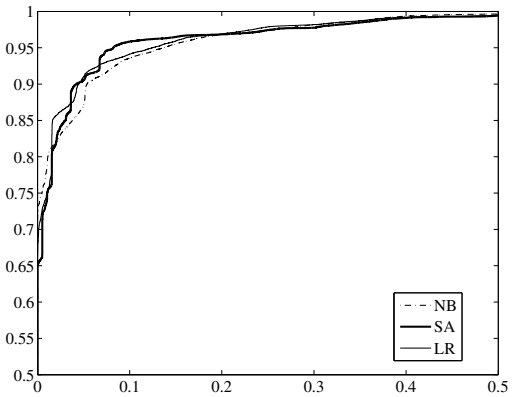
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



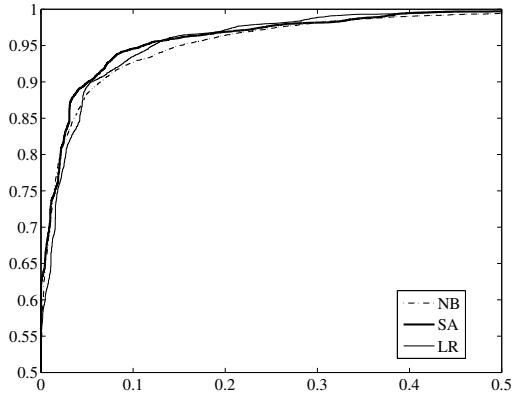
асептический некроз ГБК



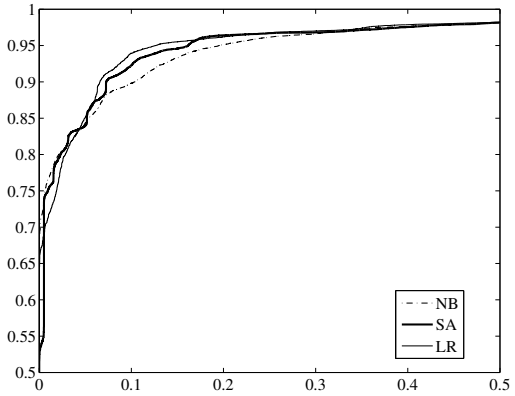
желчнокаменная болезнь



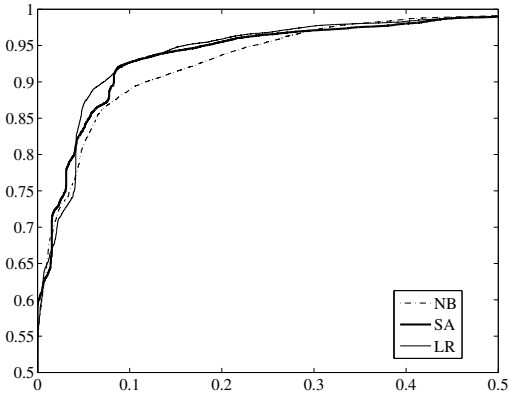
ишемическая болезнь



хронический гастрит 1



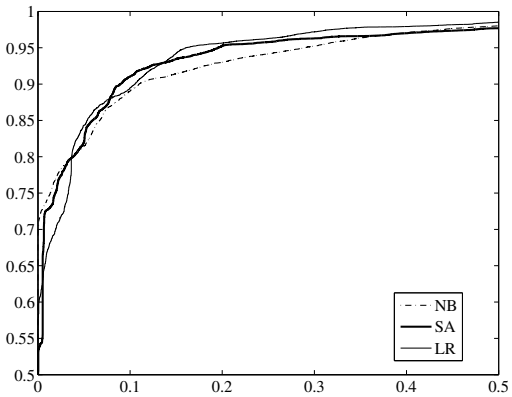
сахарный диабет



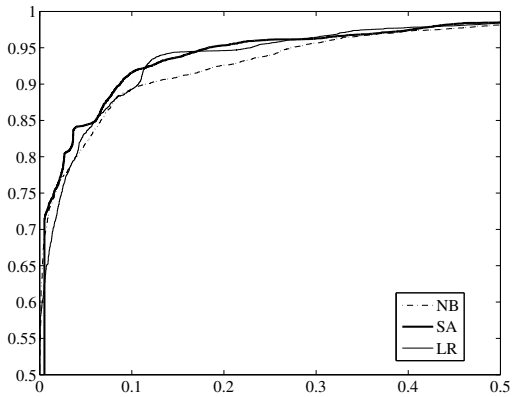
гипертония

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

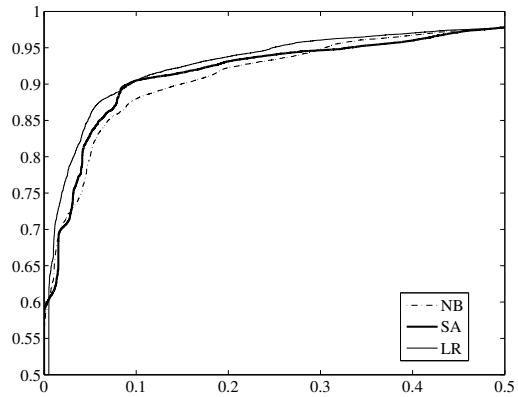
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



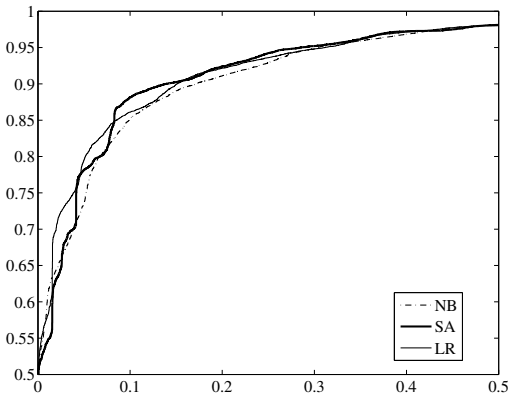
рак общий



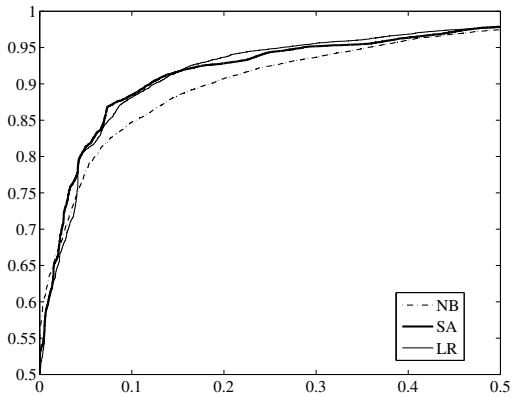
аденома простаты



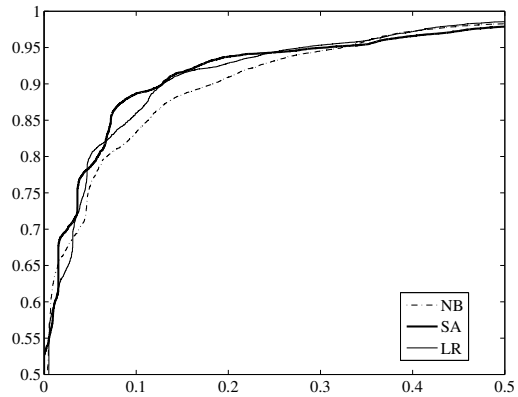
зоб щитовидной железы



хронический гастрит 2



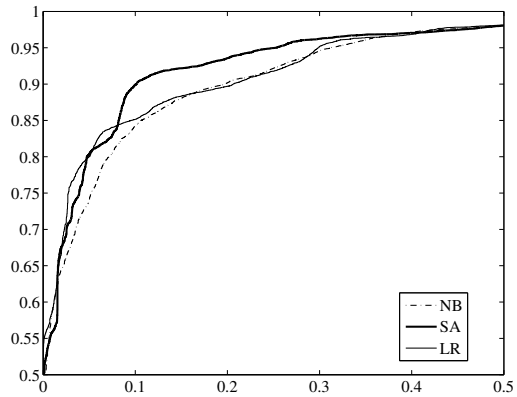
дискинезия ЖВП



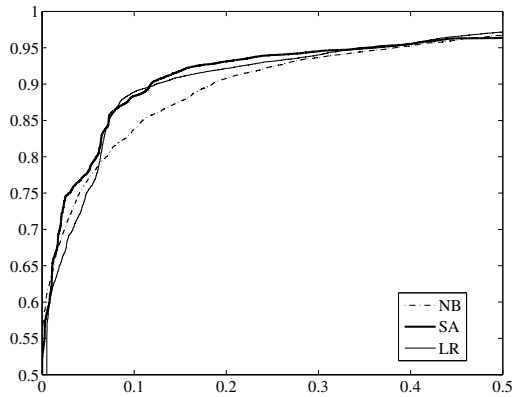
мочекаменная болезнь

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

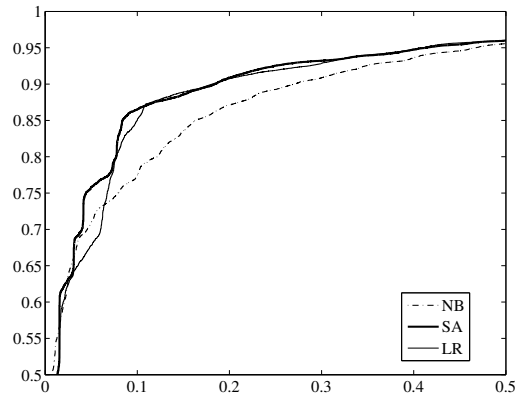
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



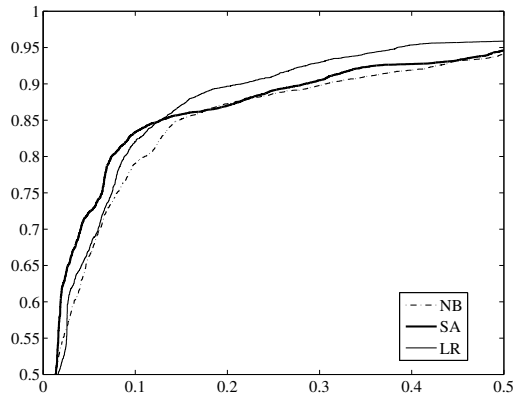
хронический холецистит



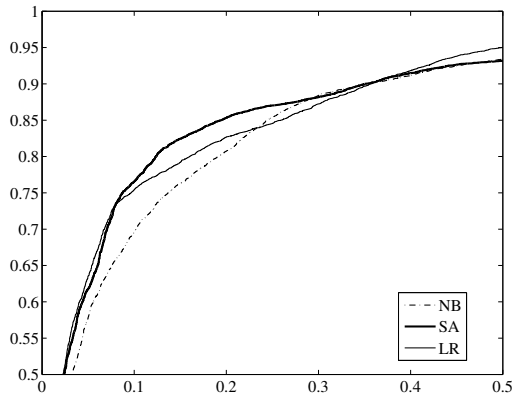
язвенная болезнь



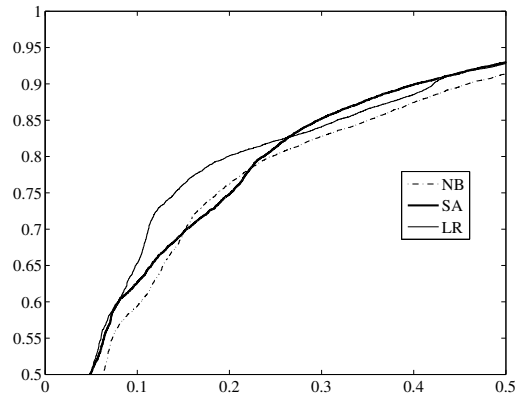
миома матки



хронический аднексит



анемия

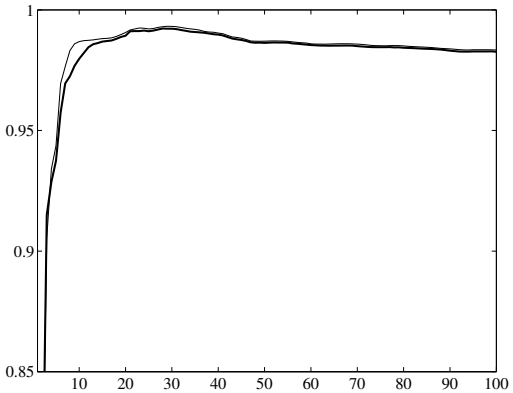


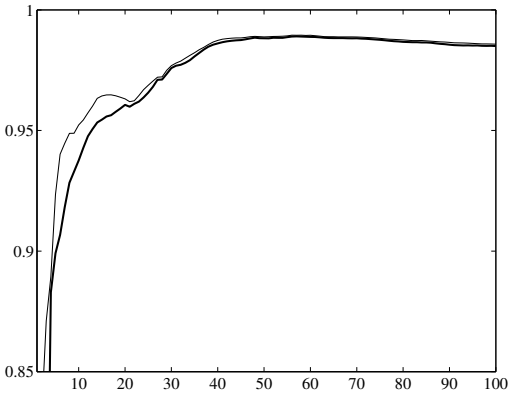
вегетососудистая дистония

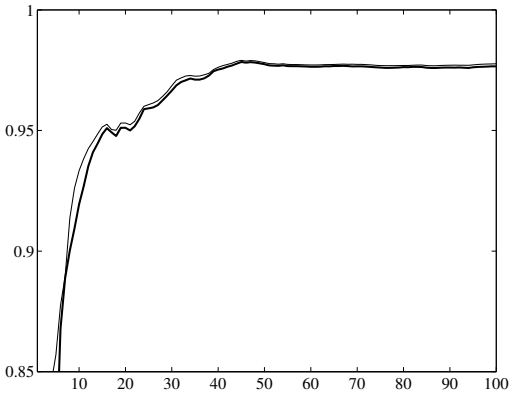
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

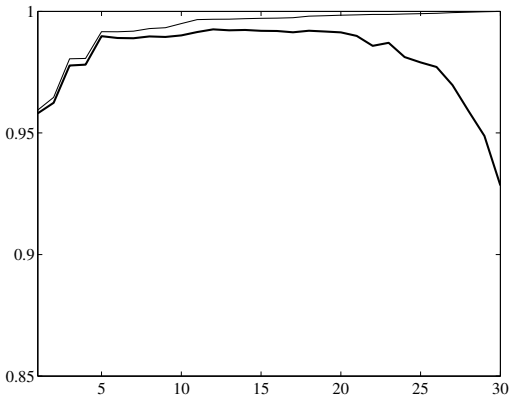
Синдромный алгоритм на K признаках:

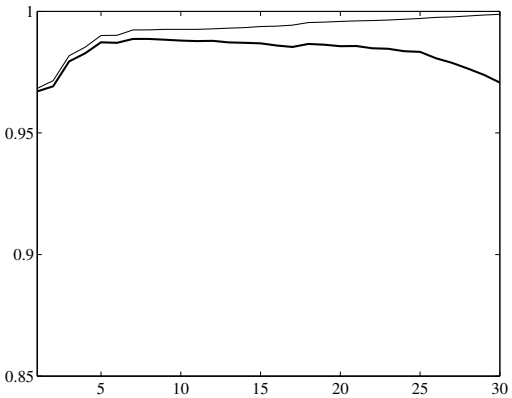


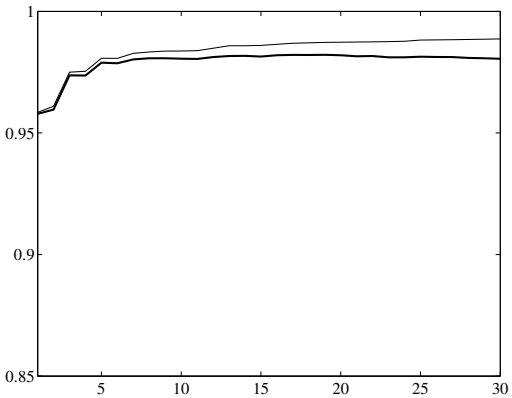




Логистическая регрессия на K главных компонентах:





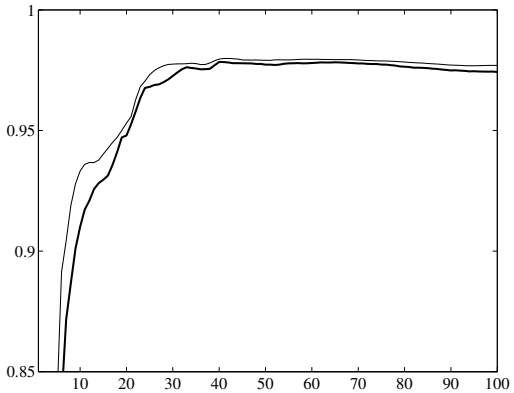


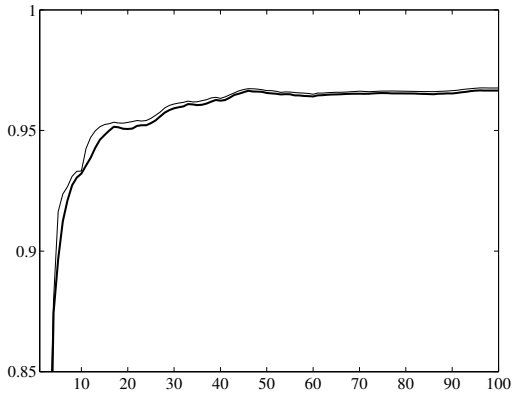
асептический некроз ГБК | желчнокаменная болезнь | ишемическая болезнь

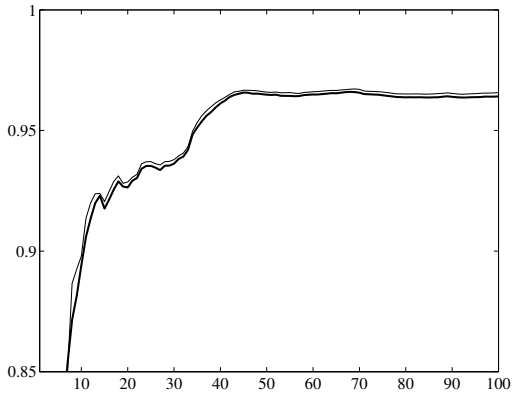
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

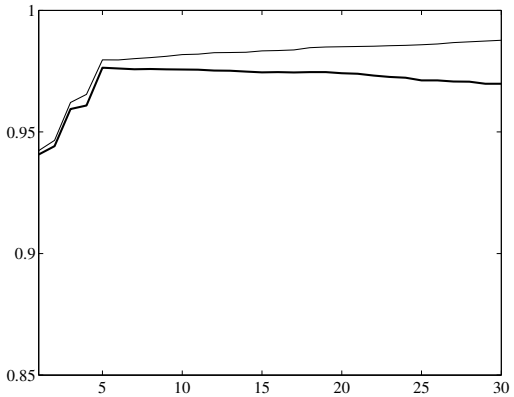
Синдромный алгоритм на K признаках:

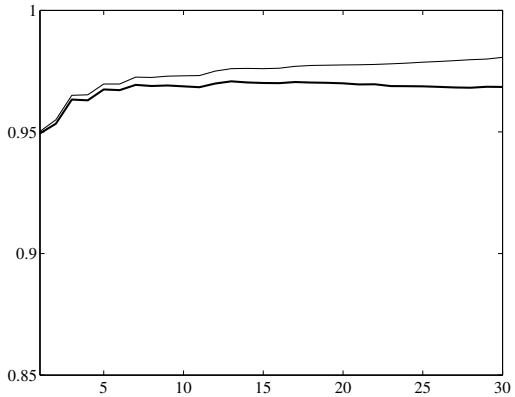


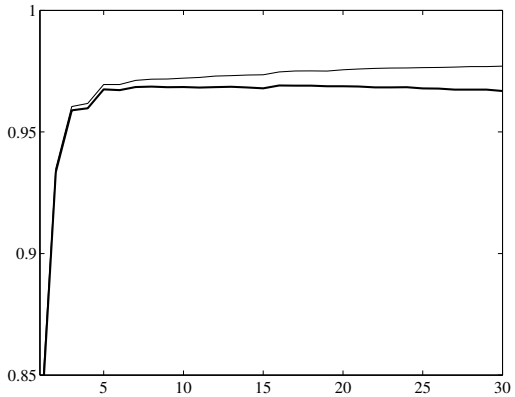




Логистическая регрессия на K главных компонентах:





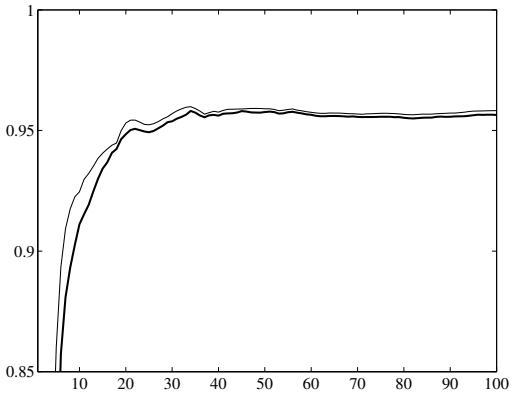


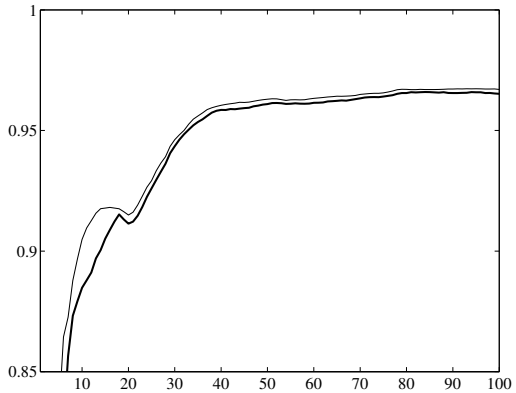
хронический гастрит 1 сахарный диабет гипертония

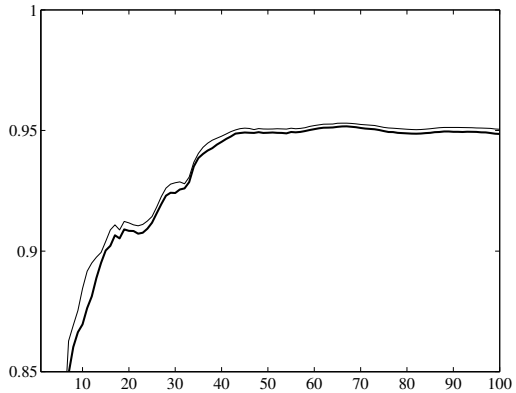
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

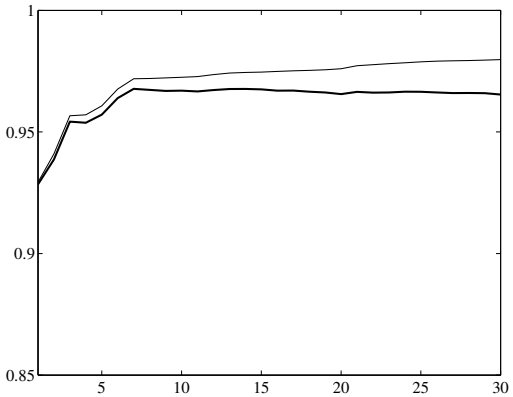
Синдромный алгоритм на K признаках:



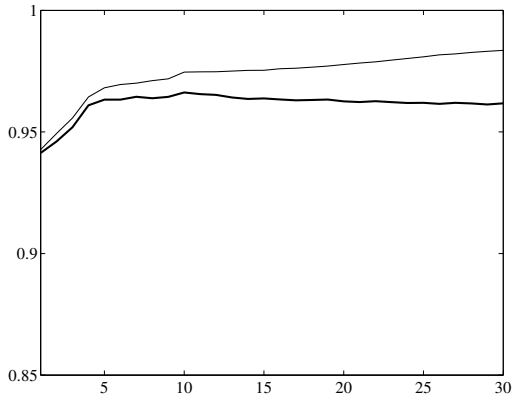




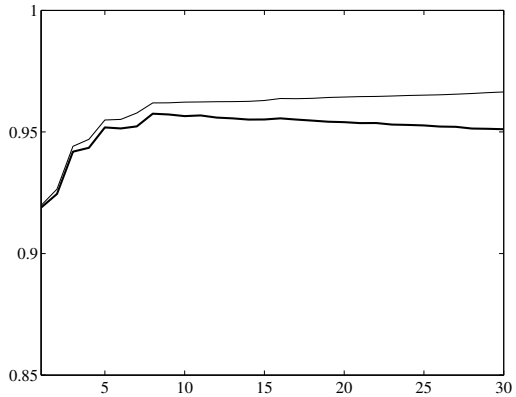
Логистическая регрессия на K главных компонентах:



рак общий



аденома простаты

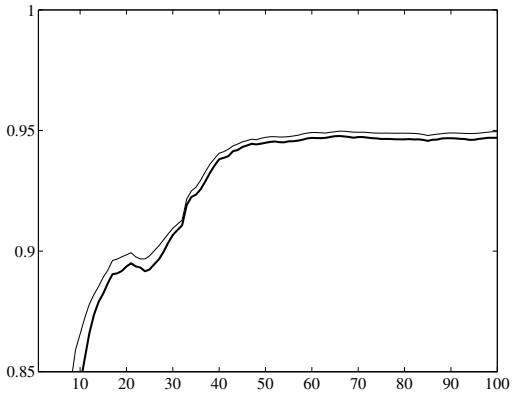
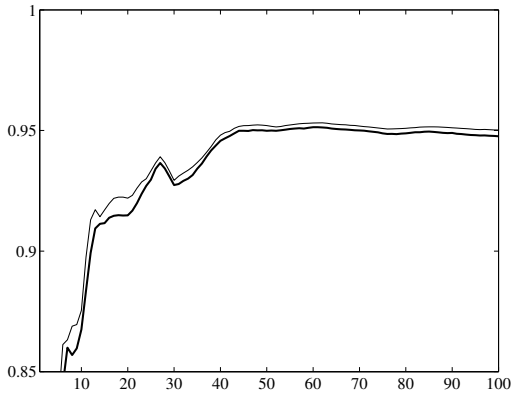
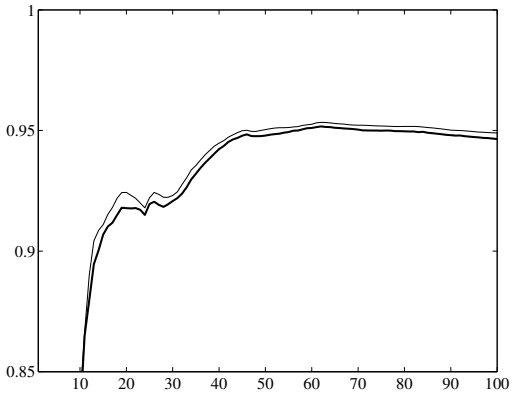


зоб щитовидной железы

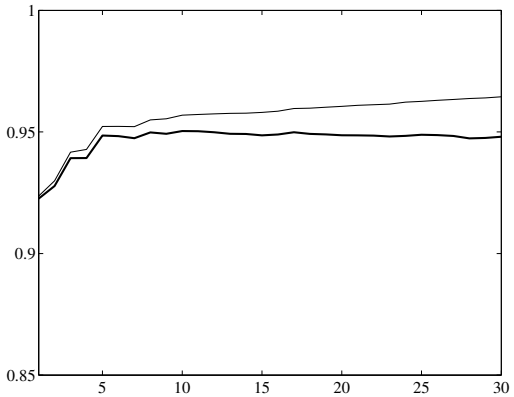
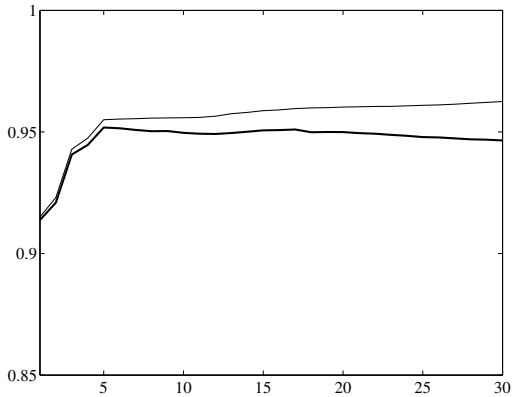
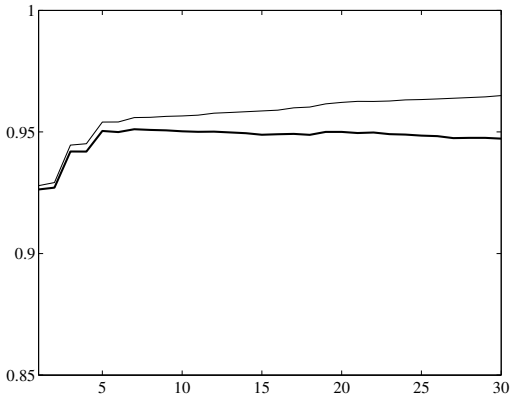
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм на K признаках:

Логистическая регрессия на K главных компонентах:

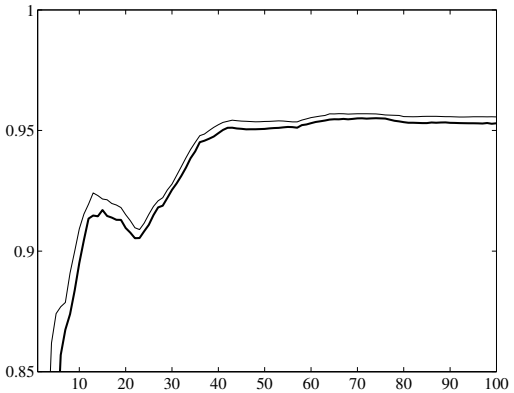
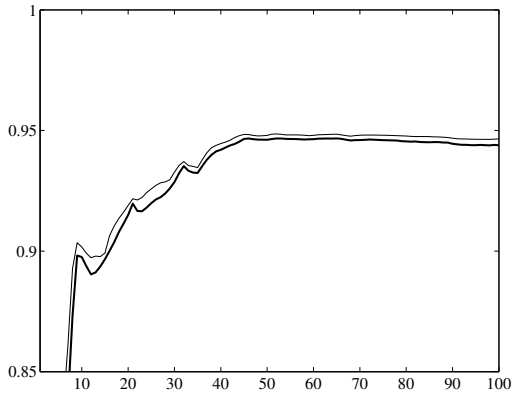
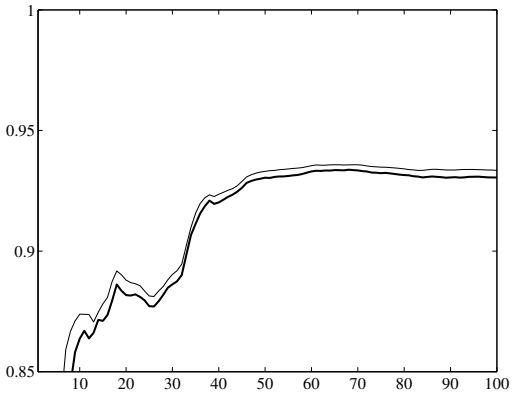




хронический гастрит 2 дискинезия ЖВП мочекаменная болезнь

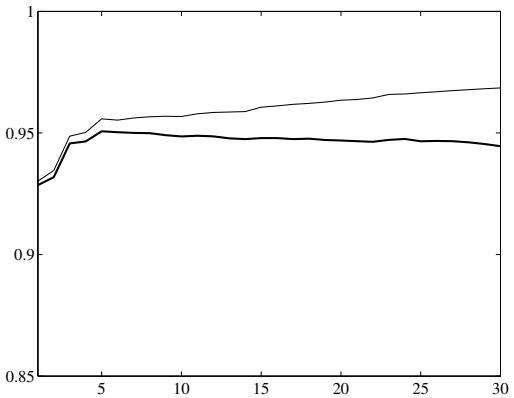
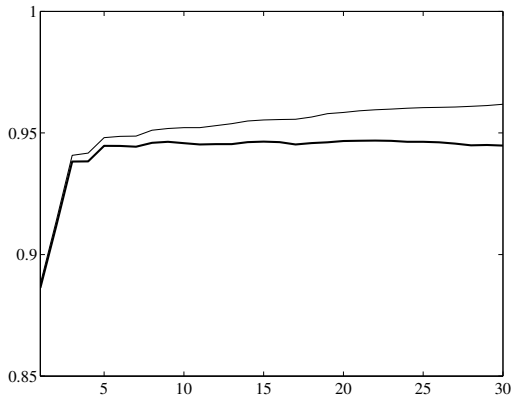
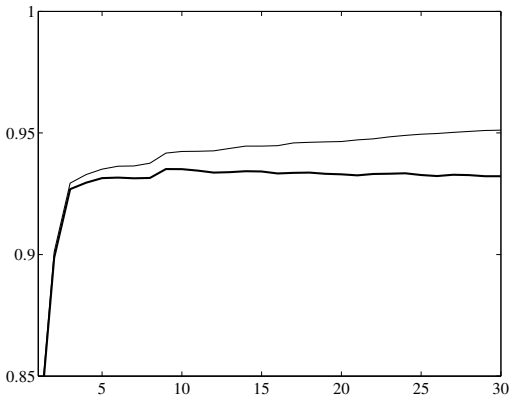
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм на K признаках:

Логистическая регрессия на K главных компонентах:

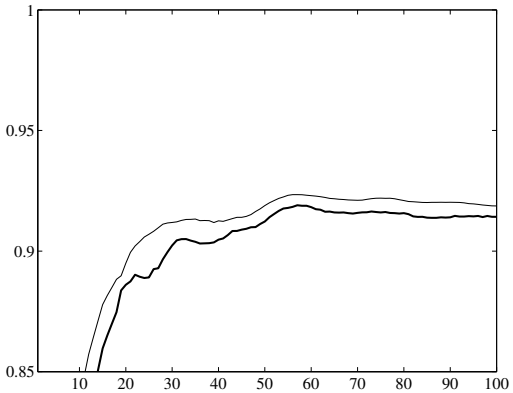
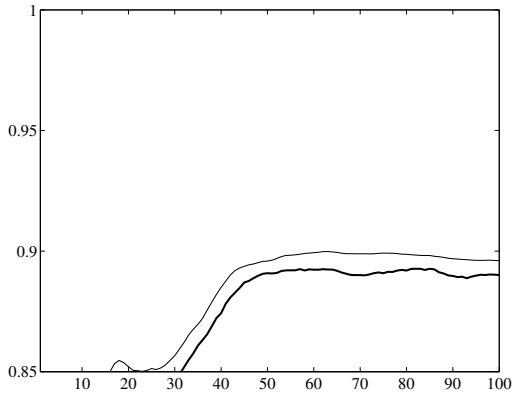
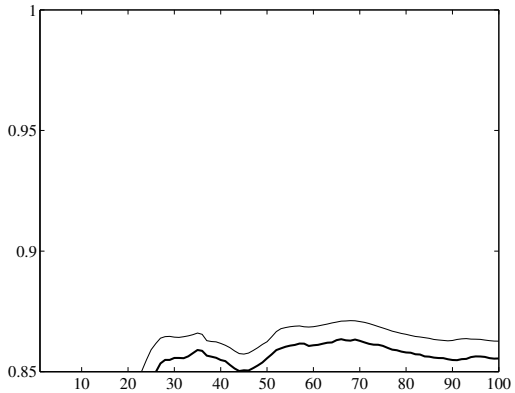




хронический холецистит | язвенная болезнь | миома матки

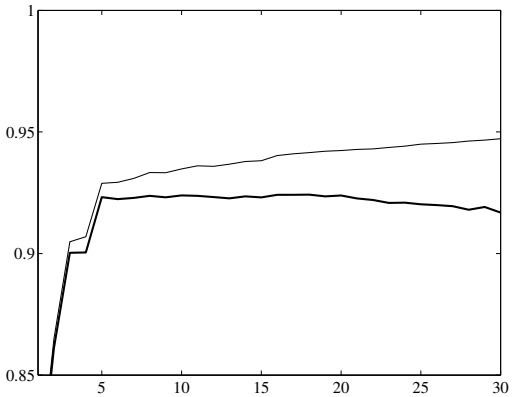
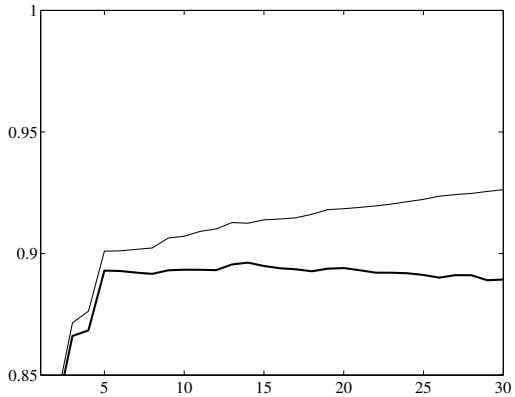
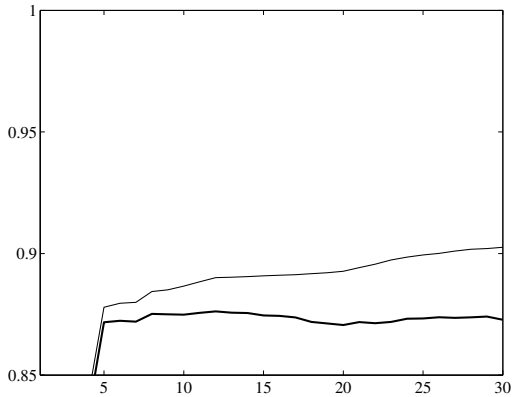
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм на K признаках:

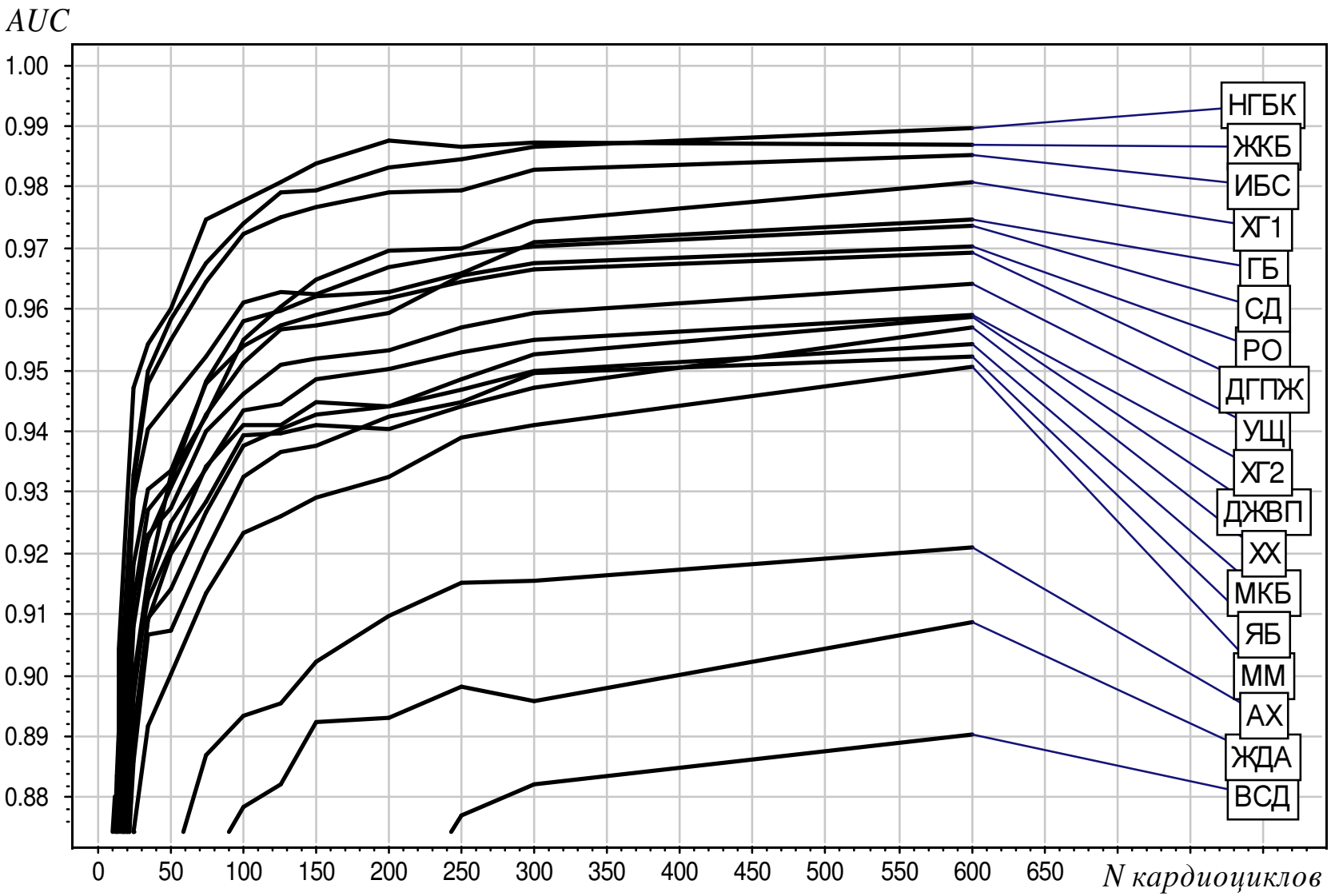
Логистическая регрессия на K главных компонентах:

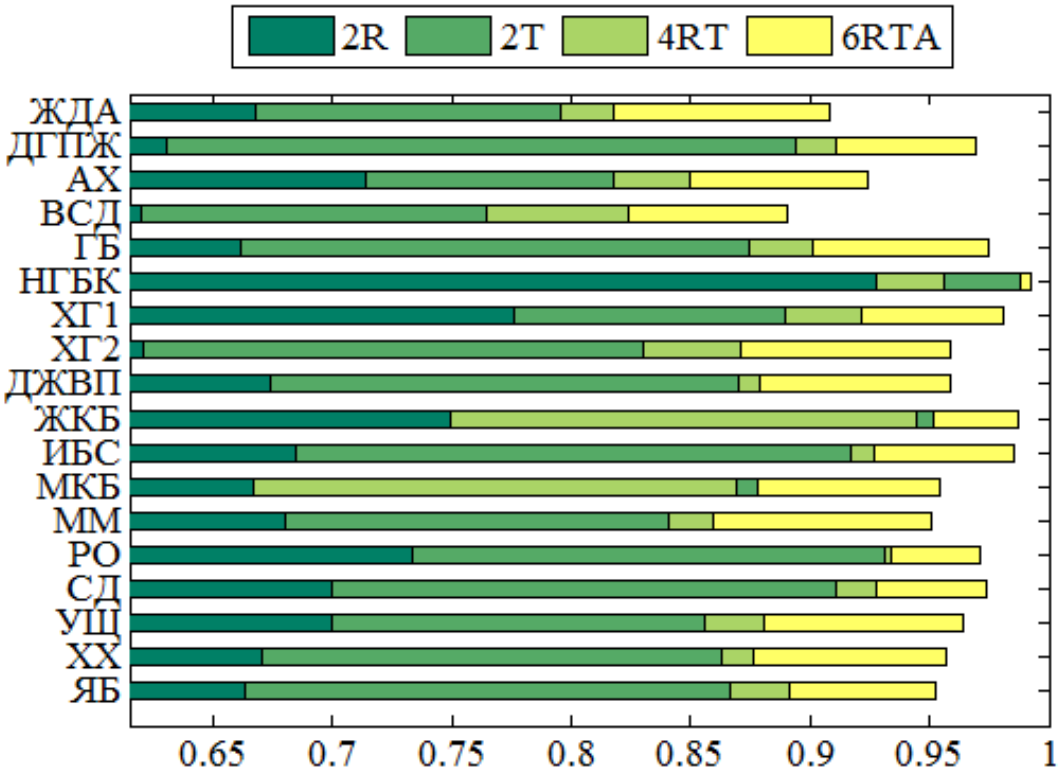
хронический аднексит | анемия | вегетососудистая дистония

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 — синдромный алгоритм

	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

- задача удивительно похожа на распознавание языка текста,
- многие болезни можно диагностировать по ЭКГ,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- причём достаточно нескольких минут записи ЭКГ.

Полезные страницы на www.MachineLearning.ru:

- Участник:Vokov
- Технология информационного анализа электрокардиосигналов
- Извлекаем пользу из Big Data (Проектная смена, СочиСириус, 2016)
- www.MachineLearning.ru/wiki/images/7/76/Voron16sirius-final.pdf