Применение машинного обучения и вычислительной лингвистики для диагностики заболеваний по электрокардиограмме

Воронцов Константин Вячеславович

Малый ШАД • 14 марта 2015

Содержание

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 «Language Identification»

Как обучить машину определять язык текста автоматически?

Зачем это нужно:

- Поисковые системы
- Системы агрегации контента
- Системы автоматического перевода

Определения и обозначения

x_i — обучающая выборка текстов, $i = 1, \ldots, \ell$
$y_i \in \{0, 1\}$ — два класса: «язык 0» и «язык 1»
(будем сравнивать языки попарно)

Векторизация текста — это преобразование текста (последовательности букв) в числовой вектор признаков

Основной принцип векторизации:
признаки должны содержать важную информацию о классах

N-грамма — подстрока текста из N последовательных букв
K — число букв в алфавите
$n = K^N$ — число N-грамм

$f_j(x_i)$ — частота N-граммы j в тексте x_i $(j = 1 \ldots n,\ i = 1 \ldots \ell)$
$\big(f_j(x_i)\big)_{j=1}^{n}$ — n-мерный вектор признаков текста x_i

Модель классификации текстов

Основное предположение:
каждый язык имеет уникальное распределение частот N-грамм

Линейная модель классификации:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j f_j(x),$$

где w_j — вес N-граммы j:

- $w_j > 0$, N-грамма более специфична для языка 1
- $w_j < 0$, N-грамма более специфична для языка 0
- $w_j = 0$, N-грамма не различает эти язвки

Методы *машинного обучения* позволяют настраивать веса w_j по обучающей выборке автоматически

Методы обучения линейных классификаторов

Линейная модель классификации:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j f_j(x),$$

Методы обучения весов w_j в линейных классификаторах

- SVM — Support Vector Machine
- LR — Logistic Regression
- RLR — Regularized Logistic Regression
- LASSO — Least Absolute Shrinkage and Selection Operator
- NB — Naïve Bayes
- и др.

Но в данной задаче простая эвристика уже работает хорошо.

Простые эвристики для выбора весов

Средняя частота N-граммы j в текстах класса y:

$$S_y^j = \frac{1}{\ell_y} \sum_{i=1}^{\ell} [y_i = y] f_j(x_i), \quad \ell_y = \sum_{i=1}^{\ell} [y_i = y]$$

Эвристика: вес N-граммы j должен быть тем больше, чем больше S_1^j и чем меньше S_0^j

Можно пробовать разные формулы для весов:

$$w_j = \frac{S_1^j + \gamma}{S_0^j + \gamma} \qquad w_j = \log \frac{S_1^j + \gamma}{S_0^j + \gamma}$$

$$w_j = \sqrt{S_1^j} - \sqrt{S_0^j} \qquad w_j = \sqrt{S_1^j/\ell_1} - \sqrt{S_0^j/\ell_0}$$

... и разрешается фантазировать!

Тексты Декларации прав человека

Цели эксперимента — проверить:

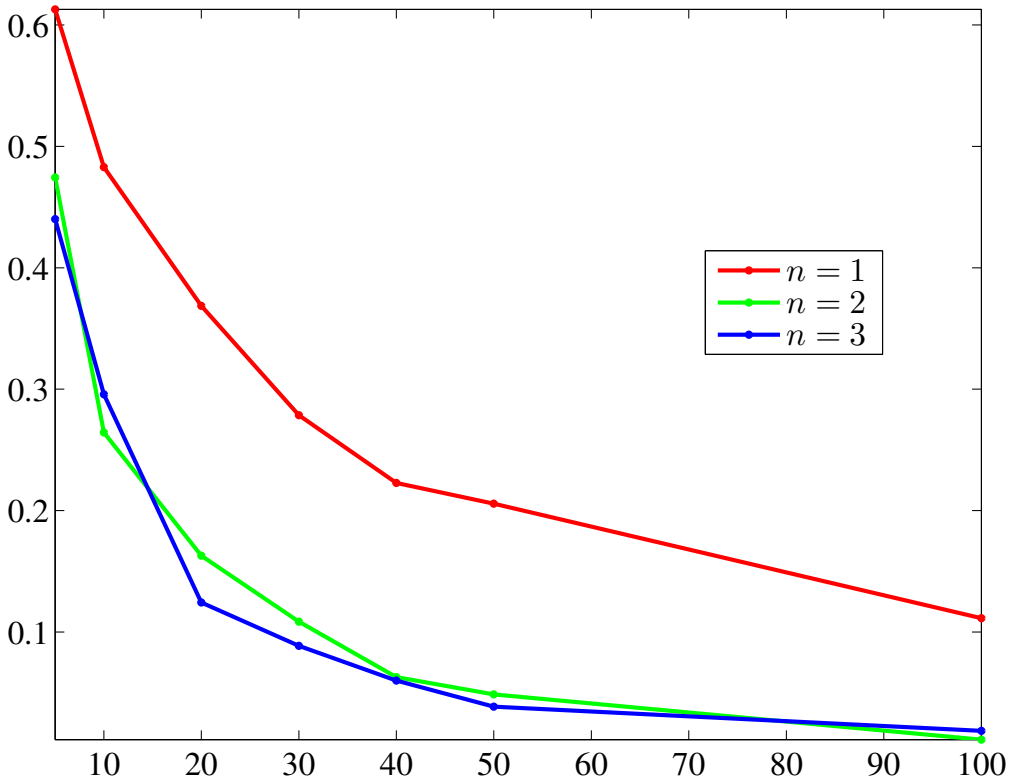
- действительно ли частоты триграмм распознают язык?
- как точность распознавания зависит от длины текста?

Методика эксперимента:

- используем тексты Декларации на 7 языках
- коэффициенты w_j определяем по обучающему фрагменту текста длины ℓ для каждого языка
- точность измеряем как долю ошибочных распознаваний языка по контрольным фрагментам текстов длины k

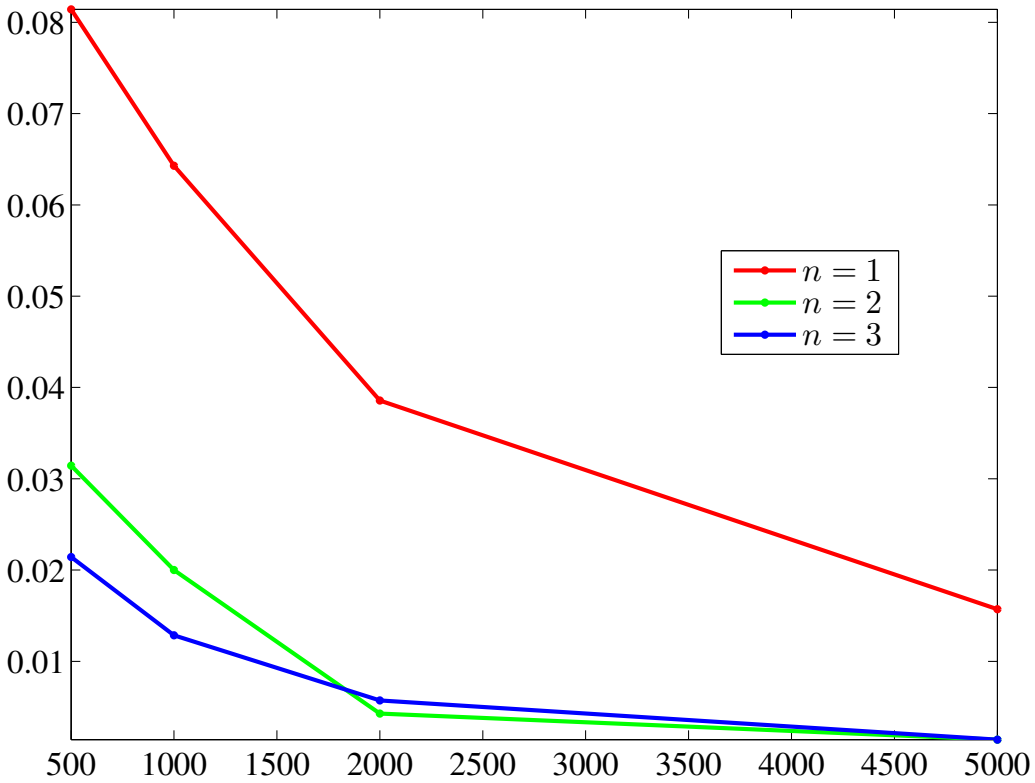
Результаты эксперимента

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



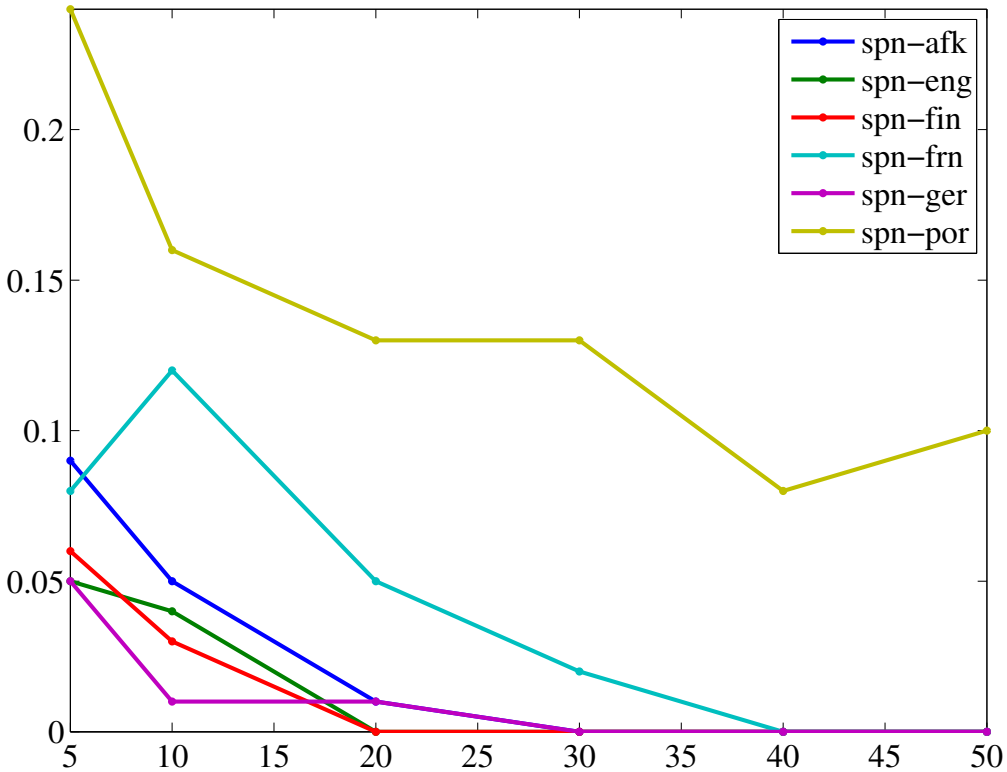
Результаты эксперимента

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:

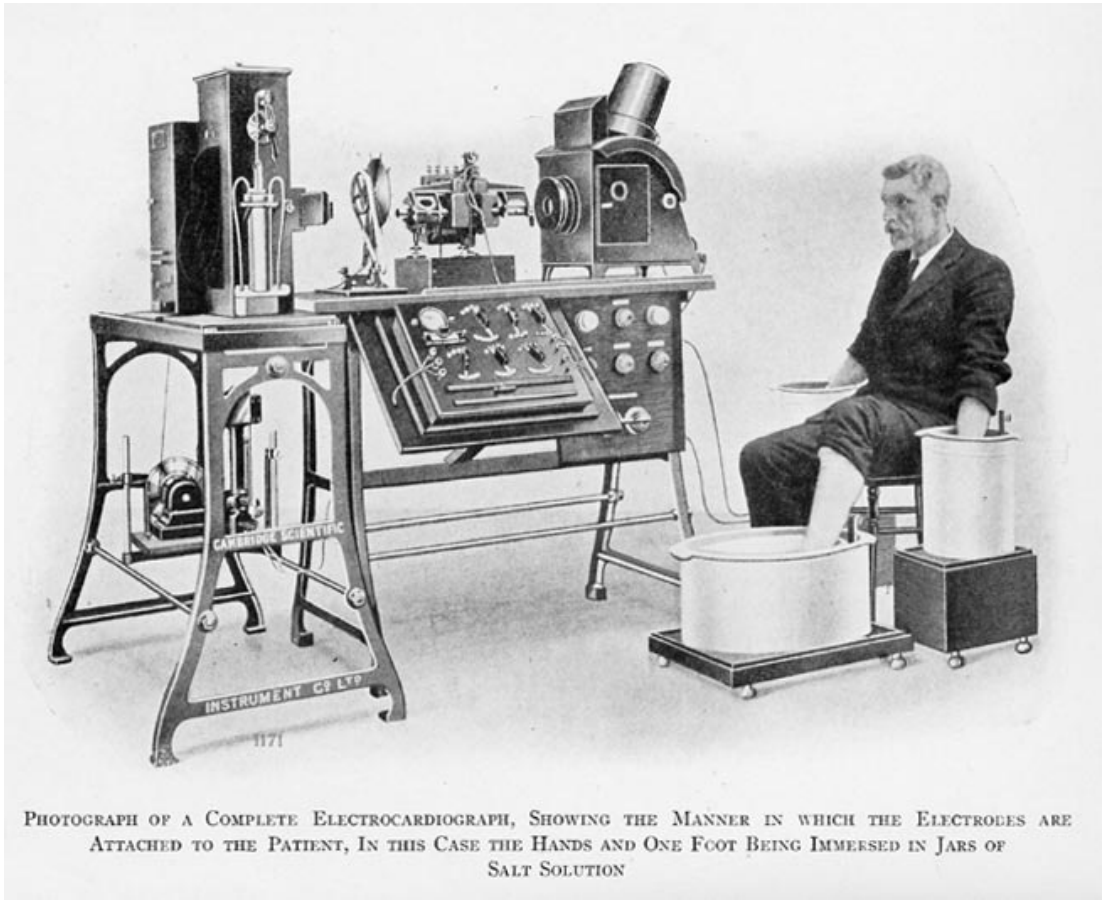
- использовались только языки на основе латиницы,
- все диакритические знаки и пробел были заменены на «-»

	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

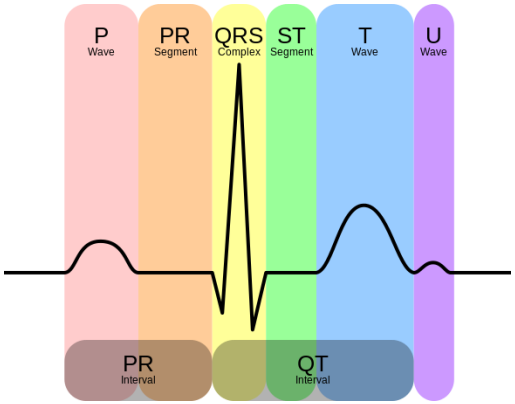
Выводы

- языки можно распознавать автоматически,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- причём точность распознавания быстро увеличивается с ростом длины контрольного текста, и сотни символов уже хватает для распознавания даже близких языков

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:

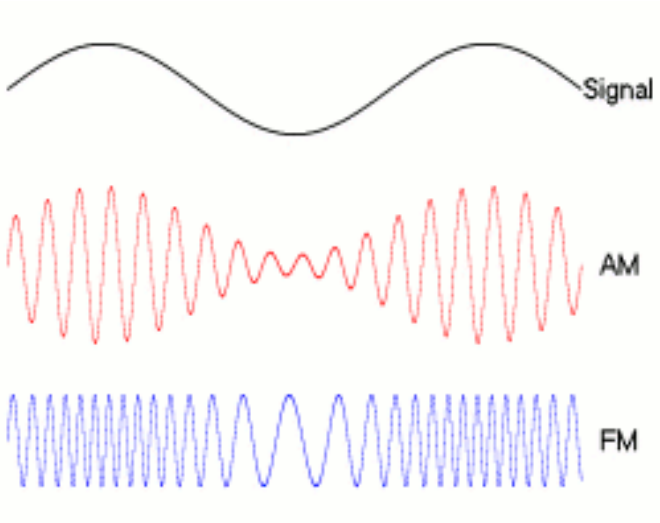
- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Электрокардиография высокого разрешения ($> 500\,\text{Гц}$)

Предположения:

- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Каждое заболевание по-своему «модулирует» ЭКГ-сигнал
- Для диагностики важны *знаки приращений интервалов и амплитуд последовательных кардиоциклов*
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика

Понятия модуляции 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.

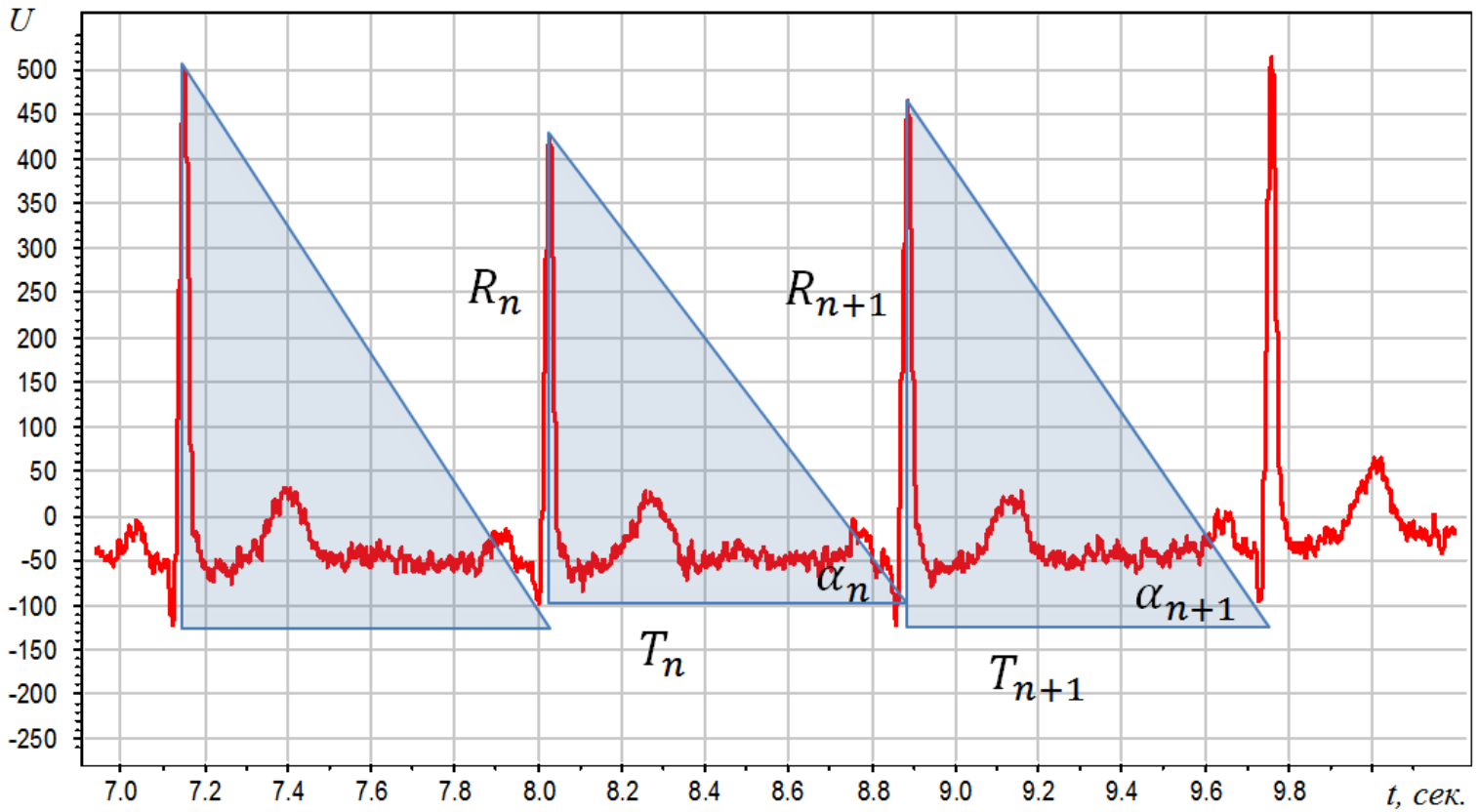
Что будет аналогом демодуляции в случае ЭКГ?

Приращения интервалов и амплитуд кардиоциклов

приращение амплитуд: $dR_n = R_{n+1} - R_n$

приращение интервалов: $dT_n = T_{n+1} - T_n$

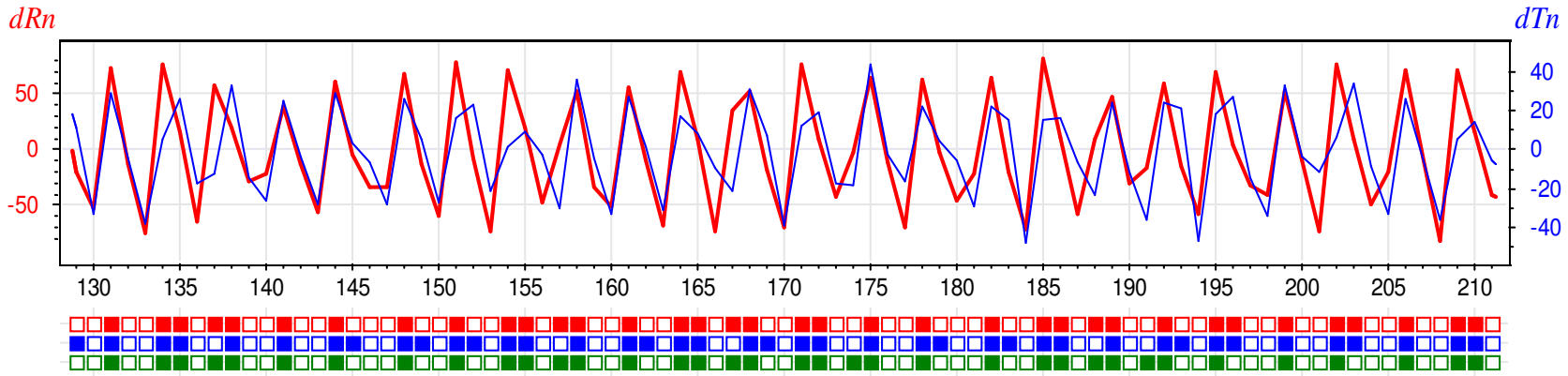
приращение углов: $d\alpha_n = \alpha_{n+1} - \alpha_n, \qquad \alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$



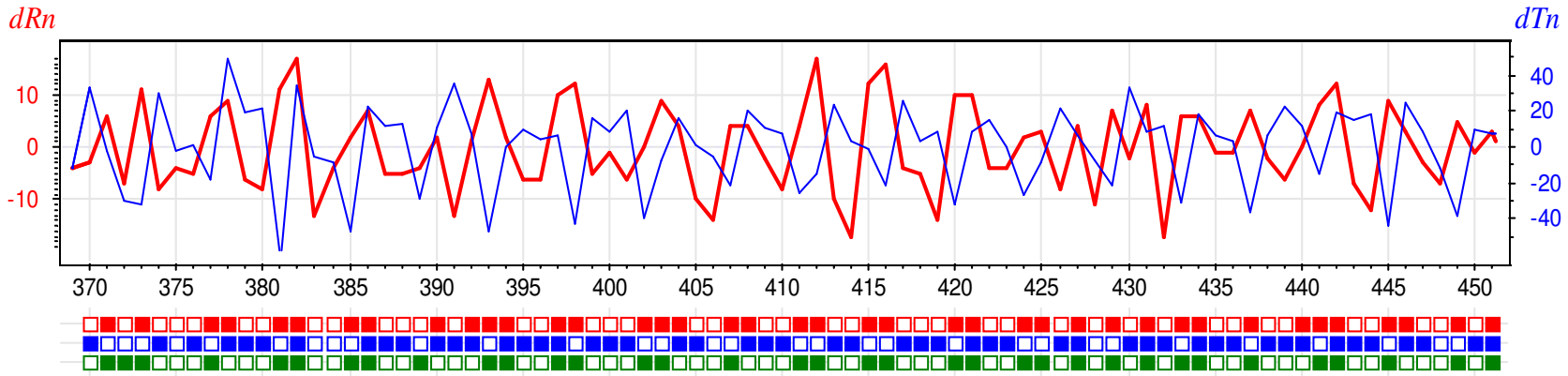
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_n, dT_n, $d\alpha_n$ в последовательных кардиоциклах n

Здоровый:



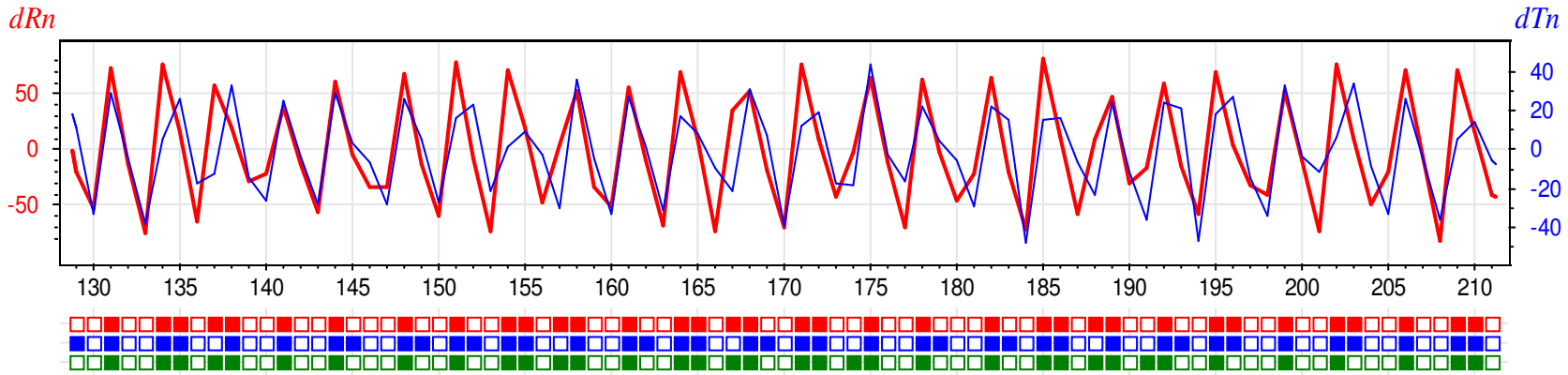
Больной (язвенная болезнь):



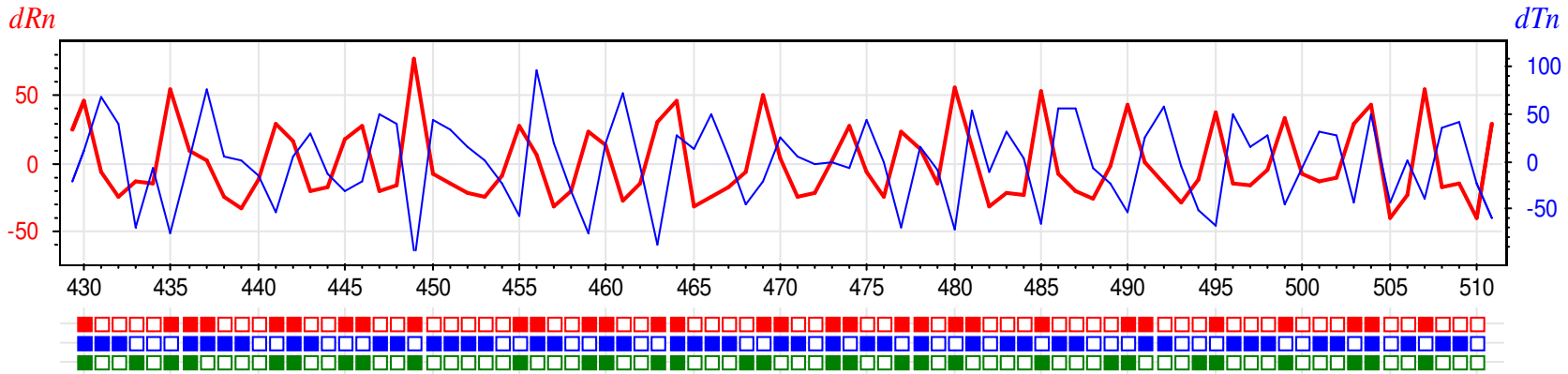
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_n, dT_n, $d\alpha_n$ в последовательных кардиоциклах n

Здоровый:



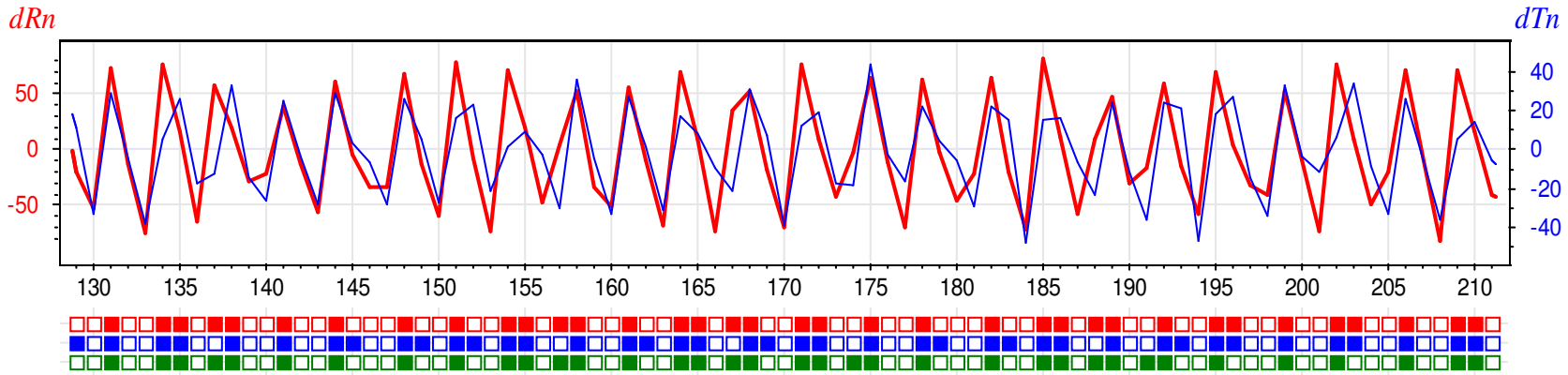
Больной (гипертония):



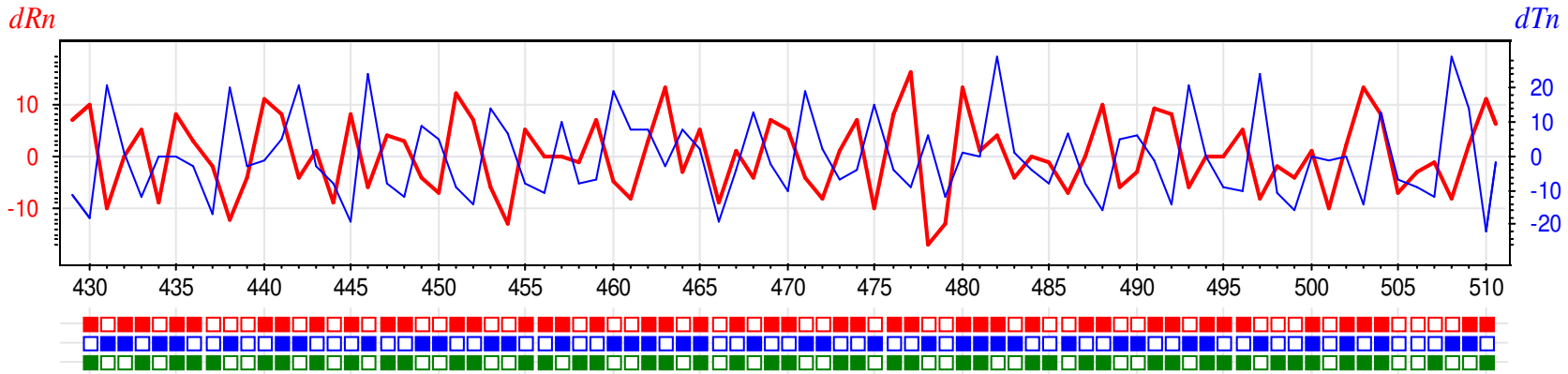
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_n, dT_n, $d\alpha_n$ в последовательных кардиоциклах n

Здоровый:



Больной (рак):



Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^{N}$

Правила кодирования:

$dR_n = R_{n+1} - R_n$	$+$	$-$	$+$	$-$	$+$	$-$
$dT_n = T_{n+1} - T_n$	$+$	$-$	$-$	$+$	$+$	$-$
$d\alpha_n = \alpha_{n+1} - \alpha_n$	$+$	$+$	$+$	$-$	$-$	$-$
s_n	A	B	C	D	E	F

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

```
DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE
```

Векторизация кодограммы ЭКГ-сигнала

Вход: *кодограмма* $x = (s_1, \ldots, s_{N-1})$ как текстовая строка

Выход: частоты триграмм $f_j(x)$ — сколько раз триграмма j появилась в кодограмме x, $j = 1, \ldots, n$, $n = 6^3 = 216$

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

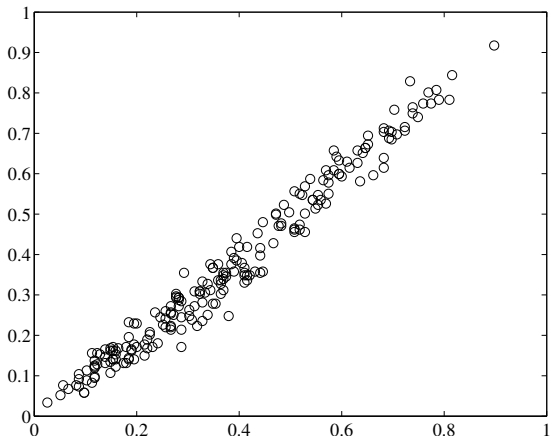
Объём исходных данных (по заболеваниям)

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

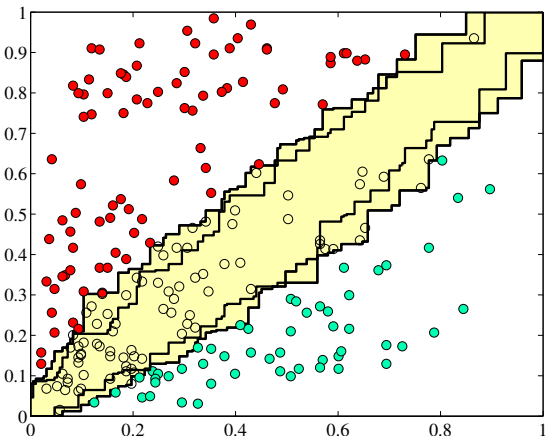
Нулевая гипотеза: частота триграммы не зависит от класса

Точки на графиках — это триграммы, $j = 1, \ldots, 216$
— ось X: доля здоровых с частотой триграммы $f_j(x) \geqslant \theta$
— ось Y: доля больных с частотой триграммы $f_j(x) \geqslant \theta$

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i

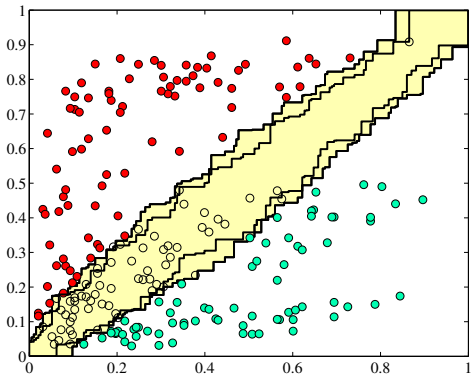


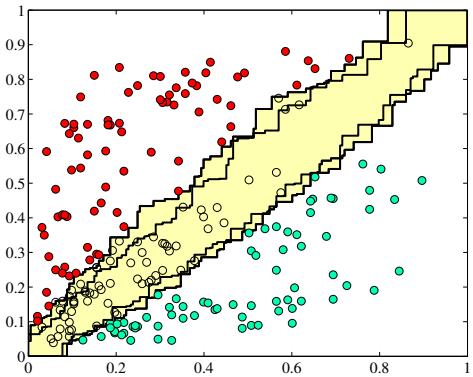
наблюдаемые y_i

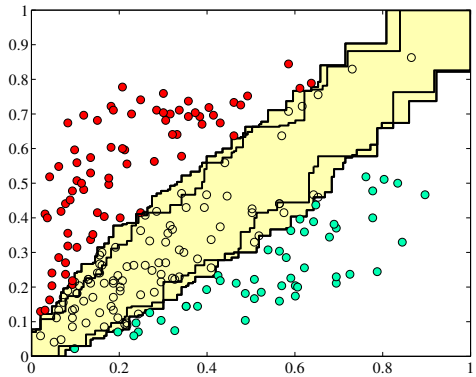
Нулевая гипотеза отвергается для большинства триграмм (красные и зелёные точки вне жёлтой области), при уровнях значимости 10% и 0.2% (20 и 1000 перемешиваний)

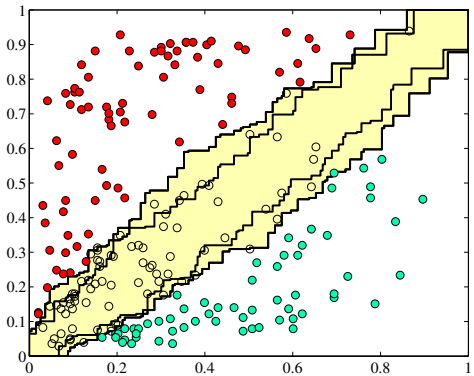
Результаты перестановочного теста для различных болезней

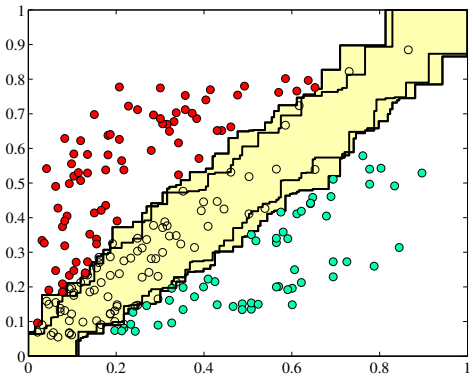
Для каждой болезни есть свои неслучайно частые триграммы

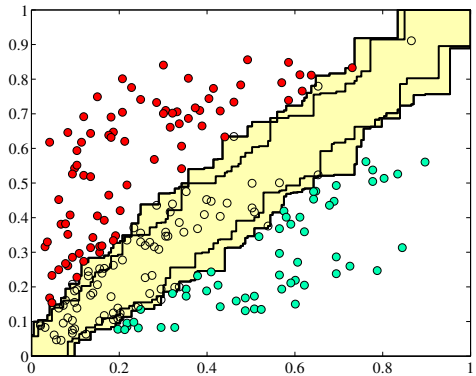

ишемия сердца


гипертония


рак


желчнокаменная болезнь

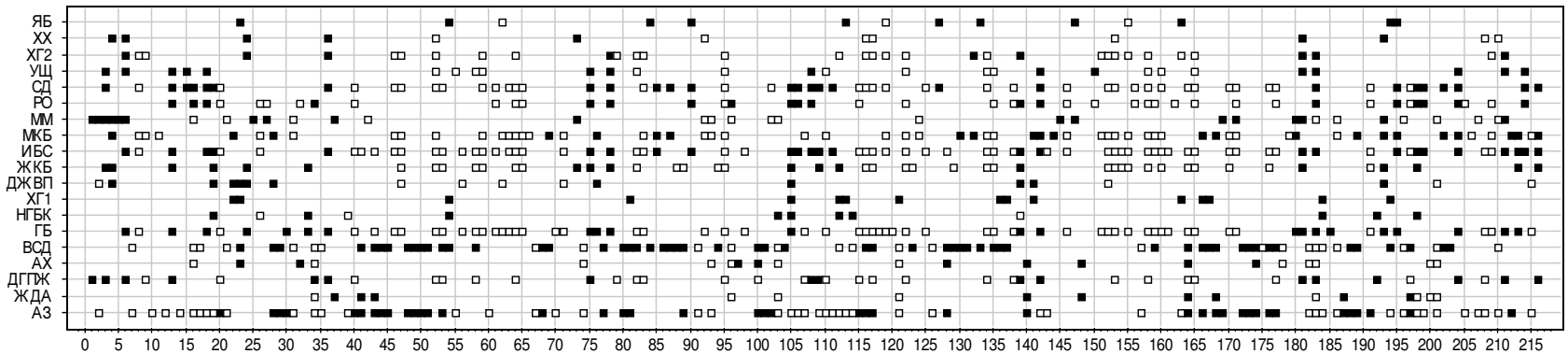

миома матки


язвенная болезнь

Болезни отличаются наборами информативных триграмм

ось X: — номера триграмм 1..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 встречаемости

Вывод 2. Болезни хорошо отличаются по наборам триграмм!

Модель классификации

x_i — обучающая выборка кодограмм, $i = 1, \ldots, \ell$
y_i — диагноз: 0 = здоровый, 1 = больной
$f_j(x_i)$ — частота триграммы j в кодограмме

Предположения:
1) для каждой болезни есть свой набор частых триграмм
2) если триграмма часто встречается, то не важно, сколько раз

Линейная модель классификации:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j \big[f_j(x) \geqslant \theta\big],$$

где w_j — вес триграммы j:

- $w_j > 0$, триграмма специфична для больных
- $w_j < 0$, триграмма специфична для здоровых
- $w_j = 0$, триграмма не релевантна для этой болезни

Простые эвристики для выбора весов

Число объектов класса y, для которых триграмма j частая

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

Число объектов класса y, для которых триграмма j редкая

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

Эвристика: вес триграммы j должен быть тем больше,
чем больше S_1^j и s_0^j,
чем меньше S_0^j и s_1^j.

Простые эвристики для выбора весов

Эвристика: вес триграммы j должен быть тем больше,
чем больше S_1^j и s_0^j,
чем меньше S_0^j и s_1^j.

Поэтому можно пробовать разные формулы для весов:

$$w_j = \frac{S_1^j}{S_0^j} \qquad w_j = \frac{S_1^j s_0^j}{S_0^j s_1^j}$$

$$w_j = \log \frac{S_1^j}{S_0^j} \qquad w_j = \log \frac{S_1^j s_0^j}{S_0^j s_1^j}$$

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

... и разрешается фантазировать!

Стандартные критерии качества диагностики

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{1}{\ell_1} \sum_{i:\, y_i=1} \big[\langle x_i, w\rangle \geqslant w_0\big]$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{1}{\ell_0} \sum_{i:\, y_i=0} \big[\langle x_i, w\rangle < w_0\big]$$

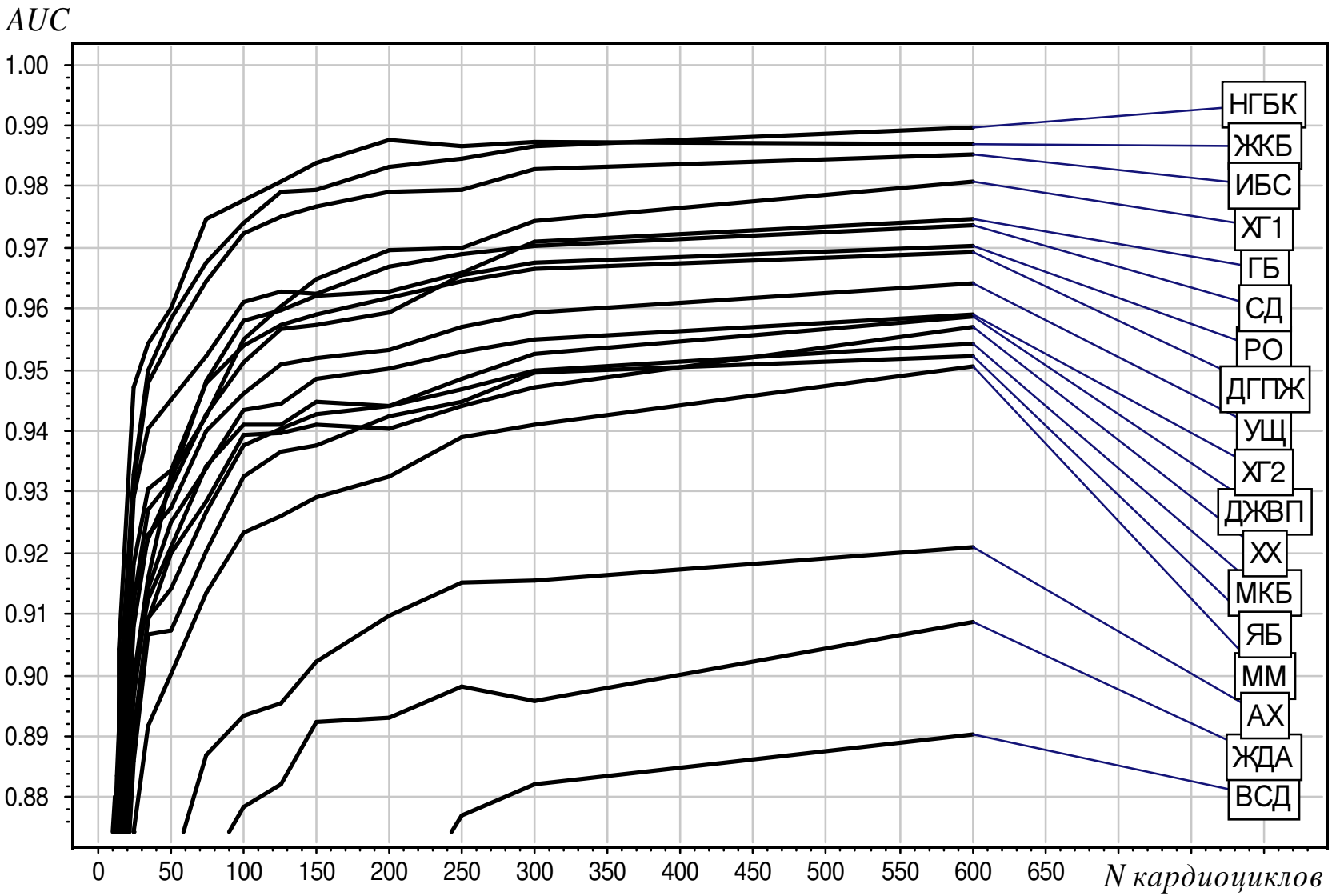
Area Under Curve — доля правильно упорядоченных пар:

$$\text{AUC} = \frac{1}{\ell_0 \ell_1} \sum_{i:\, y_i=0} \sum_{k:\, y_k=1} \big[\langle x_i, w\rangle < \langle x_k, w\rangle\big]$$

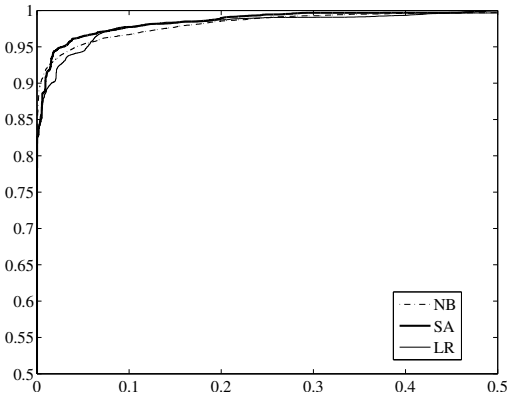
Результаты кросс-валидации

	Логистическая регрессия			Синдромный алгоритм		
	AUC, %	C (Ч=95%)	C=Ч, %	AUC, %	C (Ч=95%)	C=Ч, %
НГБК	99.26 ± 0.36	95.8 ± 1.6	95.2 ± 0.8	99.23 ± 0.05	97.4 ± 1.0	95.8 ± 0.9
ЖКБ	99.00 ± 0.25	94.9 ± 2.2	95.1 ± 1.1	98.90 ± 0.02	95.3 ± 0.5	95.5 ± 0.5
ИБС	98.21 ± 0.16	90.4 ± 1.3	93.1 ± 0.8	97.84 ± 0.03	91.8 ± 0.4	93.3 ± 0.0
ХГ1	97.64 ± 0.13	87.9 ± 1.9	91.9 ± 1.1	97.84 ± 0.09	89.4 ± 1.3	93.0 ± 0.8
СД	97.08 ± 0.14	84.1 ± 1.8	91.9 ± 0.9	96.66 ± 0.05	84.0 ± 0.9	91.2 ± 0.6
ГБ	96.91 ± 0.18	84.5 ± 2.7	91.8 ± 0.7	96.60 ± 0.05	81.6 ± 1.8	91.5 ± 0.4
РО	96.77 ± 0.17	82.7 ± 2.9	90.6 ± 0.9	95.81 ± 0.14	80.2 ± 3.0	90.5 ± 0.8
ДГПЖ	96.62 ± 0.40	77.2 ± 4.7	91.0 ± 1.2	96.59 ± 0.10	79.8 ± 3.7	91.2 ± 0.7
УЩ	95.75 ± 0.14	75.0 ± 2.7	90.2 ± 0.6	95.17 ± 0.10	66.7 ± 2.2	90.4 ± 0.6
ХГ2	95.22 ± 0.18	72.0 ± 2.0	88.4 ± 0.8	94.77 ± 0.11	71.7 ± 2.8	88.8 ± 1.0
ДЖВП	95.18 ± 0.15	73.4 ± 1.9	88.9 ± 0.9	95.14 ± 0.08	70.9 ± 2.2	89.1 ± 1.0
МКБ	95.11 ± 0.28	71.9 ± 3.5	88.6 ± 1.0	95.17 ± 0.07	69.0 ± 4.2	89.0 ± 0.3
ХХ	95.07 ± 0.21	73.4 ± 2.8	88.8 ± 1.3	95.51 ± 0.10	76.3 ± 1.9	90.1 ± 0.5
ЯБ	94.69 ± 0.40	66.2 ± 3.2	88.6 ± 1.4	94.67 ± 0.05	64.3 ± 2.5	89.6 ± 0.5
ММ	93.52 ± 0.30	60.5 ± 2.8	87.1 ± 1.1	93.37 ± 0.10	59.0 ± 2.1	87.6 ± 1.0
АХ	92.42 ± 0.48	62.7 ± 6.3	85.5 ± 2.0	91.90 ± 0.29	49.0 ± 3.4	85.6 ± 1.0
ЖДА	90.04 ± 0.60	54.4 ± 7.2	81.2 ± 1.8	89.27 ± 0.28	35.9 ± 6.1	83.0 ± 1.2
ВСД	87.62 ± 0.67	42.2 ± 5.0	79.9 ± 1.1	86.35 ± 0.24	39.5 ± 4.5	77.9 ± 1.0

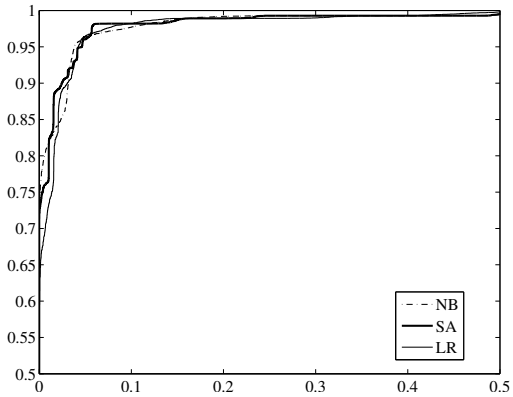
Зависимость AUC от длительности регистрации ЭКГ



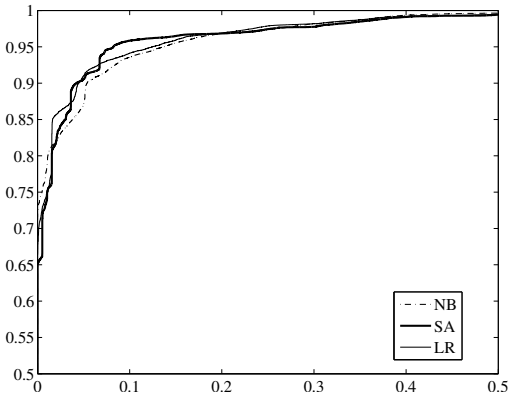
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



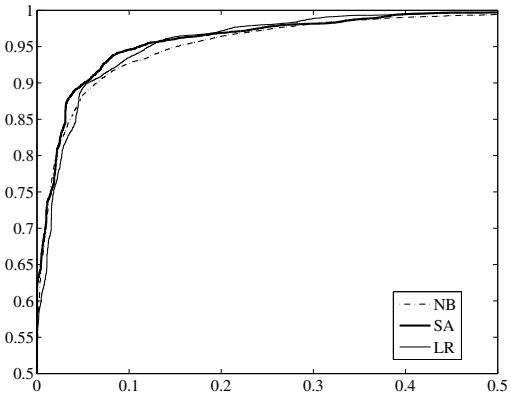
асептический некроз ГБК



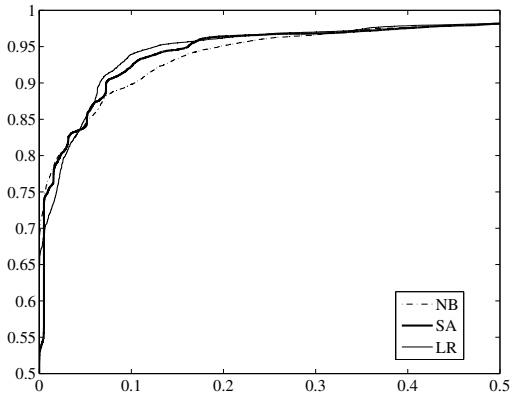
желчнокаменная болезнь



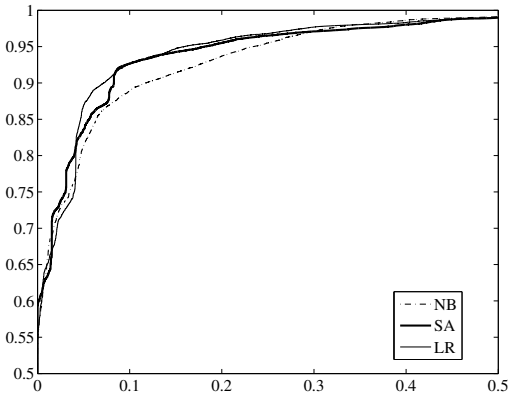
ишемическая болезнь



хронический гастрит 1



сахарный диабет

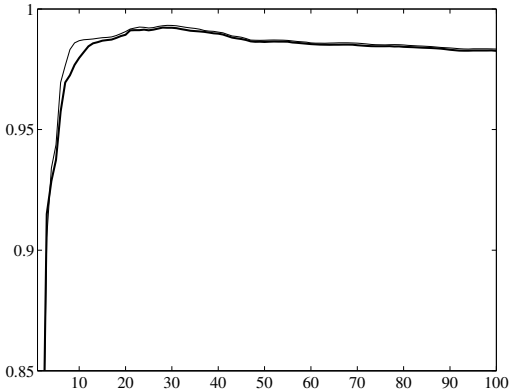


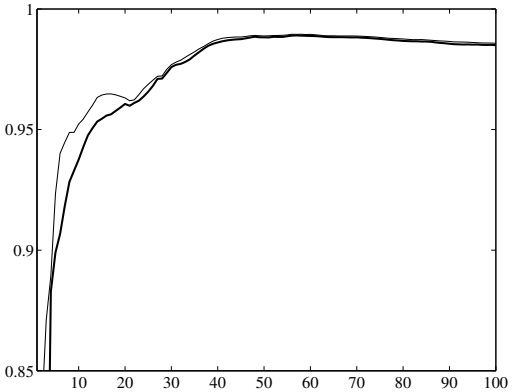
гипертония

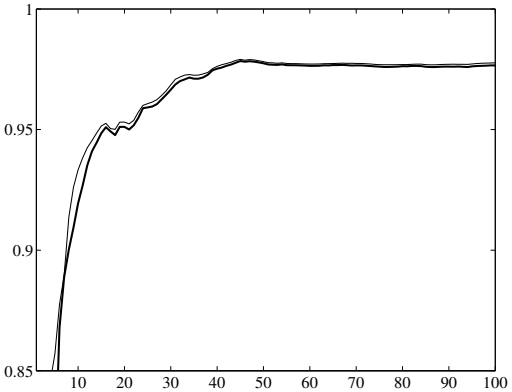
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

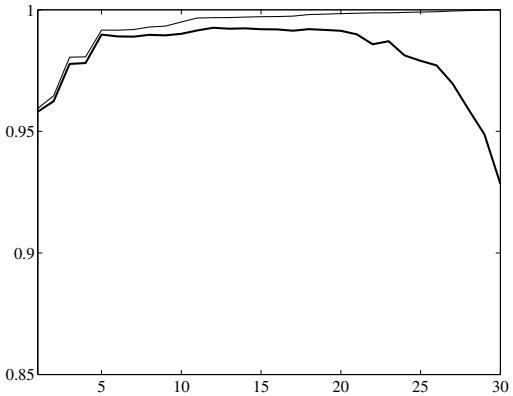
Синдромный алгоритм (K — число признаков):

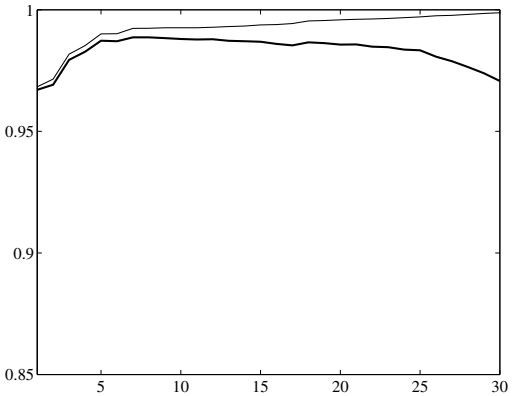


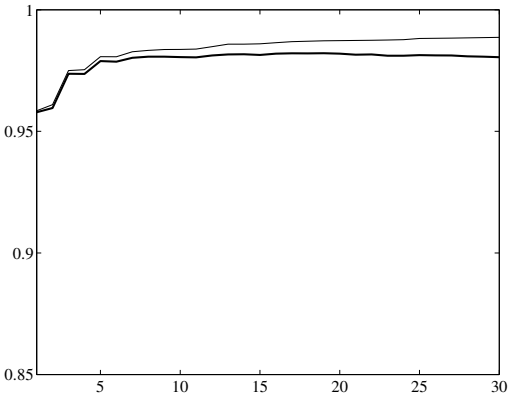




Логистическая регрессия (K — число главных компонент):







асептический некроз ГБК желчнокаменная болезнь ишемическая болезнь

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

Выводы

- многие болезни можно диагностировать по ЭКГ,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- причём точность диагностики увеличивается с ростом времени регистрации ЭКГ, и нескольких сотен кардиоциклов уже хватает для распознавания
- методы машинного обучения ведут себя по-разному,
- некоторые подвержены эффекту переобучения

Задача обучения по прецедентам (машинного обучения)

$\mathbb{X}$ — объекты; $\mathbb{Y}$ — ответы (классы, прогнозы);
$y^*: \mathbb{X} \to \mathbb{Y}$ — неизвестная зависимость.

Дано: $x_i = \big(f_1(x_i), \ldots, f_n(x_i)\big)$ — *обучающие объекты* с известными ответами $y_i = y^*(x)$, $i = 1, \ldots, \ell$:

$$\begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: алгоритм $a: \mathbb{X} \to \mathbb{Y}$, способный давать правильные ответы на новых объектах $\tilde{x}_i = \big(f_1(\tilde{x}_i), \ldots, f_n(\tilde{x}_i)\big)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} f_1(\tilde{x}_1) & \ldots & f_n(\tilde{x}_1) \\ \ldots & \ldots & \ldots \\ f_1(\tilde{x}_k) & \ldots & f_n(\tilde{x}_k) \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - x — характеристики недвижимости; y — цена;

Внимание, конкурс!

Дано:
матрица «объекты–признаки» по одной болезни
(некроз головки бедренной кости),
первый столбец — метки классов (0–здоровый, 1–больной),
остальные столбцы — 216 признаков,
строки — объекты (99 здоровых, 153 больных)

Найти:
оценки объектов тестовой выборки, 253 объекта

Критерий: AUC (площадь под ROC-кривой)

Задача давалась на VI Традиционной молодёжной летней школе «Управление, информация и оптимизациия» 26-06-2014
http://www.MachineLearning.ru/wiki?title=User:Vokov
http://www.machinelearning.ru/wiki/images/e/e1/School-VI-2014-task-3.rar

Литература

William B. Cavnar , John M. Trenkle. N-Gram-Based Text Categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994. Pp. 161–175.

Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008. 116 с.

Успенский В. М. Информационная функция сердца. *Клиническая медицина*. 2008. Т. 86. № 5. С. 4–13.

Благодарности

- профессору **Вячеславу Максимилиановичу Успенскому** — за предоставленные данные многолетних исследований и плодотворное сотрудничество
- студентке 6 курса кафедры «Интеллектуальные системы» ФУПМ МФТИ **Владе Целых** — за проведение всех вычислительных экспериментов в этой презентации

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Если что-то было не понятно,
не стесняйтесь спрашивать :)