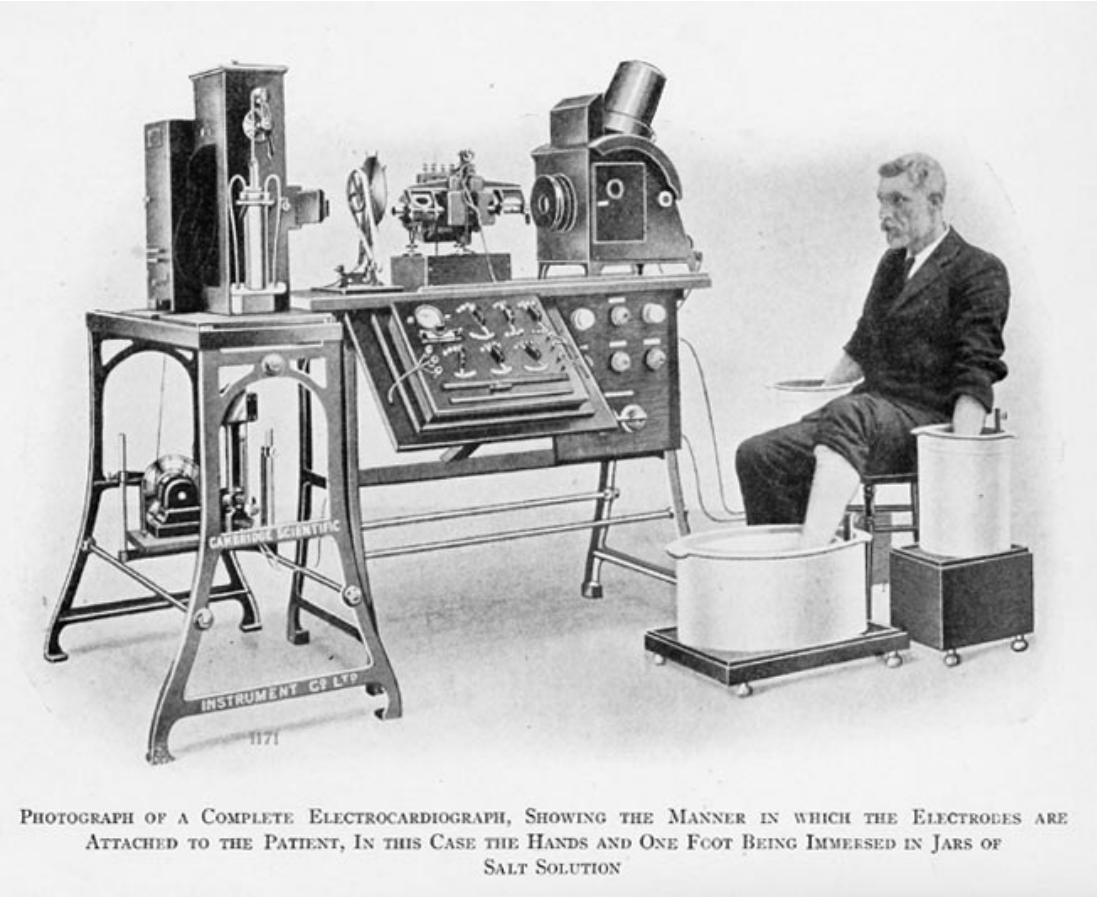
ПРОЕКТ
Медицинская диагностика по электрокардиограмме

Воронцов Константин Вячеславович
ФУПМ МФТИ • ВМК МГУ • Яндекс • FORECSYS

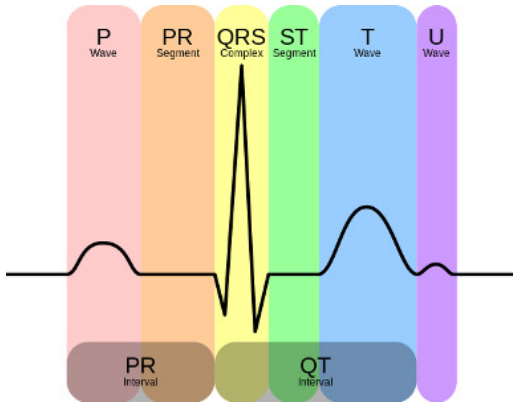
Семинар по проекту • 6 июля 2016
Сочи, Сириус • Проектная смена • 1–24 июля 2016

Содержание

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

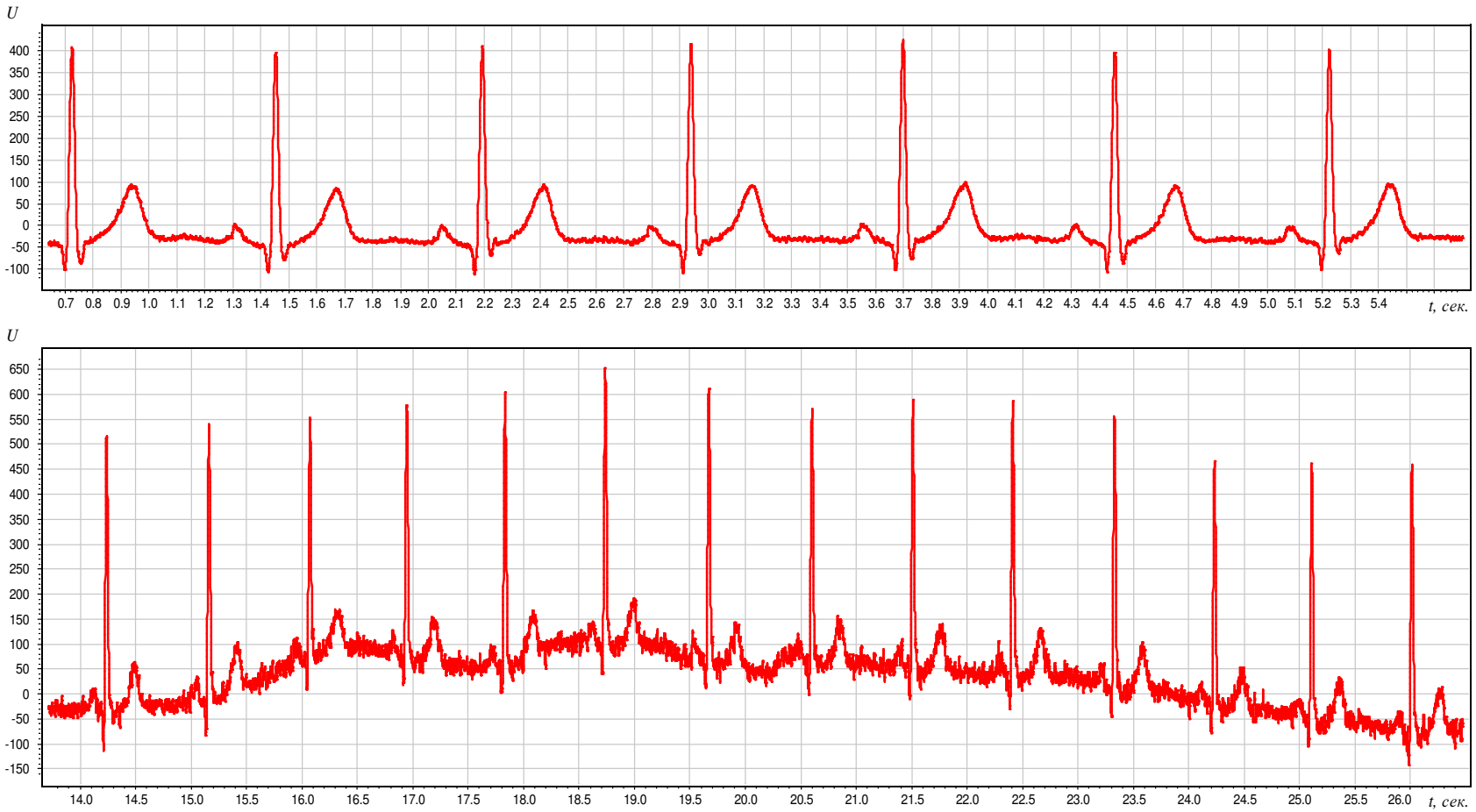


1872 — первые записи электрической активности сердца

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:

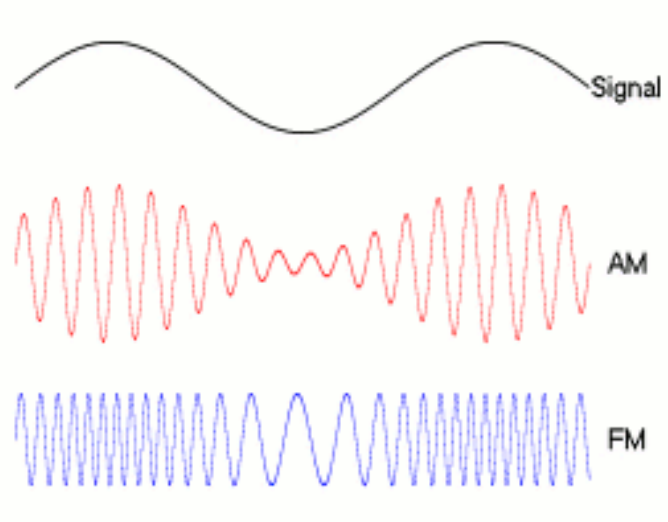
- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Появление цифровой электрокардиографии

Предположения:

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

Краткий экскурс 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.

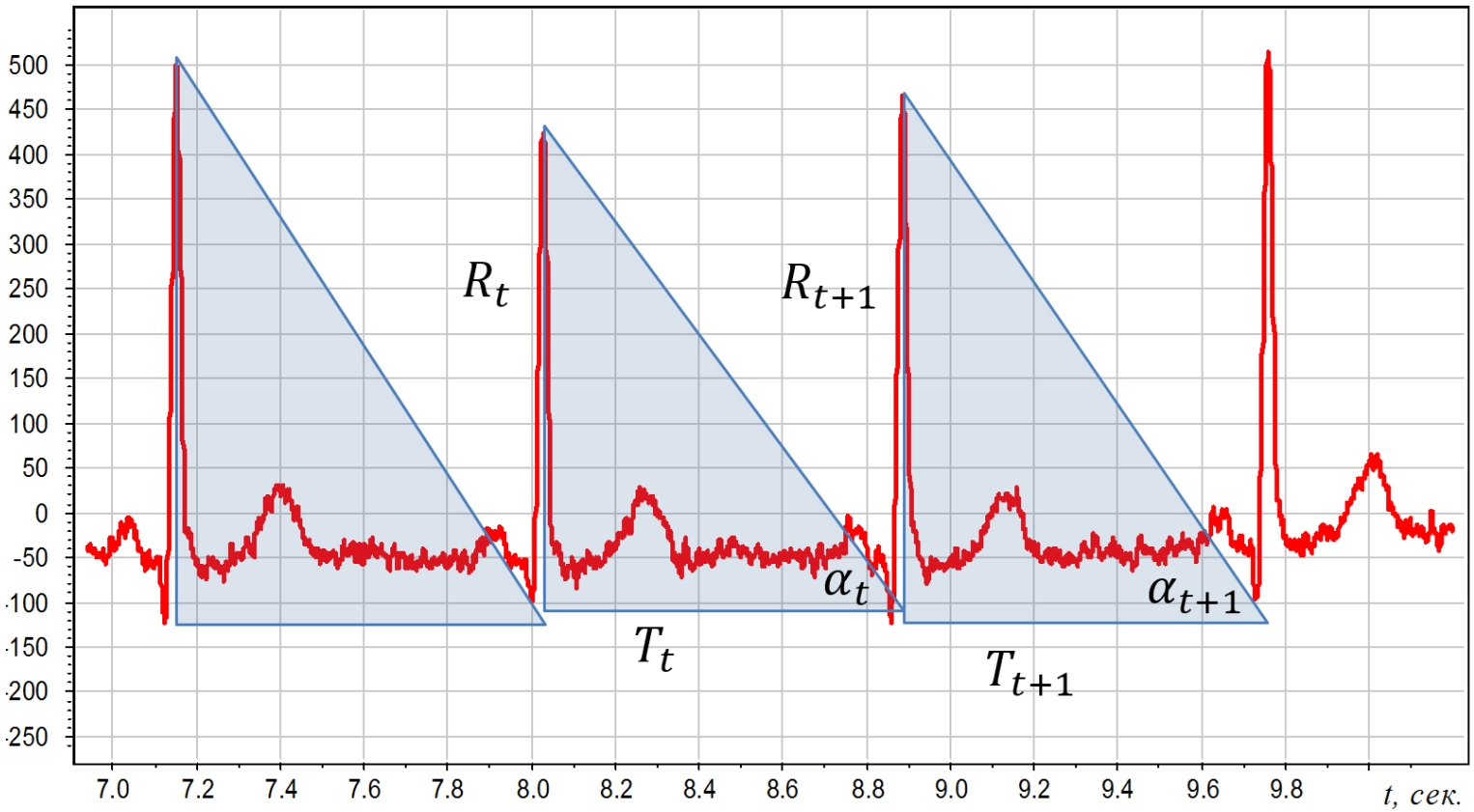
В случае ЭКГ несущая частота — биения сердца, ~ 1 Гц
А что будет аналогом модуляции и демодуляции?

Вариабельность интервалов и амплитуд кардиоциклов

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

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

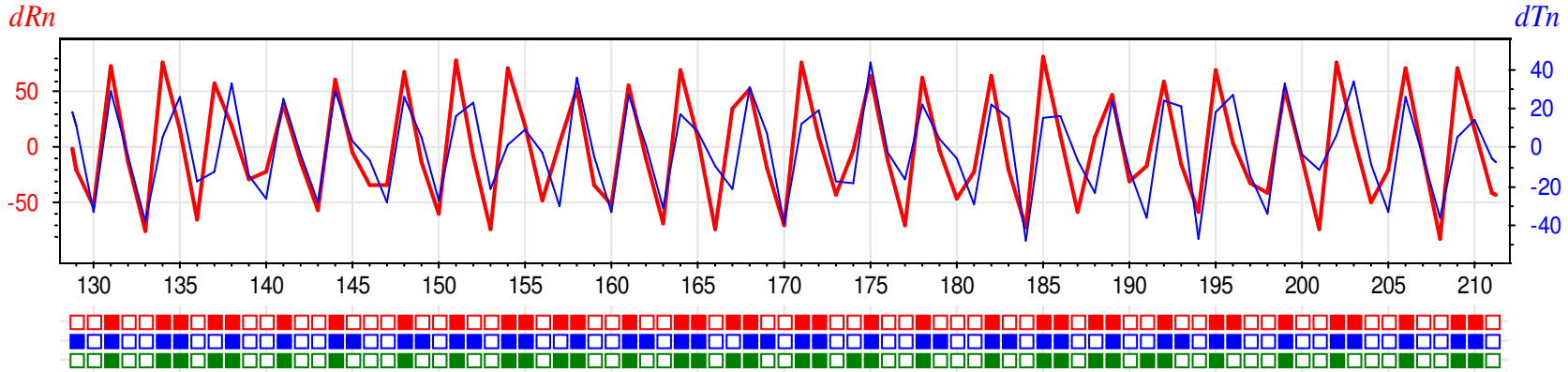
приращение углов: $d\alpha_t = \alpha_{t+1} - \alpha_t, \quad \alpha_t = \operatorname{arctg} \frac{R_t}{T_t}$



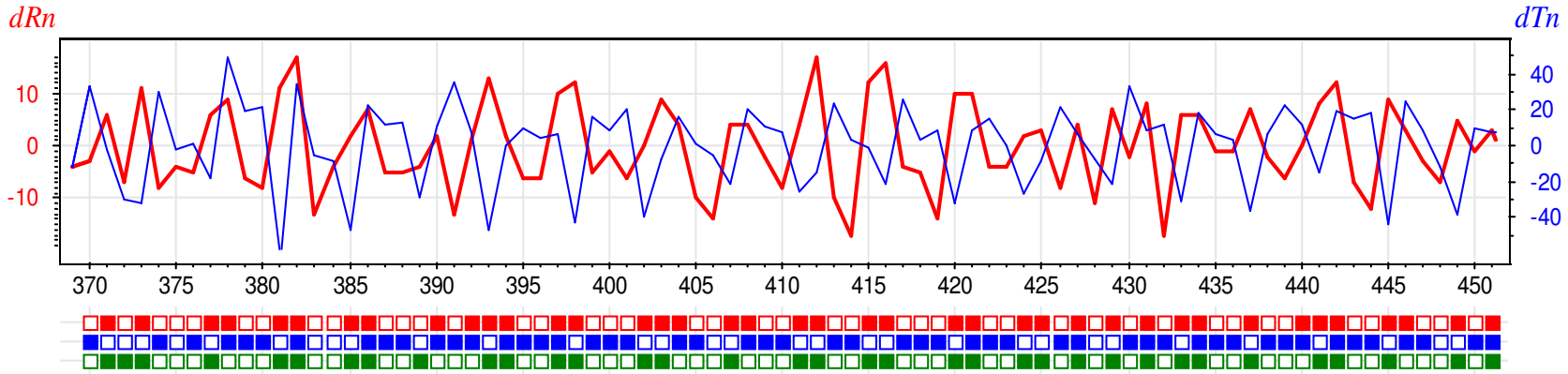
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:



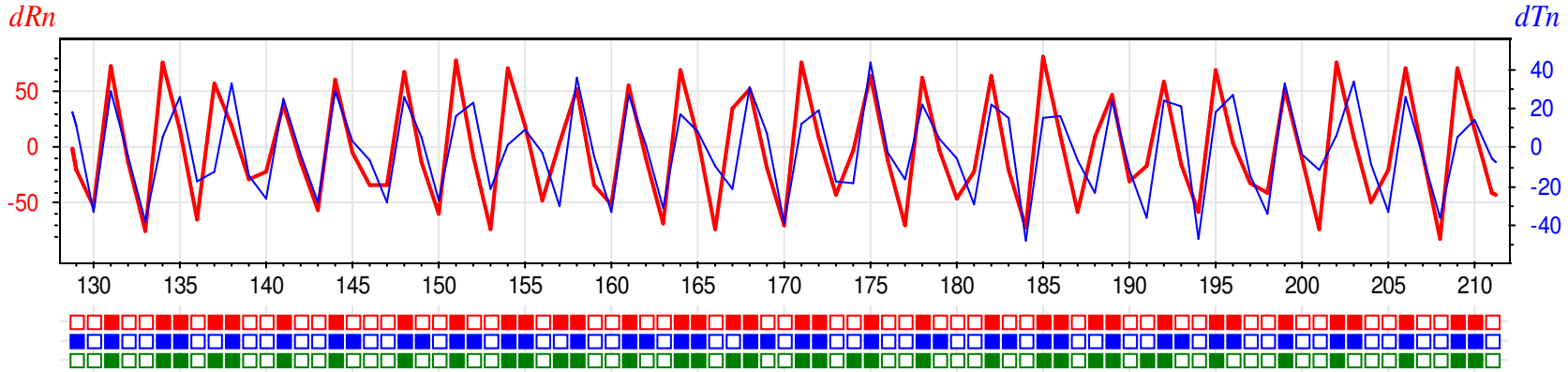
Больной (язвенная болезнь):



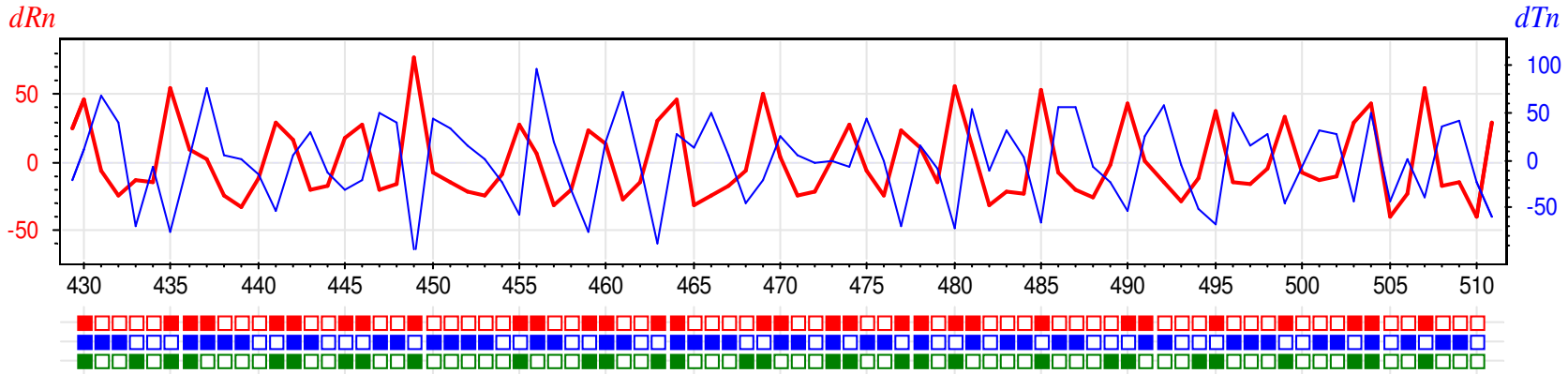
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:



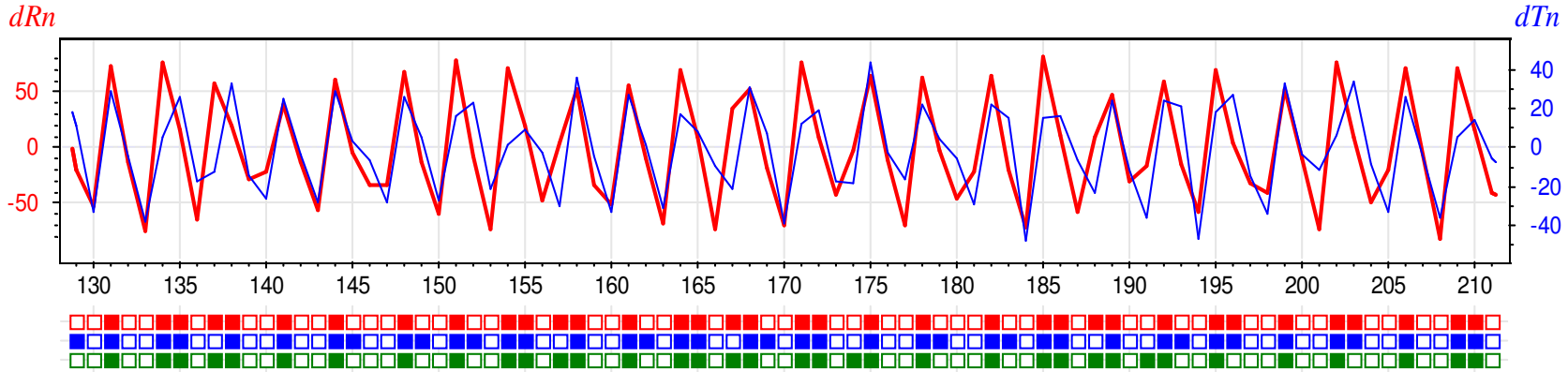
Больной (гипертония):



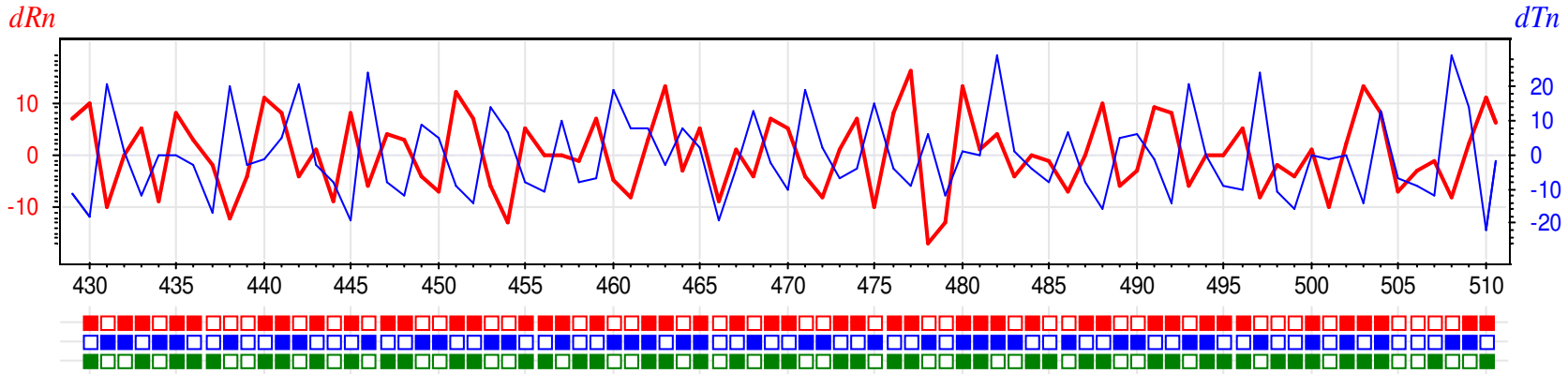
Есть ли различия в знаках приращений у больных и здоровых?

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

Здоровый:

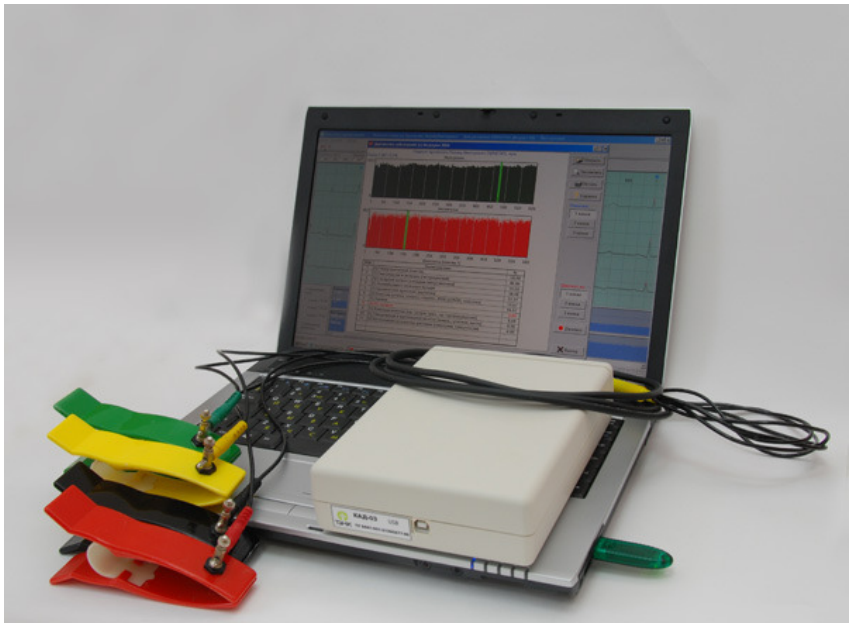


Больной (рак):



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

Технология информационного анализа ЭКГ по В.М.Успенскому

Этапы предварительной обработки ЭКГ-сигнала:

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

Этапы машинного обучения:

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

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*:

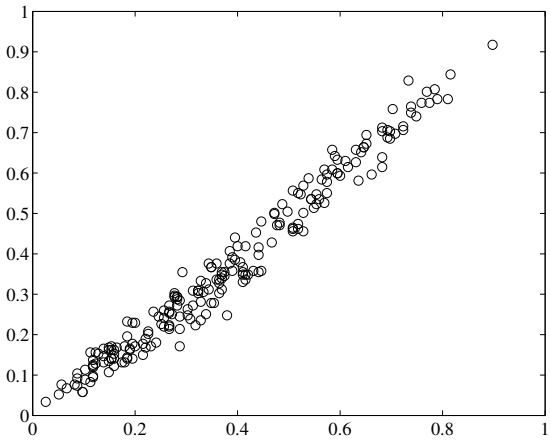
Частоты триграмм — число вхождений триграммы в кодограмму:

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

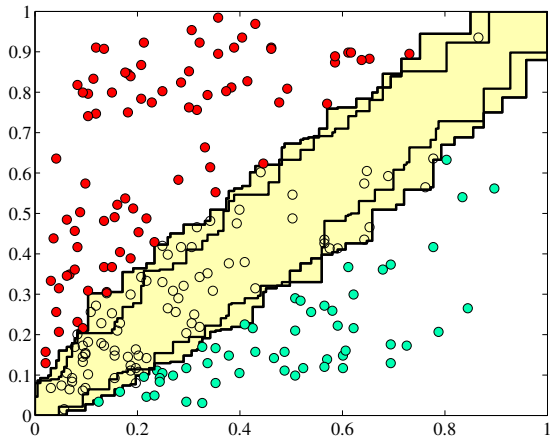
Существуют сочетания триграмм, специфичные для болезней

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

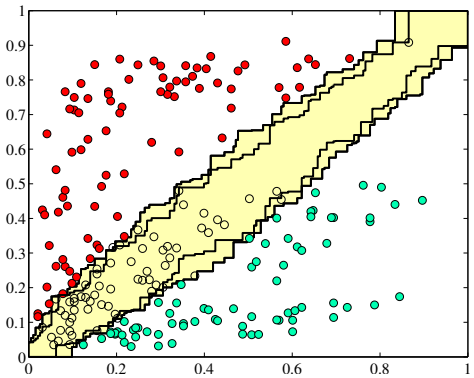


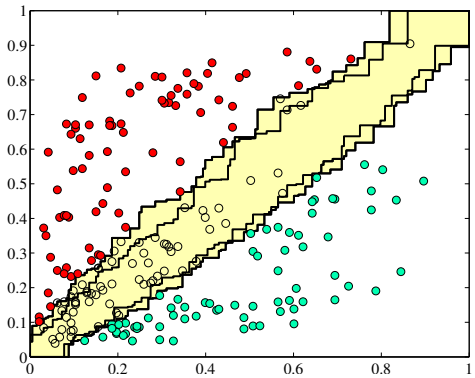
наблюдаемые y_i

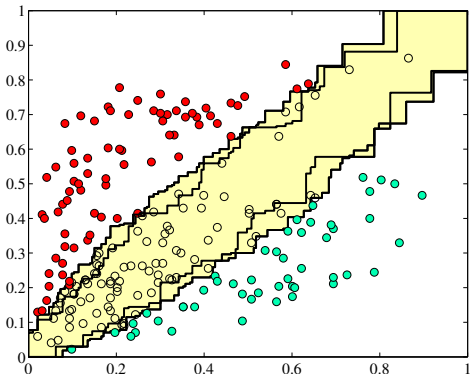
Слева: как распределятся точки, если объектам x_i назначить случайные (случайно перемешанные) метки классов y_i.
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

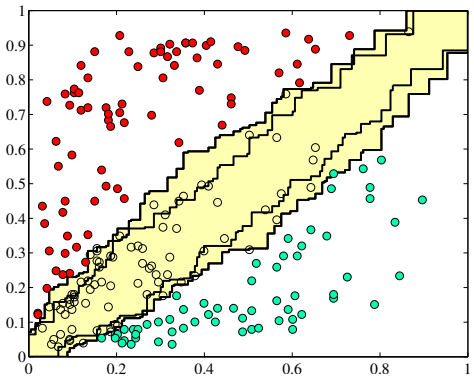
Существуют сочетания триграмм, специфичные для болезней

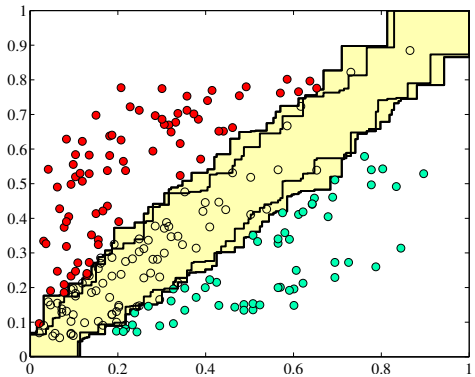
Для каждой болезни есть свои неслучайно частые триграммы

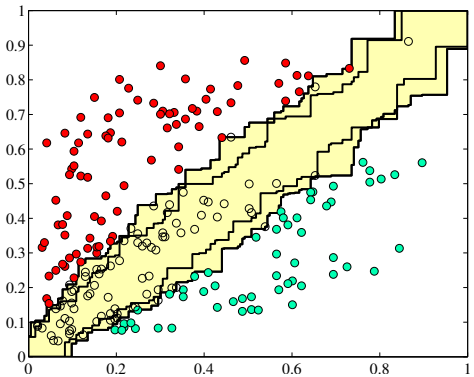

ишемия сердца


гипертония


рак


желчнокаменная болезнь

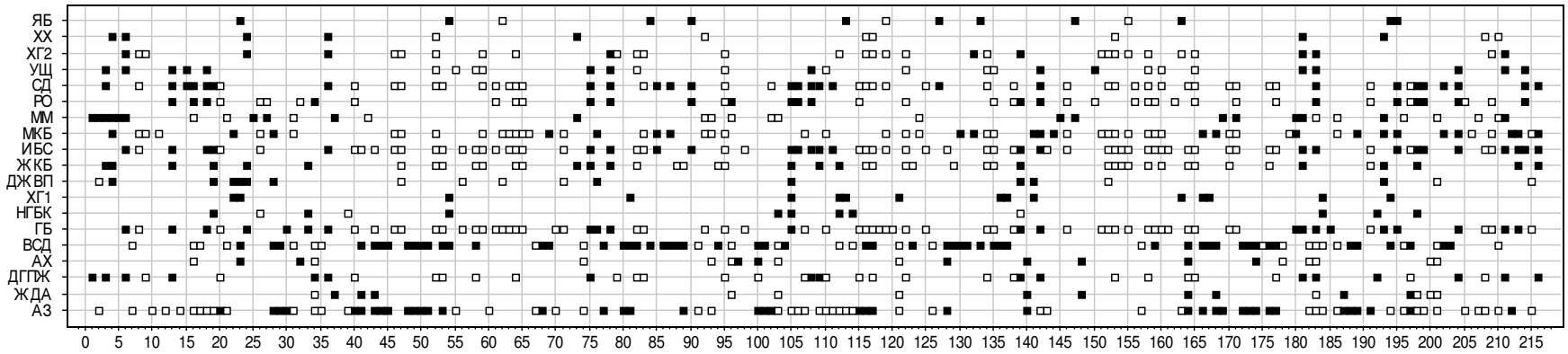

миома матки


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

Вывод 2. *Диагностический эталон* болезни — специфичное подмножество триграмм с неслучайно высокой частотой

Задача статистического (машинного) обучения с учителем

Восстановление зависимости $y = f(x)$ по *обучающей выборке* $(x_i, y_i)_{i=1}^{\ell}$, объекты $x_i = (x_i^1, \ldots, x_i^n)$, ответы $y_i = f(x_i)$.

Этап обучения (train)

Метод обучения μ строит алгоритм a:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} a$$

Этап применения (test)

Алгоритм a выдаёт ответы $a(\tilde{x}_i)$ для новых объектов $\tilde{x}_1, \ldots, \tilde{x}_k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_\ell^1 & \ldots & \tilde{x}_\ell^n \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Линейная модель классификации с двумя классами

Для простоты рассмотрим только одно заболевание:
$y_i = 1$ — больной, $y_i = 0$ — здоровый

- чем выше частота триграммы x^j, тем она информативнее
- есть триграммы, более характерные для больных, и есть триграммы, более характерные для здоровых

Линейная модель классификации:

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

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

Бинаризация признаков

Эвристика: важно, что триграмма часто встречается,
но не так важно, *настолько* часто.

Исходные целочисленные признаки:
x_i^j — сколько раз j-я триграмма встретилась в i-й кодограмме

Бинаризованные признаки:
$b_i^j = [x_i^j \geqslant A]$ (позже выяснилось, что лучше брать $A = 2$)

Число обучающих объектов класса y, для которых $b_i^j = z$:

$$N_{yz}^j = \sum_{i=1}^{\ell} [y_i = y]\,[b_i^j = z]$$

N_{11}^j — число больных, у которых j-я триграмма частая

N_{01}^j — число здоровых, у которых j-я триграмма частая

Выбор весов

Эвристика: вес j-й триграммы тем больше,

1) чем больше N_{11}^j и N_{00}^j,

2) чем меньше N_{01}^j и N_{10}^j,

Можно пробовать разные формулы для весов:

$$w_j = \frac{N_{11}^j}{N_{01}^j} \qquad w_j = \frac{N_{11}^j N_{00}^j}{N_{01}^j N_{10}^j}$$

$$w_j = \log \frac{N_{11}^j}{N_{01}^j} \qquad w_j = \log \frac{N_{11}^j N_{00}^j}{N_{01}^j N_{10}^j}$$

$$w_j = \sqrt{N_{11}^j} - \sqrt{N_{01}^j} \qquad w_j = \sqrt{N_{11}^j N_{00}^j} - \sqrt{N_{01}^j N_{10}^j}$$

... и разрешается фантазировать!

Отбор признаков

Гипотеза: если в линейный классификатор $\langle x, w \rangle$ добавить кучу неинформативных признаков, никак не связанных с данной болезнью, то получится $\langle x, w \rangle$ + шум.

Эвристика 1: отсортировать признаки по убыванию весов $|w_j|$ и взять первые K лучших. Для остальных положить $w_j = 0$.

Эвристика 2: отсортировать по весам из одной формулы, а в классификаторе использовать веса из другой формулы.

Модификации модели (примочки, костыли,...), сделанные из нестрогих соображений, по-научному называются *эвристиками*.

Эвристическое мышление в прикладных исследованиях необходимо, оно ближе к мышлению физика, чем математика.

Итак, наш первый алгоритм обучения!

Вход: выборка $(x_i, y_i)_{i=1}^{\ell}$; $\mathscr{A} = \{A_1, ..., A_N\}$, $\mathscr{K} = \{K_1, ..., K_M\}$
Выход: веса w_j, параметры A, K

для всех значений параметров $A \in \mathscr{A}$, $K \in \mathscr{K}$
    вычислить N_{yz}^j для всех $j = 1, \ldots, n$, $y, z \in \{0, 1\}$;
    вычислить веса w_j для всех признаков $j = 1, \ldots, n$;
    отсортировать признаки по убыванию $|w_j|$;
    обнулить веса признаков $w_{K+1}, \ldots, w_n$;
    оценить качество классификации $Q(A, K)$;
оставить A, K, при которых $Q(A, K) \to \max$;

Перебирать можно также виды формул весов, порог w_0, и другие параметры (если их придумать)

Следующий вопрос: как измерять качество?

Терминология диагностики

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.

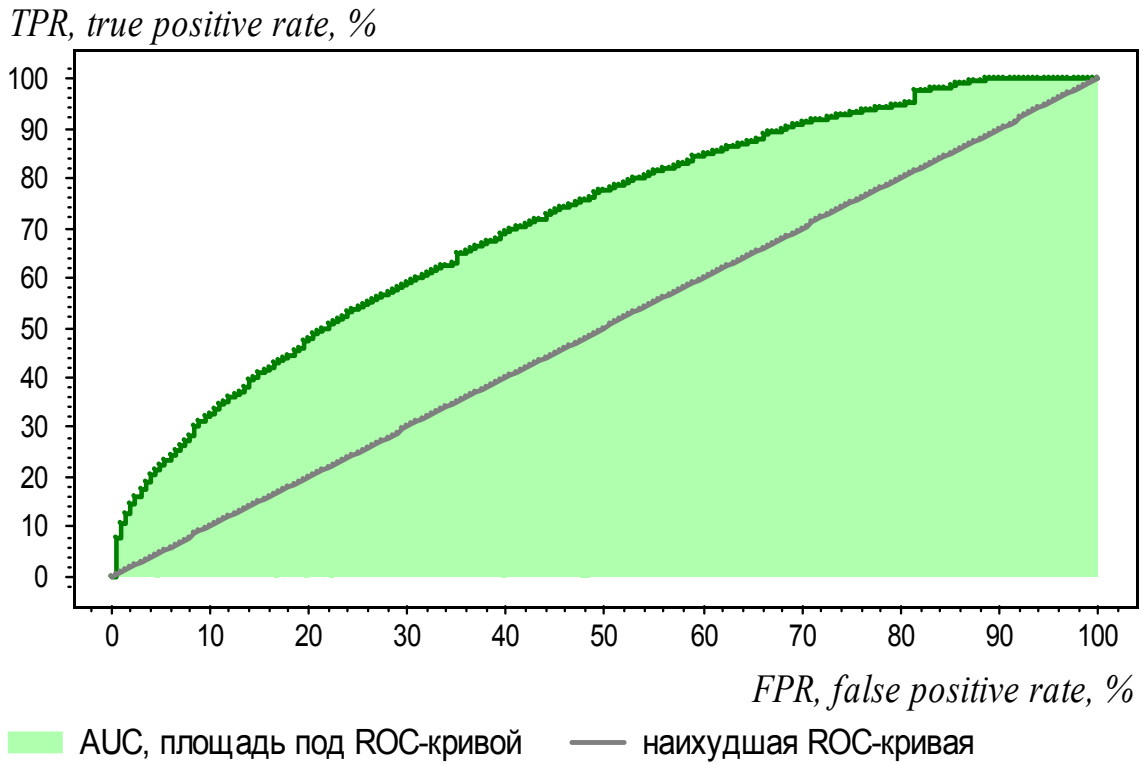
⊕ Они не зависят от соотношения мощностей классов.

⊕ Хорошо подходят для несбалансированных выборок.

Определение ROC-кривой

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»),



AUC — площадь под ROC-кривой

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

AUC равна доле правильно упорядоченных пар (x_i, x_j) (докажите):

$$\text{AUC} = \frac{1}{\ell_0 \ell_1} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \big[y_i < y_j\big] \big[\langle x_i, w\rangle < \langle x_j, w\rangle\big]$$

Преимущества AUC:

$\oplus$ не зависит от порога w_0, оценивает только качество w;

$\oplus$ не зависит от численности классов;

$\oplus$ это общепринятая мера качества классификации;

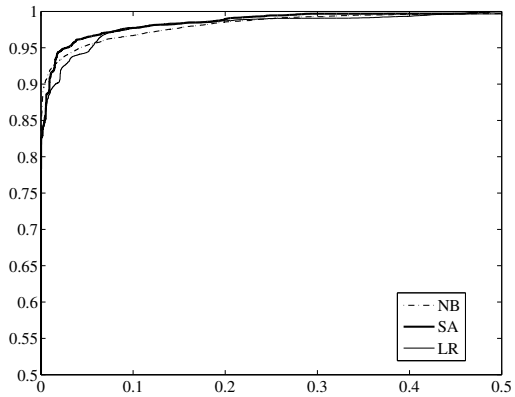
Метод обучения μ по выборке $(x_i, y_i)_{i=1}^{\ell}$ строит алгоритм a. Чтобы измерить предсказательную способностью μ, будем вычислять AUC на контрольной выборке.

Результаты кросс-валидации

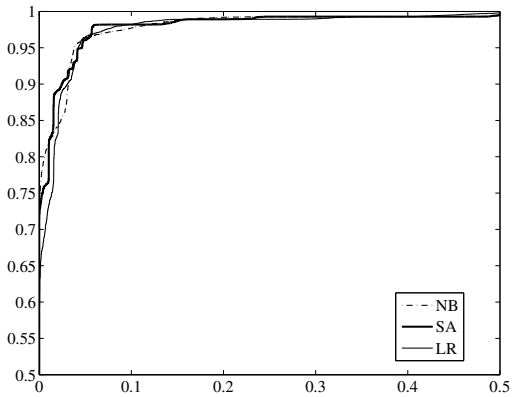
Обучающая выборка — для оптимизации параметров модели
Тестовая выборка — для оценивания чувс., спец., AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	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

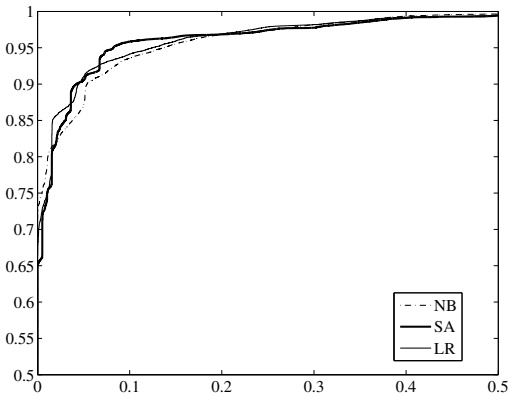
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



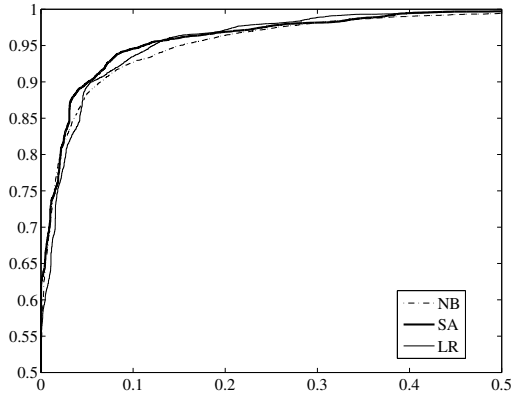
асептический некроз ГБК



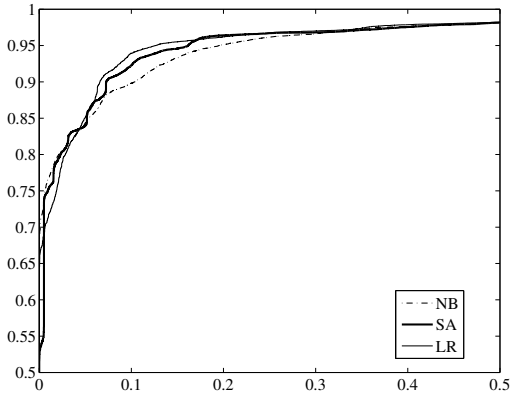
желчнокаменная болезнь



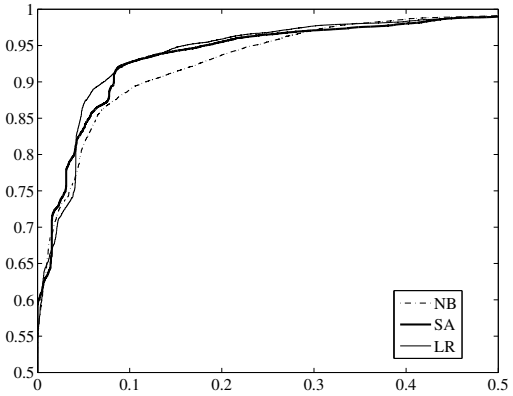
ишемическая болезнь



хронический гастрит 1



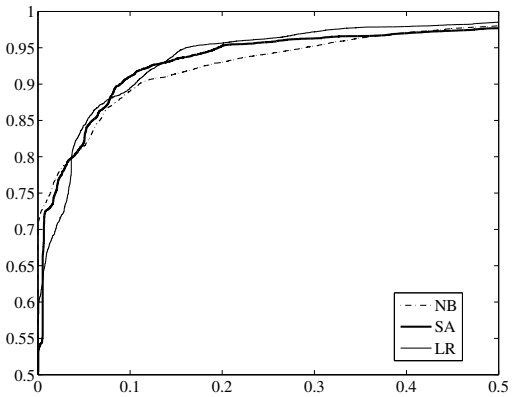
сахарный диабет



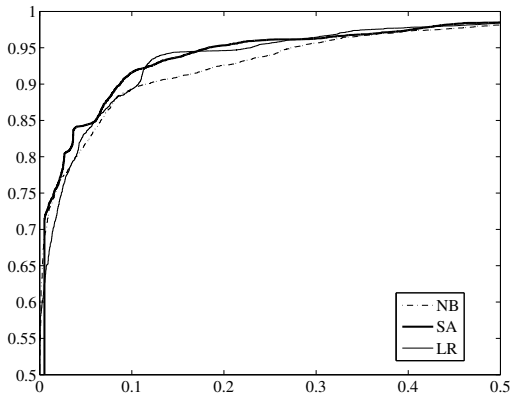
гипертония

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

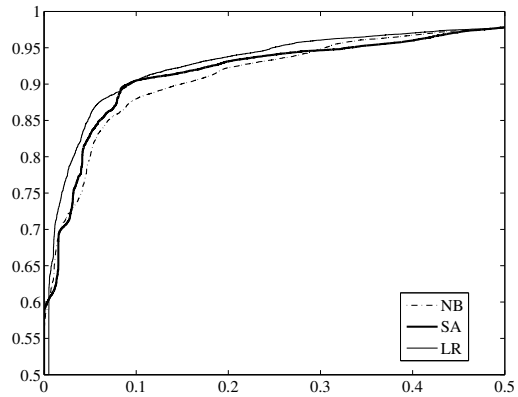
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



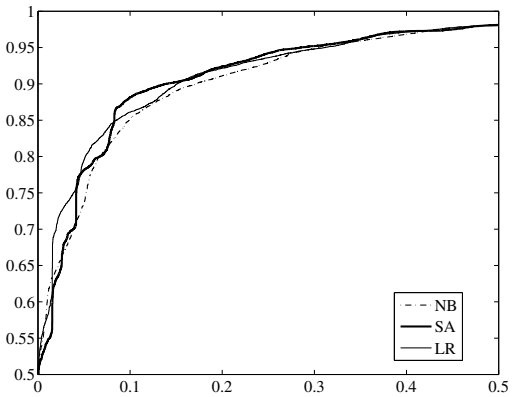
рак общий



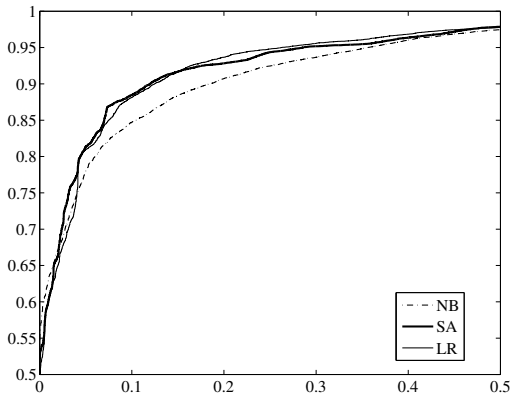
аденома простаты



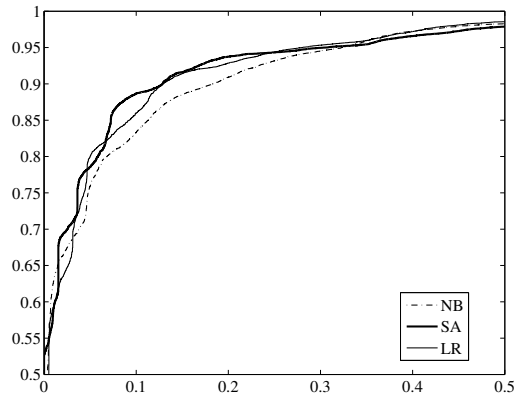
зоб щитовидной железы



хронический гастрит 2



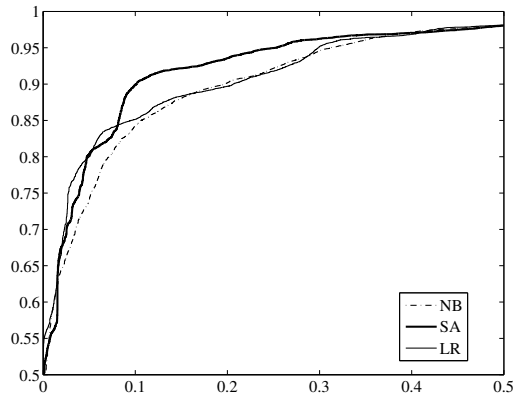
дискинезия ЖВП



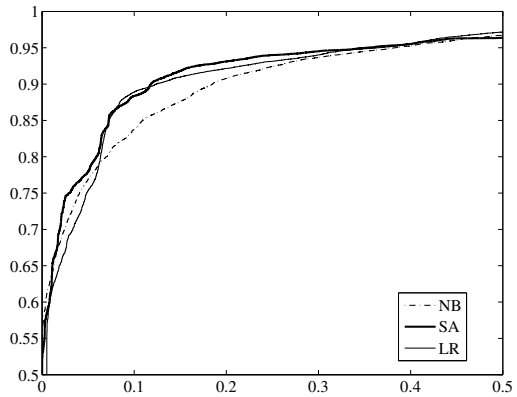
мочекаменная болезнь

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

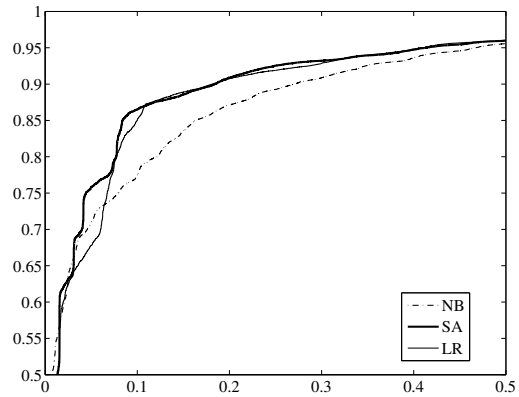
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



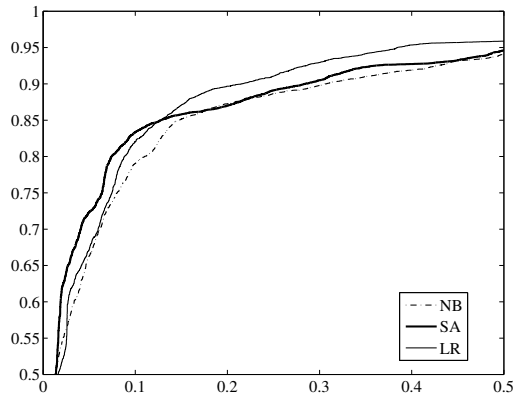
хронический холецистит



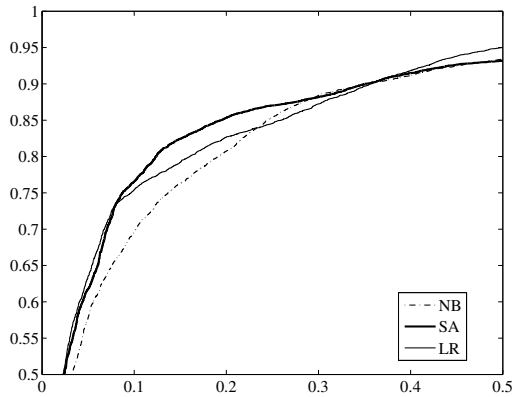
язвенная болезнь



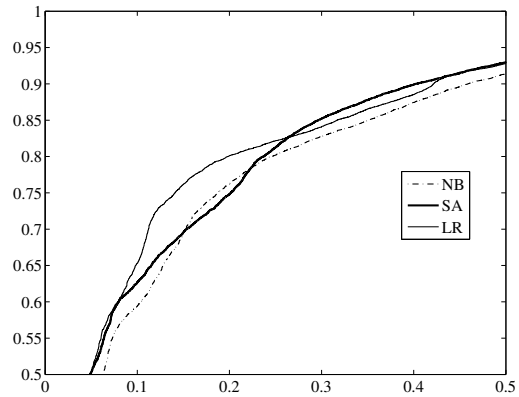
миома матки



хронический аднексит



анемия

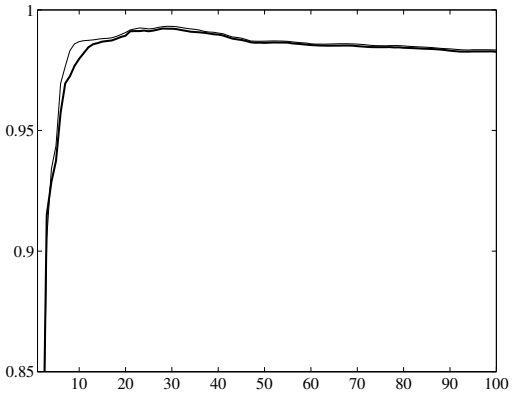


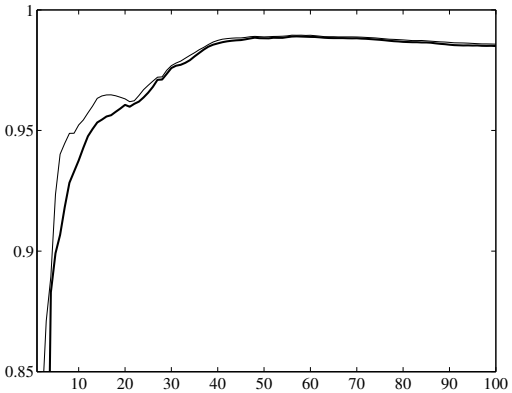
вегетососудистая дистония

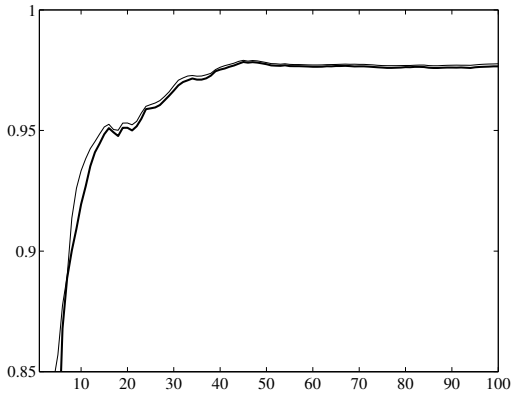
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

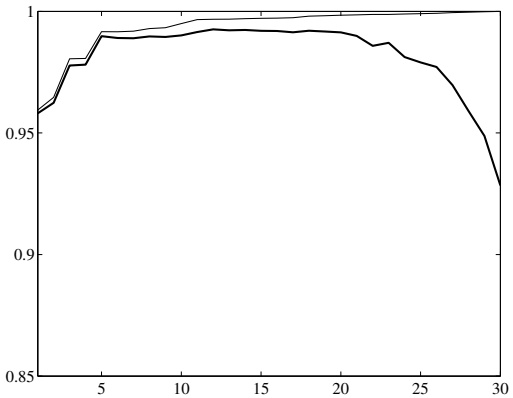
Синдромный алгоритм (наивный Байес на K признаках):

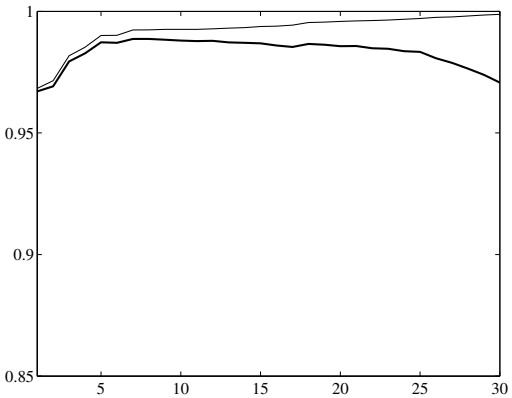


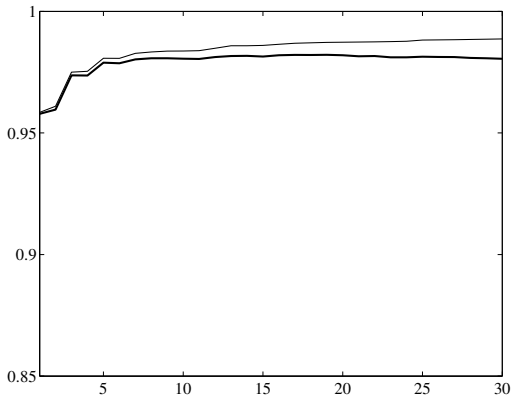




Логистическая регрессия (K — число главных компонент):





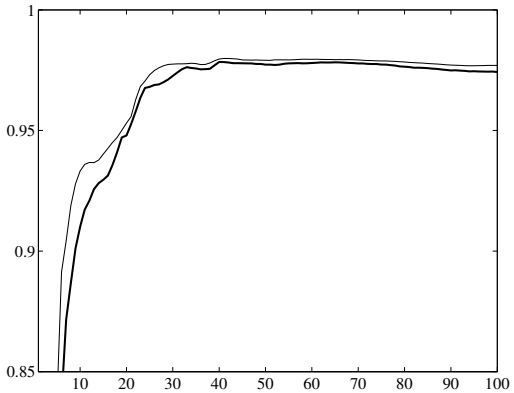


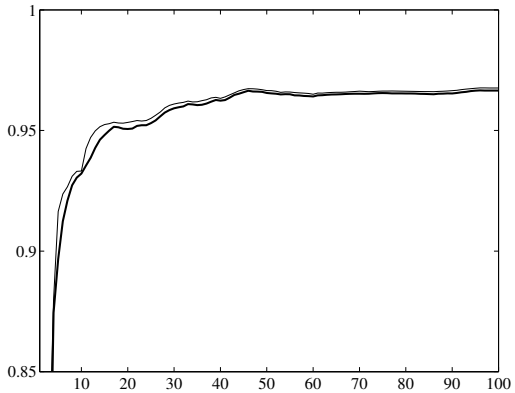
асептический некроз ГБК желчнокаменная болезнь ишемическая болезнь

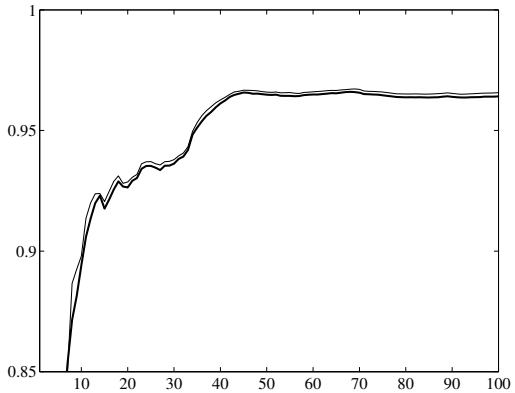
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

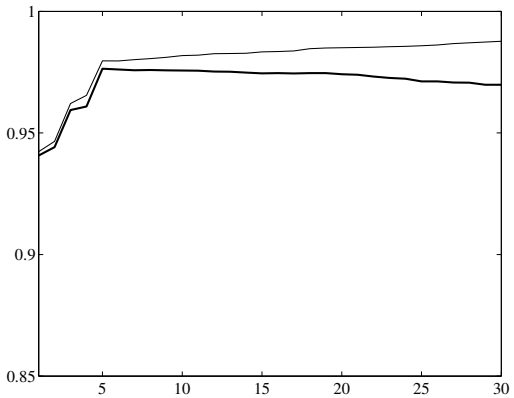
Синдромный алгоритм (K — число признаков):

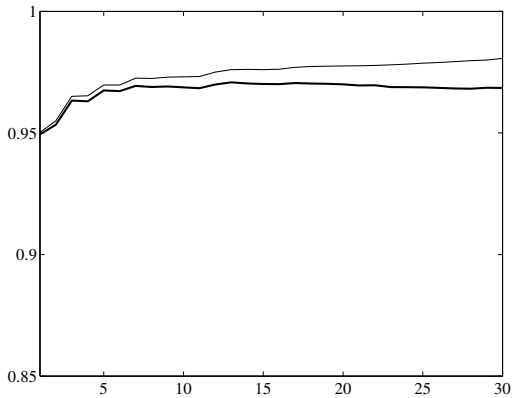


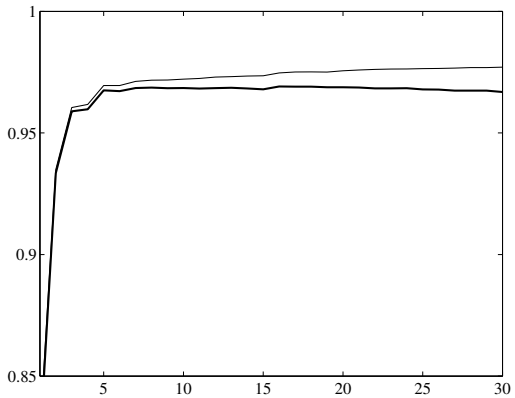




Логистическая регрессия (K — число главных компонент):





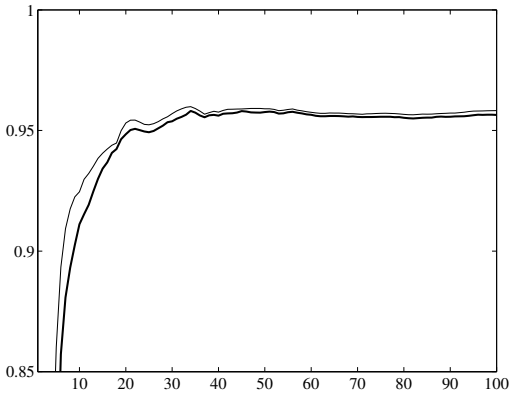


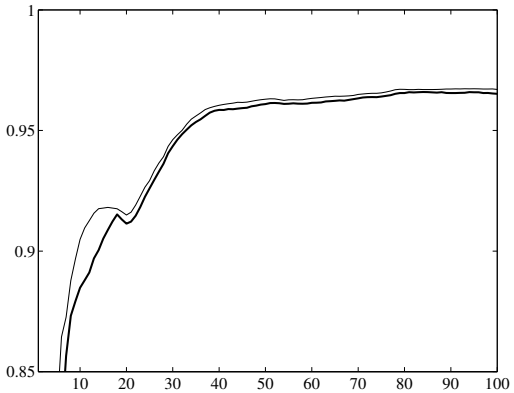
хронический гастрит 1 | сахарный диабет | гипертония

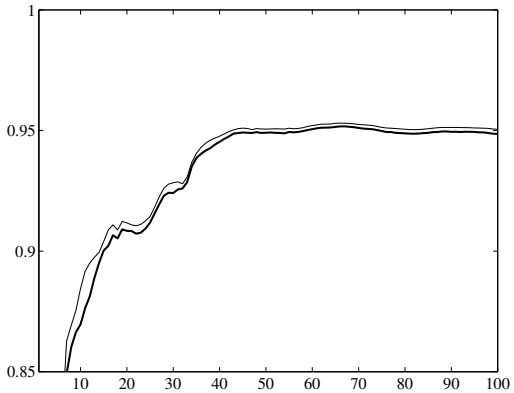
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

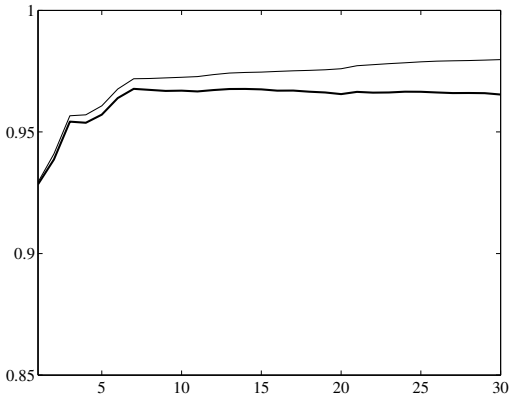
Синдромный алгоритм (K — число признаков):

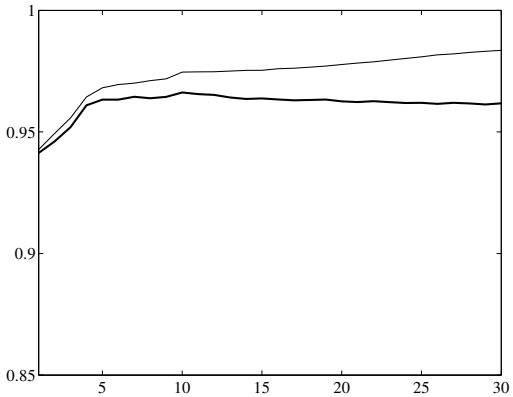


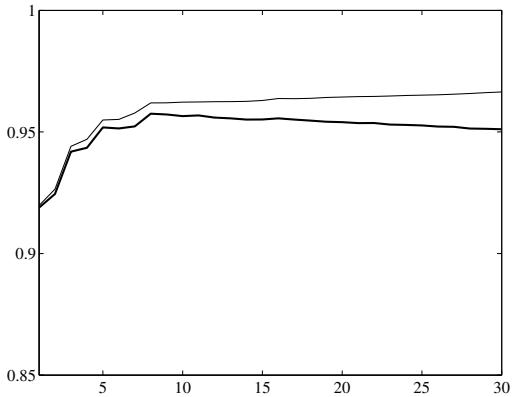




Логистическая регрессия (K — число главных компонент):





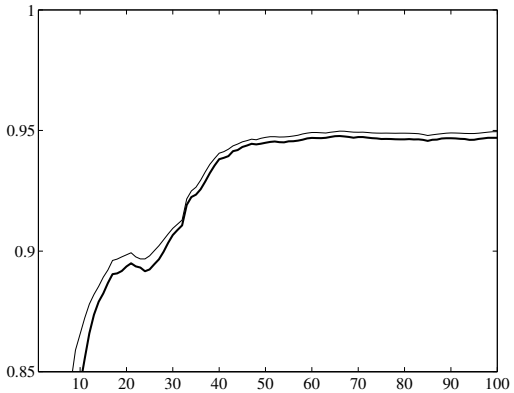


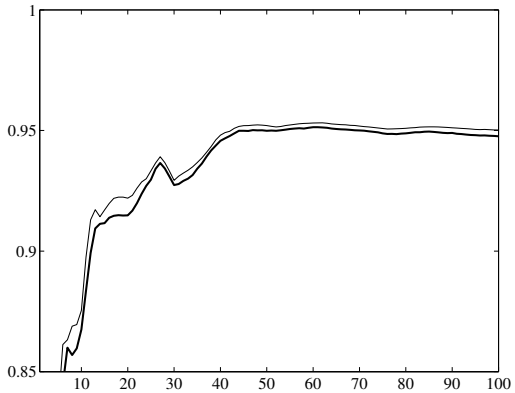
рак общий | аденома простаты | зоб щитовидной железы

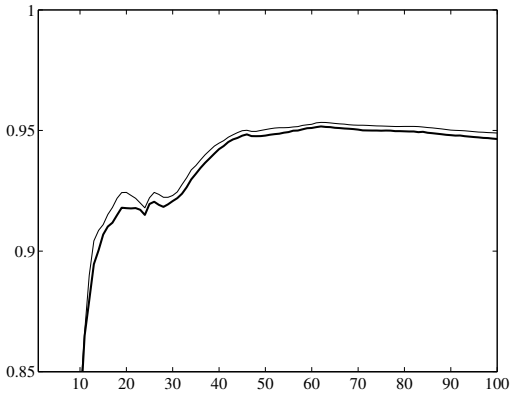
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

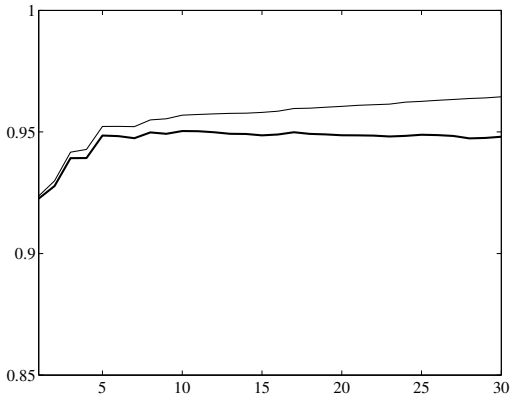
Синдромный алгоритм (K — число признаков):

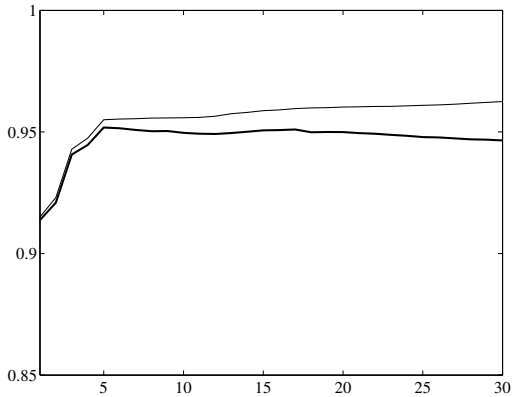


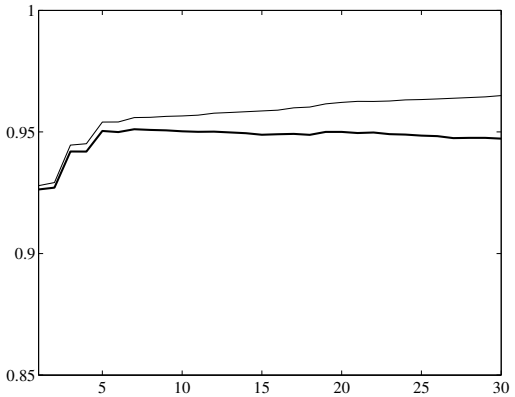




Логистическая регрессия (K — число главных компонент):





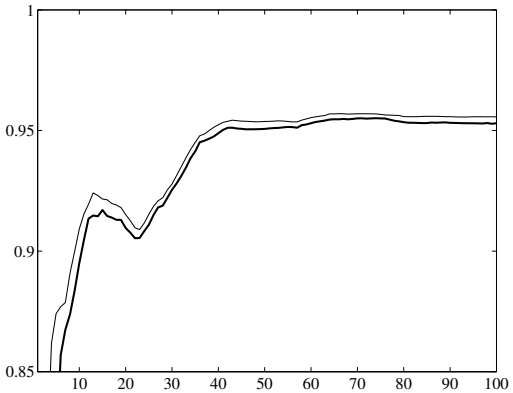


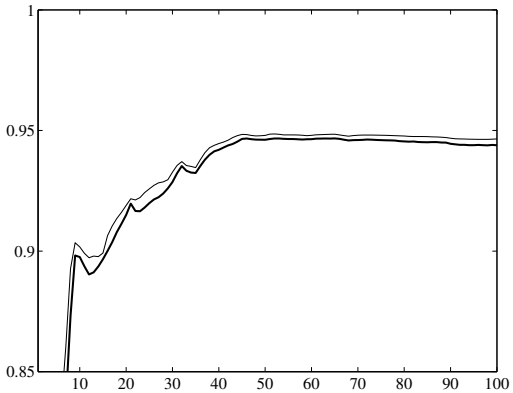
хронический гастрит 2 дискинезия ЖВП мочекаменная болезнь

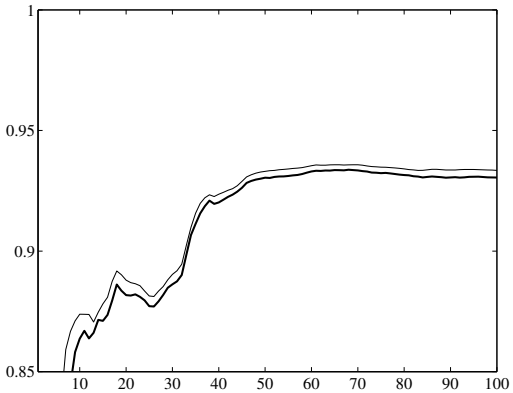
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

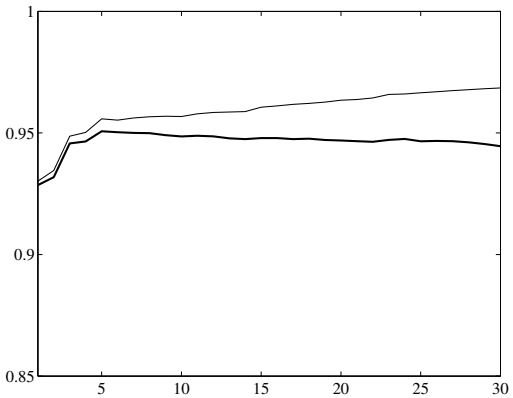
Синдромный алгоритм (K — число признаков):

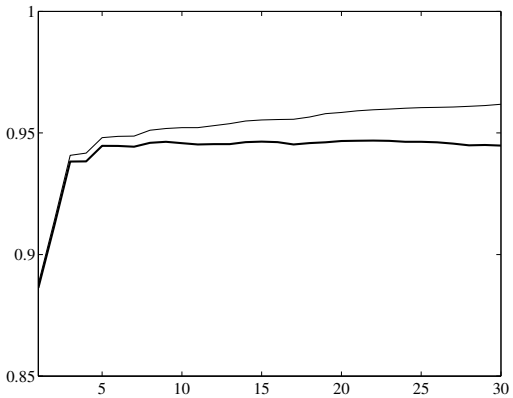


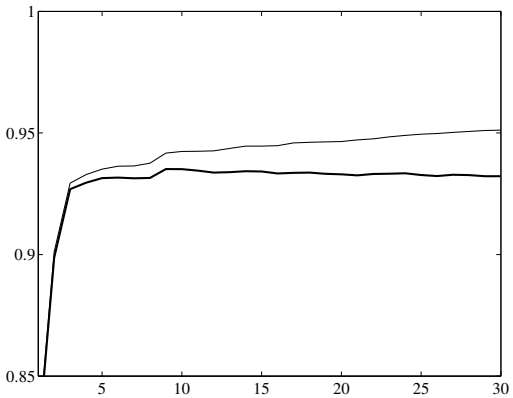




Логистическая регрессия (K — число главных компонент):





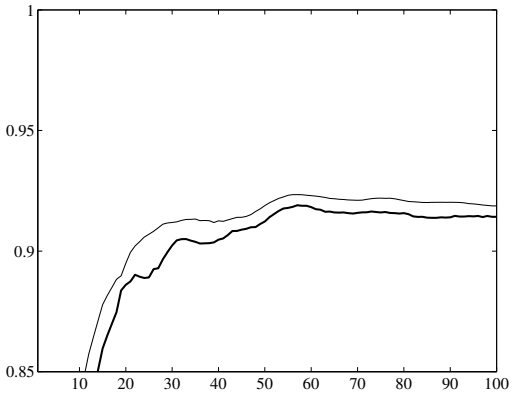


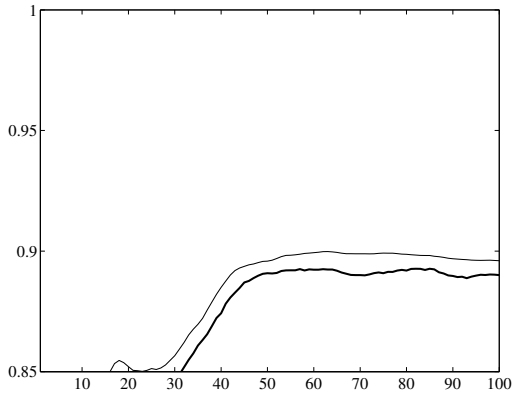
хронический холецистит | язвенная болезнь | миома матки

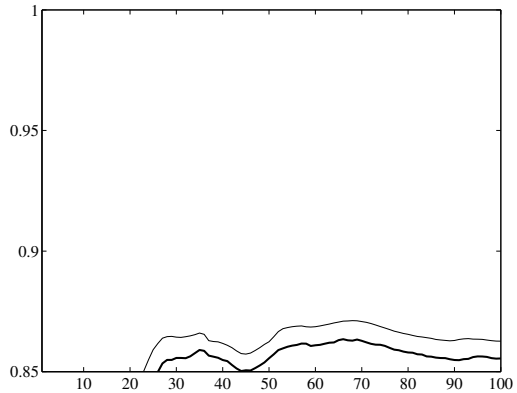
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

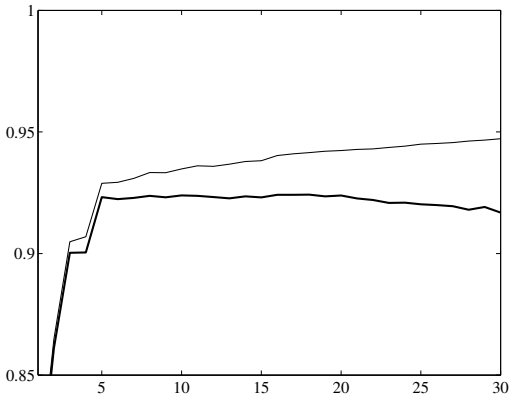
Синдромный алгоритм (K — число признаков):

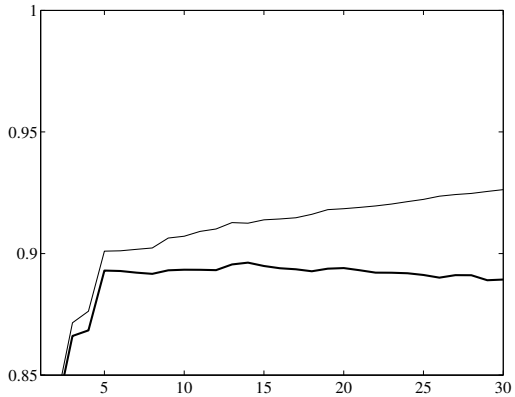


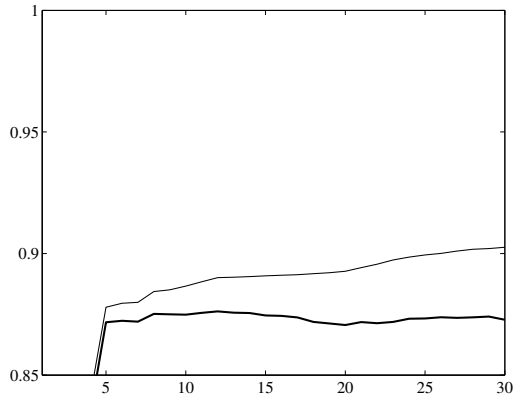




Логистическая регрессия (K — число главных компонент):







хронический аднексит | анемия | вегетососудистая дистония

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

Ваши цели в этом проекте

- добиться улучшения точности диагностики
- освоить программирование в Python
- освоить «машинное обучение из коробки»
- научиться приёмам работы с данными
- научиться «смотреть на данные»
- научиться проводить мозговые штурмы
- научиться оформлять результаты исследований
- выступить на конференции

Кооперация со смежными проектами:

- «Антитела» — алгоритмы множественного выравнивания
- «Случайные явления» — фрактальный анализ ЭКГ
- «Оптимизация» — градиентные методы

Наши планы

Первая фаза проекта — конкурс идей

- аккуратно реализовать наивный линейный классификатор
- добавить отбор признаков
- добавить метод стохастического градиента
- добавить максимизацию AUC
- добавить регуляризацию
- добавить признаки множественного выравнивания
- добавить фрактальные признаки ВСР-анализа

Вторая фаза проекта — кооперация

- поиск успешных сочетаний идей
- решение проблемы многоклассовой классификации

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Когда что-то не понятно,
не стесняйтесь подходить и спрашивать :)