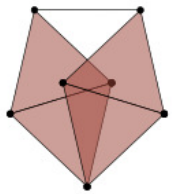
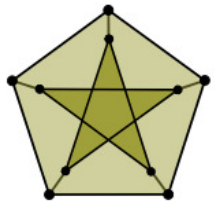
Теория и практика машинного обучения
• Лекция 1 •
Задача диагностики заболеваний по электрокардиограмме

Воронцов Константин Вячеславович
МФТИ • МГУ • ВШЭ • ВЦ РАН • Яндекс • FORECSYS



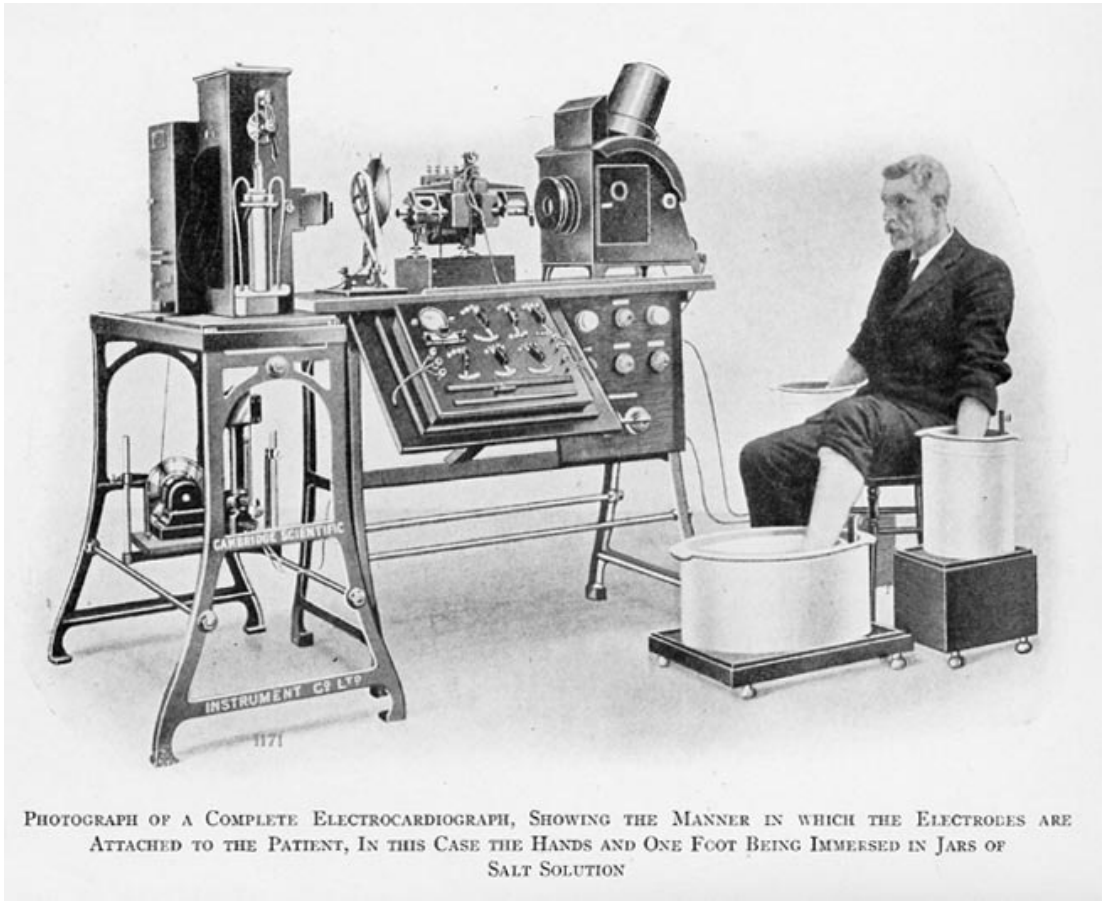
Комбинаторика и алгоритмы для школьников



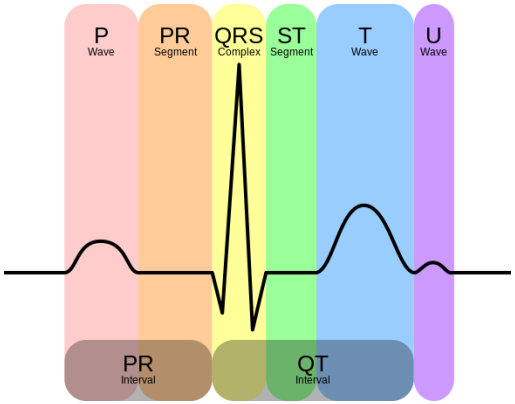
• Летняя школа — 2014 •
21 августа 2014

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

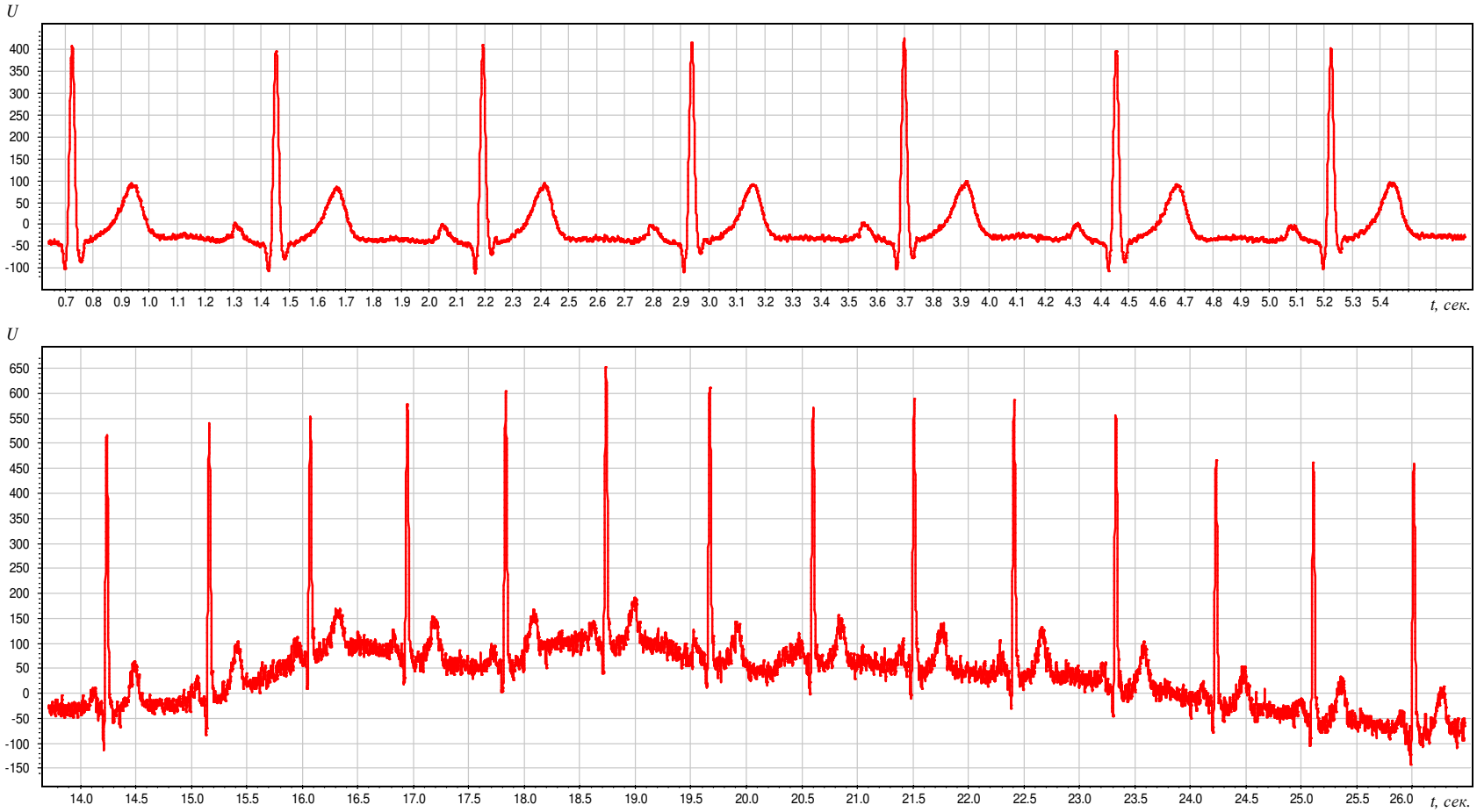


1872 — первые записи электрической активности сердца

1911 — коммерческий электрокардиограф (фото)

1924 — нобелевская премия по медицине, Виллем Эйнтховен

Примеры электрокардиограмм



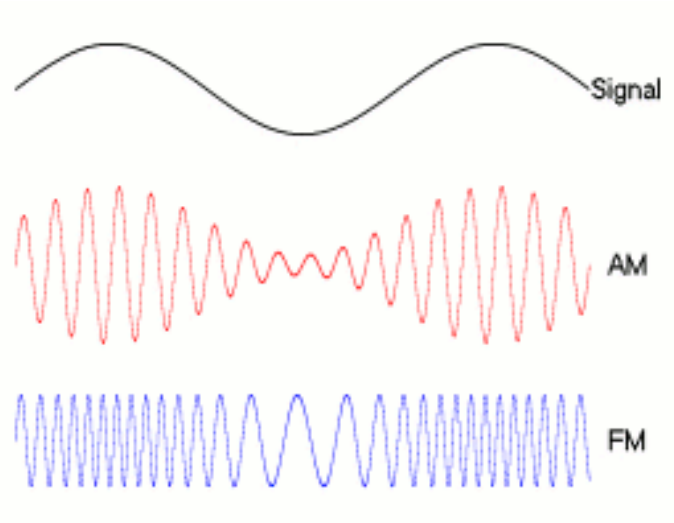
В основе диагностики заболеваний сердца — многочисленные наблюдения за особенностями PQRST-комплекса

Идеи, догадки, гипотезы...

- ЭКГ-сигнал может нести информации о функционировании не только сердца, но и всех систем организма.
- В Китайской Традиционной Медицине давно успешно применяется *пульсовая диагностика*.
- Информация о заболевании должна появляться раньше явных клинических проявлений (когда уже поздно).
  — отсюда заманчивая возможность *ранней* диагностики.
- Каждое заболевание может по-своему «модулировать» ЭКГ-сигнал.
- *Модуляция сигналов* бывает амплитудная и частотная (в радиотехнике — см. следующий слайд),
  их аналоги, наверное, есть в ЭКГ-сигнале.
- **Наша цель** — найти их.
  Хорошая новость — выборка данных уже собрана!

Понятия модуляции сигналов

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



— низкочастотный сигнал

— амплитудная модуляция

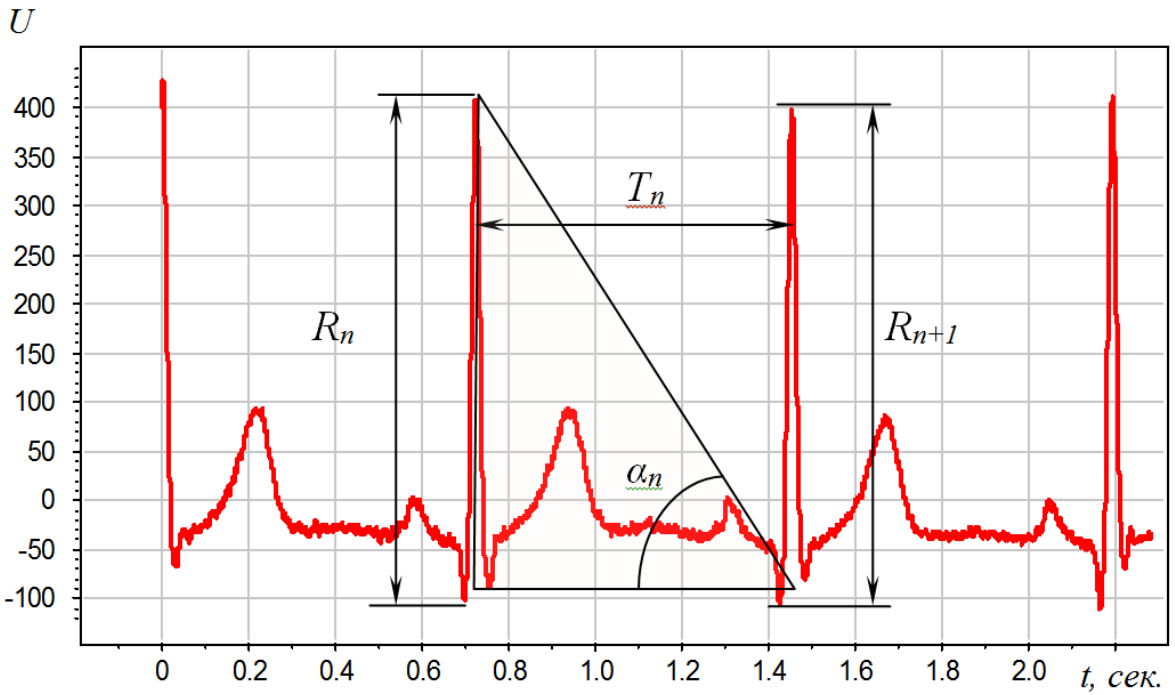
— частотная модуляция

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

Что будет аналогом демодуляции в случае ЭКГ?

Информационный анализ электрокардиосигналов

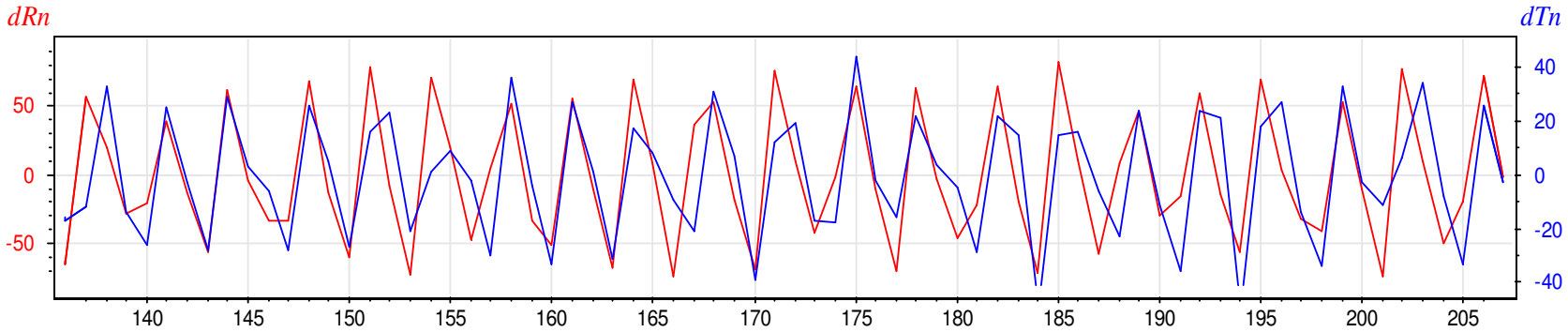
Открытие проф. Вячеслава Максимилиановича Успенского:
для диагностики болезней важны знаки приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.



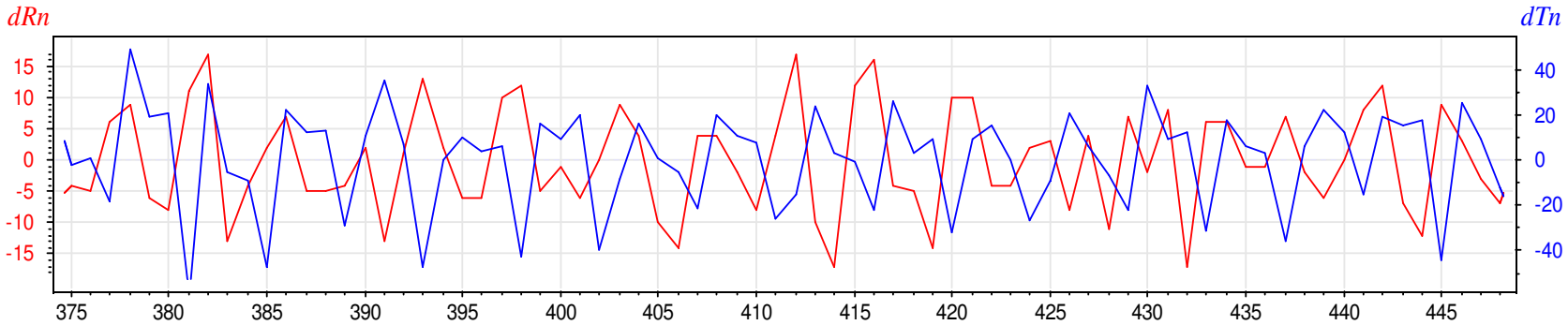
$$\alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$$

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



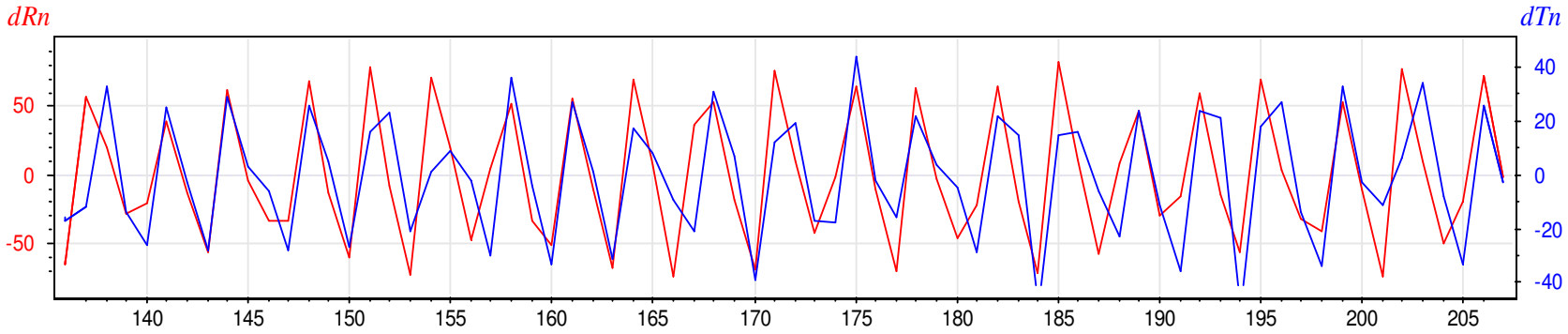
Больной (язвенная болезнь)



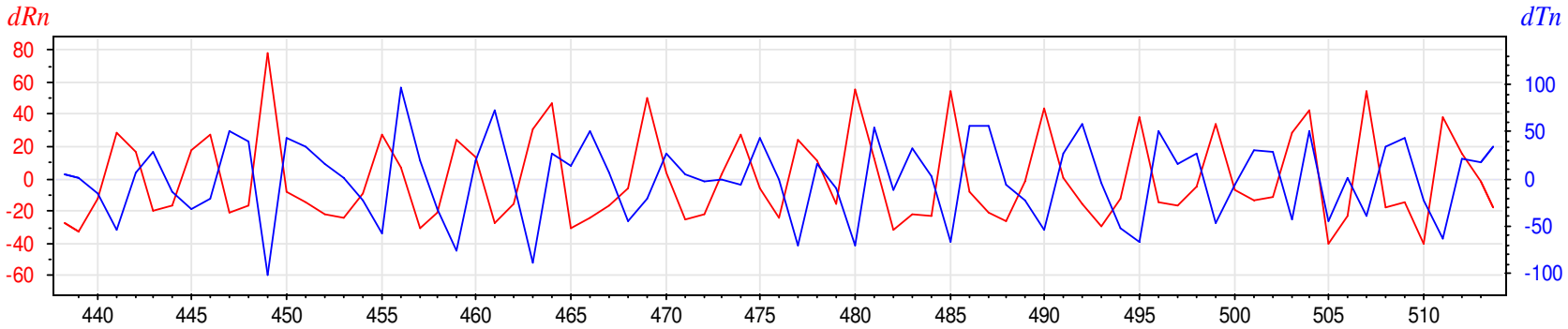
$dRn = R_{n+1} - R_n$, $dTn = T_{n+1} - T_n$ от номера кардиоцикла

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



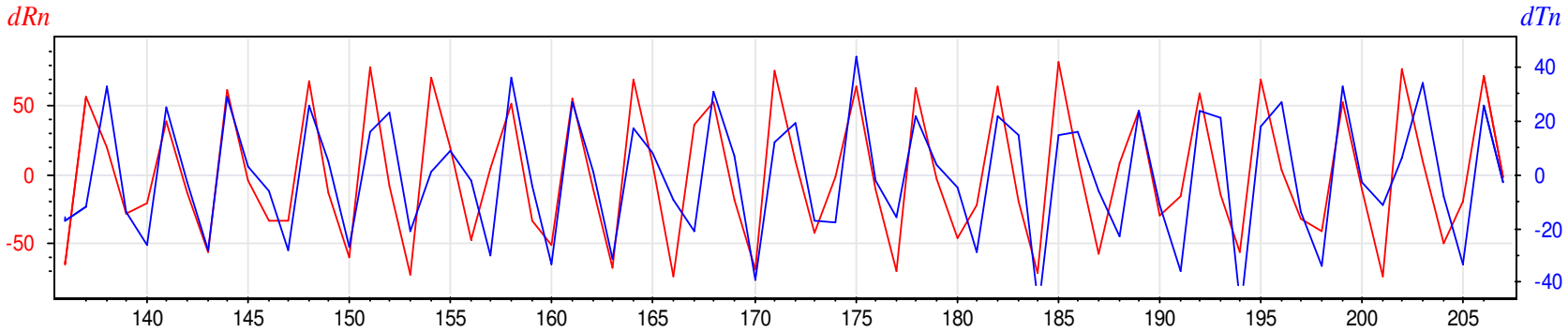
Больной (гипертония)



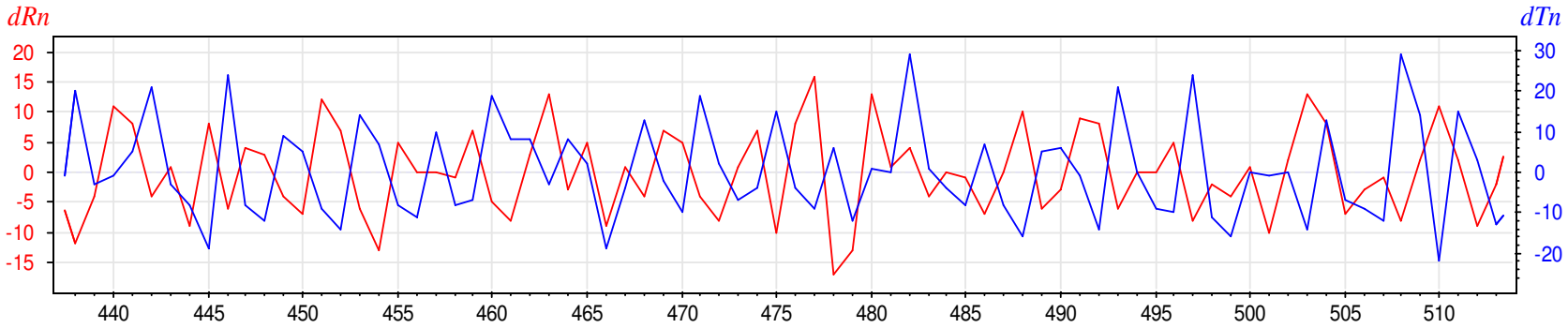
$dRn = R_{n+1} - R_n$, $dTn = T_{n+1} - T_n$ от номера кардиоцикла

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



Больной (рак)



$dRn = R_{n+1} - R_n$, $dTn = T_{n+1} - T_n$ от номера кардиоцикла

Диагностическая система «Скринфакс» (2-е поколение)

- более 10 лет эксплуатации (начало исследований: 1978)
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 40 заболеваний
- из них более 20 имеют отобранные эталонные выборки

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 выборок здоровых и больных
2. Обучение алгоритма классификации
3. Оценивание качества диагностики

Дискретизация и векторизация ЭКГ-сигнала

Дискретизация ЭКГ-сигнала:

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^{N}$;
Выход: *кодограмма* $S = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

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

Векторизация кодограммы ЭКГ-сигнала:

Вход: кодограмма S;
Выход: вектор частот $n = 6^3 = 216$ триграмм $x = (x^1, \ldots, x^n)$, x^j — сколько раз j-я триграмма встретилась в S

Векторизация кодограммы ЭКГ-сигнала

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

Задача статистического (машинного) обучения с учителем

Задача восстановления зависимости $y = f(x)$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: векторы $x_i = (x_i^1, \ldots, x_i^n)$ — объекты обучающей выборки,
$y_i = f(x_i)$ — ответы (диагнозы), $i = 1, \ldots, \ell$:

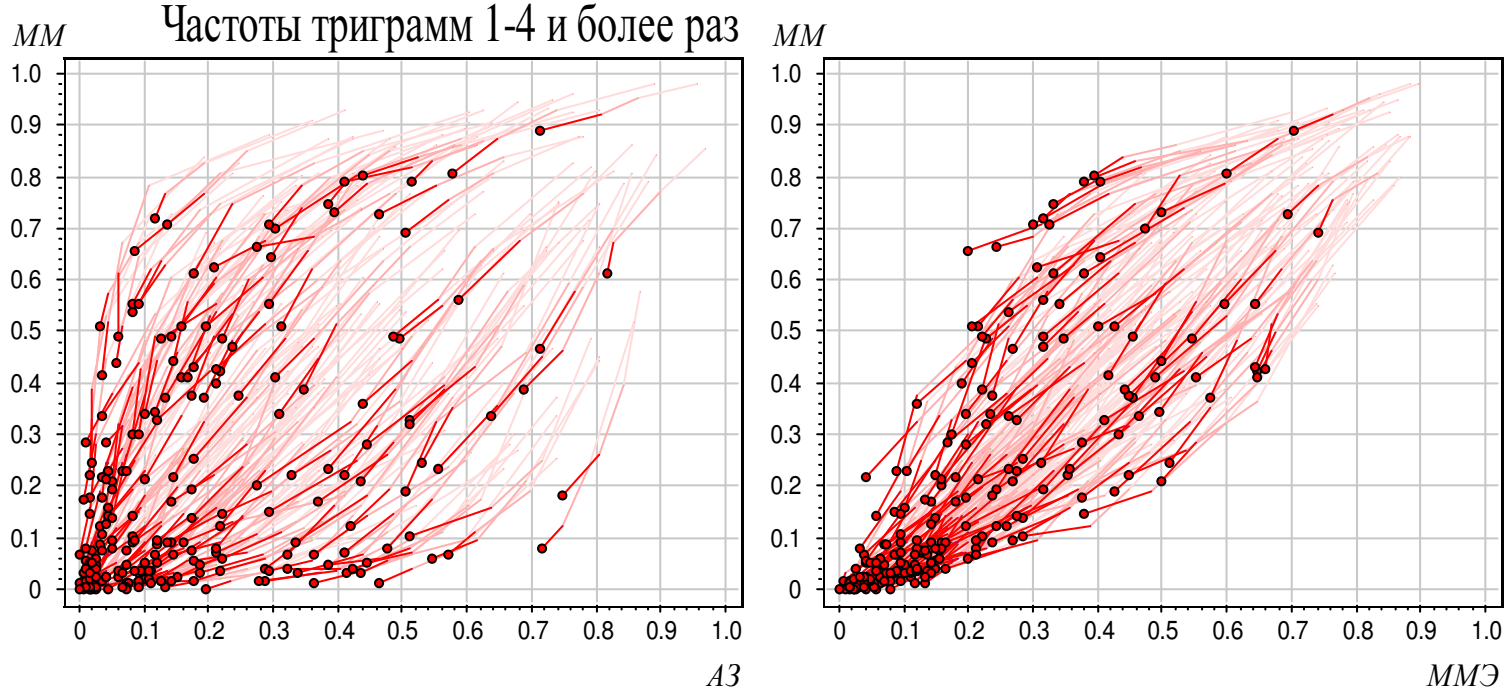
$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: классификатор $a(x)$, способный давать правильные
ответы на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

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

Гипотеза существования информативных признаков-триграмм

Гипотеза: для каждой болезни есть своё множество триграмм, часто встречающихся у больных, и редко — у здоровых.

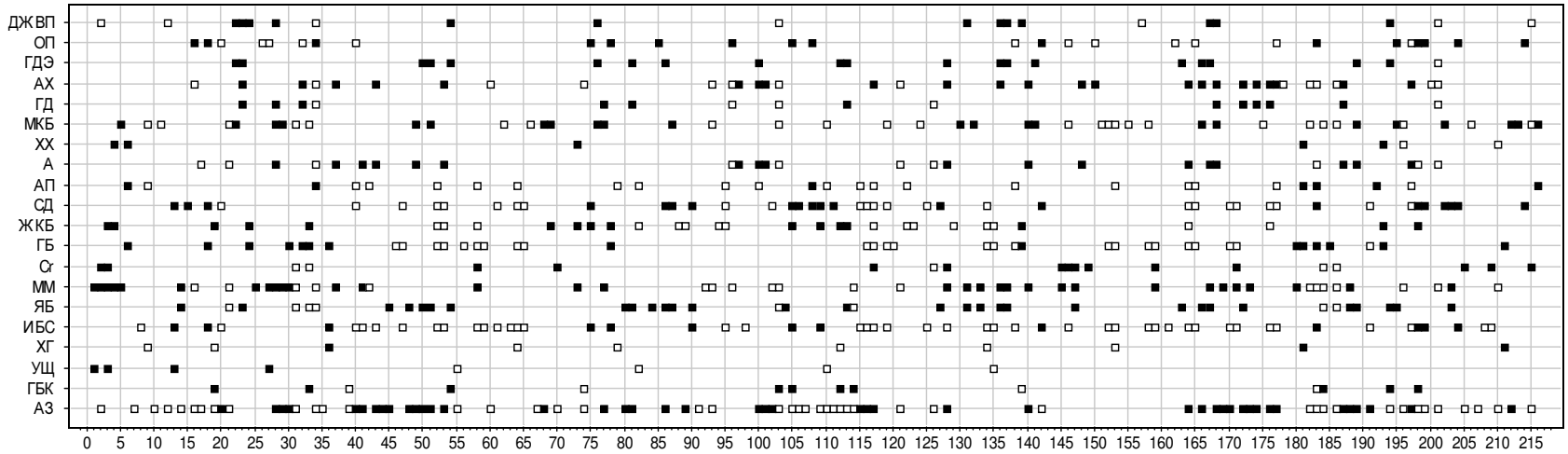


Слева: триграммы в осях «доля здоровых» — «доля больных».
Справа: триграммы в осях «доля больных» — «доля больных».

Гипотеза существования информативных признаков-триграмм

По оси абсцисс — номера триграмм 1..216

По оси ординат — болезни (АЗ — абсолютно здоровые)



□ — аномально низкая частота триграммы

■ — аномально высокая частота триграммы

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

Вывод 2. Болезни хорошо отличаются по наборам триграмм!

Линейная модель классификации

- рассматриваем только одно заболевание,
  $y_i = 1$ — больной, $y_i = 0$ — здоровый
- чем выше частота j-й триграммы x^j, тем она «сильнее»
- есть триграммы, более характерные для больных, и
  есть триграммы, более характерные для здоровых

Линейная модель классификации:

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

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

Бинаризацияя признаков

Гипотеза: 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

Гипотеза: вес j-й триграммы тем больше, чем
1) больше N_{11}^j и N_{00}^j,
2) меньше N_{01}^j и N_{10}^j,

Поэтому 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 $SCORE(x, w)$ набрать кучу неинформативных признаков (не связанных с данным заболеванием), то получится $SCORE(x, w)$ + шум.

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

Идея 2: Отсортировать можно 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:

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

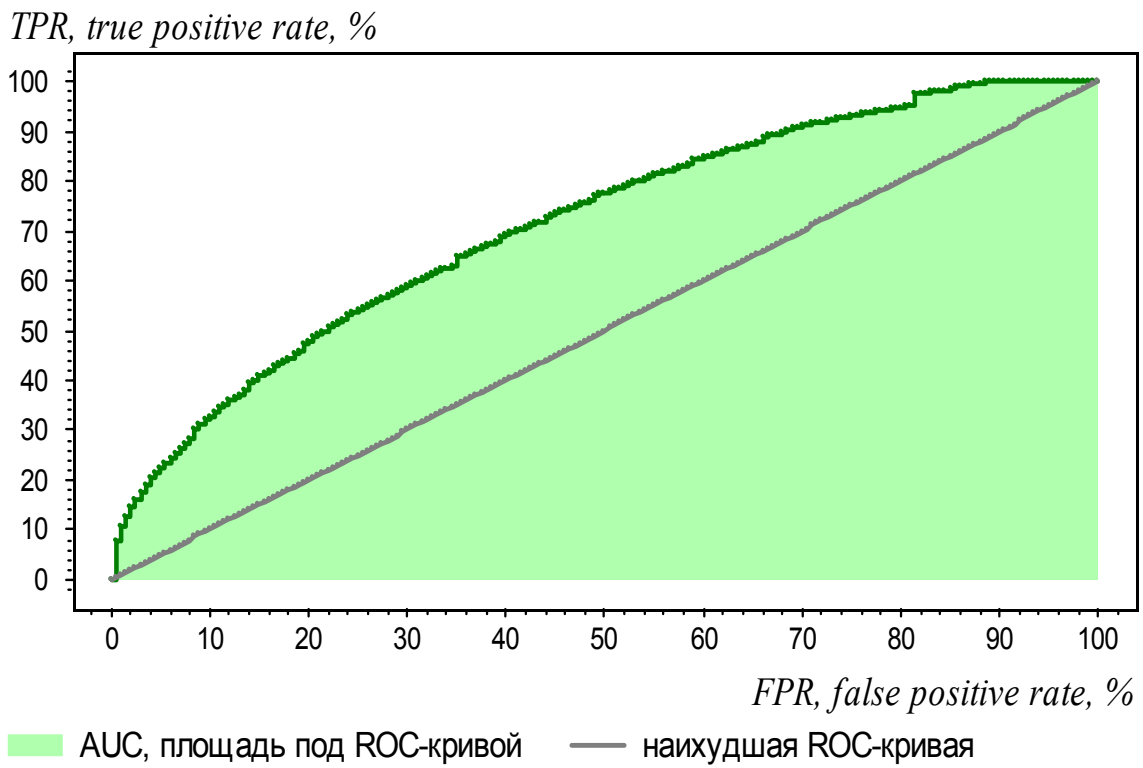
Чувствительность и специфичность хотим максимизировать.

⊕ Они не зависят от соотношения мощностей классов.

⊕ Хорошо подходят для несбалансированных выборок.

Определение ROC-кривой

Каждая точка ROC-кривой соответствует значению порога w_0
(ROC — «receiver operating characteristic»),
по оси X: 1 – специфичность = FPR, false positive rate,
по оси Y: чувствительность = TPR, true positive rate



AUC — площадь под ROC-кривой

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

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

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

Преимущества AUC:

⊕ не зависит от порога w_0, оценивает только качество w;

⊕ не зависит от численности классов;

⊕ это общепринятая мера качества классификации;

Алгоритм построения ROC-кривой и вычисления AUC за $O(\ell)$

Вход: выборка $(x_i, y_i)_{i=1}^{\ell}$; функция $SCORE(x, w)$;

Выход: $\{(\mathrm{FPR}_i, \mathrm{TPR}_i)\}_{i=0}^{\ell}$, AUC — площадь под ROC-кривой.

$\ell_0 := \sum_{i=1}^{\ell} [y_i = 0]$; $\ell_1 := \sum_{i=1}^{\ell} [y_i = 1]$;
упорядочить выборку по убыванию $SCORE(x_i, w)$;
$(\mathrm{FPR}_0, \mathrm{TPR}_0) := (0, 0)$; AUC $:= 0$;
для $i := 1, \ldots, \ell$
  если $y_i = 0$ **то**
    сместиться на один шаг вправо:
    $\mathrm{FPR}_i := \mathrm{FPR}_{i-1} + \frac{1}{\ell_0}$; $\mathrm{TPR}_i := \mathrm{TPR}_{i-1}$;
    AUC $:=$ AUC $+ \frac{1}{\ell_0}\mathrm{TPR}_i$;
  иначе
    сместиться на один шаг вверх:
    $\mathrm{TPR}_i := \mathrm{TPR}_{i-1} + \frac{1}{\ell_1}$; $\mathrm{FPR}_i := \mathrm{FPR}_{i-1}$;

Задача

Дано:
матрица «объекты–признаки» по одной болезни
(некроз ГБК — 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: площадь под ROC-кривой.

Описание задачи — на странице
http://www.MachineLearning.ru/wiki/index.php?title=User:Vokov
http://www.machinelearning.ru/wiki/images/e/e1/School-VI-2014-task-3.rar

Подсказки

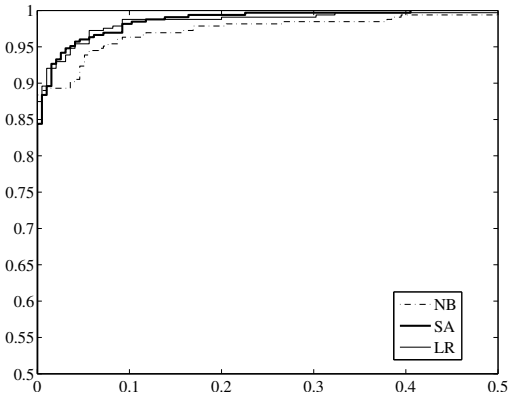
Задача хорошо решается многими методами...

- метод ближайшего соседа (kNN)
- метод опорных векторов (SVM)
- логистическая регрессия (LR)
- нейронная сеть BackProp (ANN)
- наивный байесовский классификатор (NB)
- деревья решений (DT)
- градиентный бустинг (GB) над деревьями решений

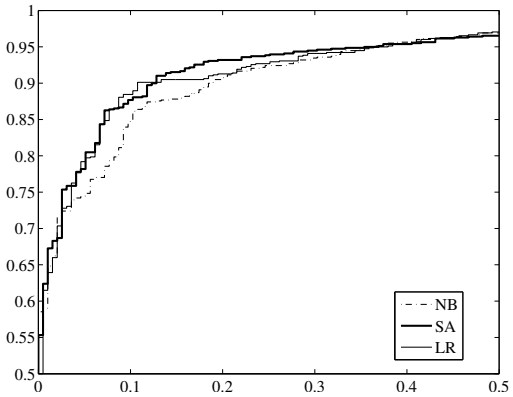
...если приложить усилия и попробовать

- разные одномерные преобразования признаков
- разные эвристики для отбора признаков
- разные оптимизации параметров

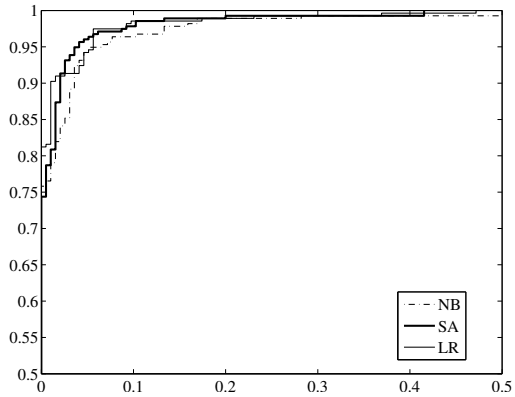
ROC-кривые для некоторых заболеваний



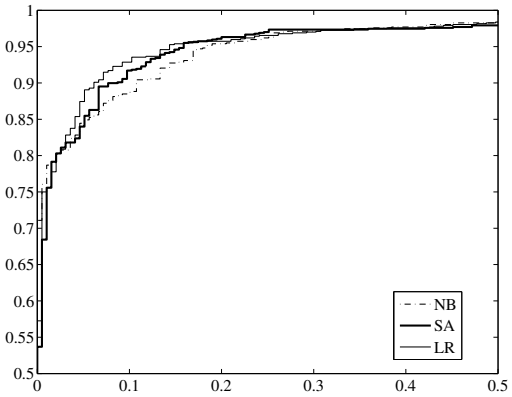
некроз ГБК



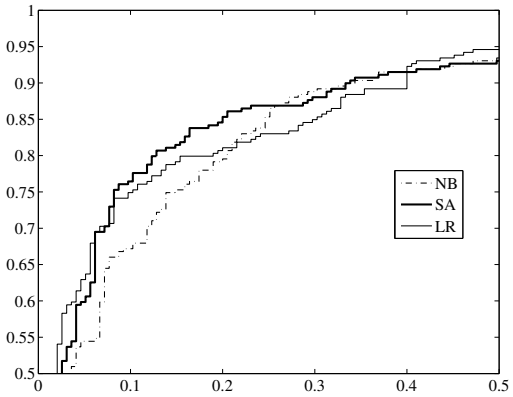
язвенная болезнь



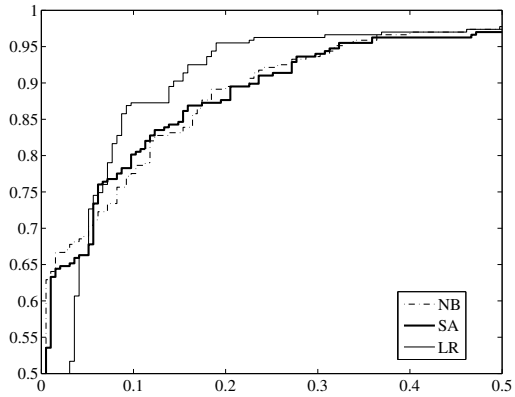
желчекаменная болезнь



сахарный диабет

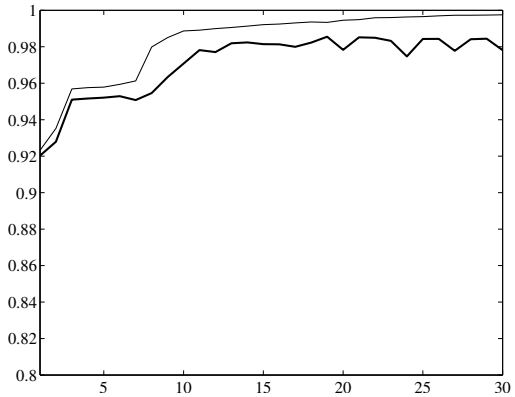


анемия железодефицитная

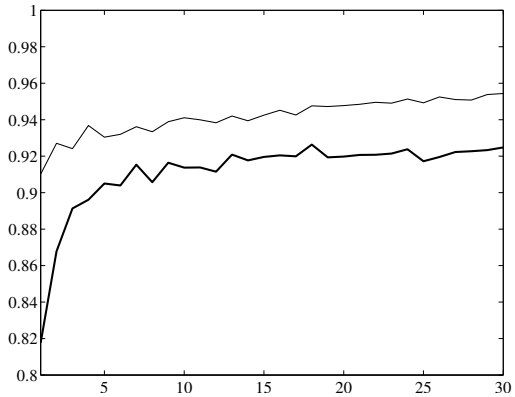


рак

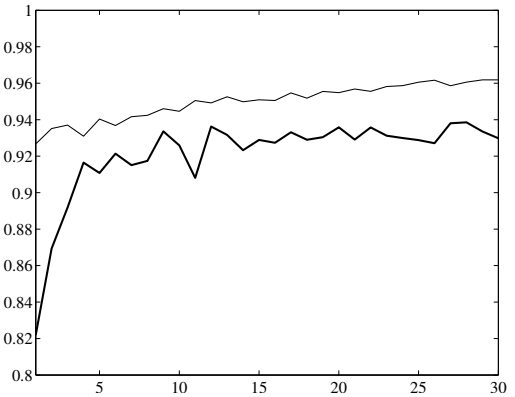
Зависимости AUC от числа используемых признаков K



некроз ГБК



хронический гастрит



зоб щитовидный железы

Тонкая (верхняя) линия — на обучающей выборке

Толстая (нижняя) линия — на тестовой выборке

Выводы:

- Переобучение есть всегда
- Обычно хватает 10–20 признаков
- Точность диагностики выше 90% поразительна!

Резюме

Про задачу диагностики многих болезней по одной ЭКГ:

- Точность диагностики превосходит многие методы!
- Живое исследование на стыке медицины и математики,
- на грани науки и религии
- Так почему такая штука не стоит в каждой поликлинике?

Про машинное обучение:

- Не более чем проведение функции через заданные точки
- Но решает сложные задачи прогнозирования и принятия решений, которые человек решить не в состоянии!
- 2-я лекция — про разные методы машинного обучения
- 3-я лекция — про комбинаторные оценки переобучения

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

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