

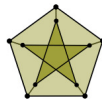
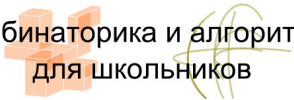
Теория и практика машинного обучения. Задача диагностики заболеваний по электрокардиограмме

Воронцов Константин Вячеславович

МФТИ • МГУ • ВШЭ • ВЦ РАН • Яндекс • FORECSYS



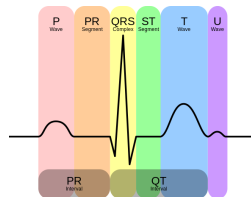
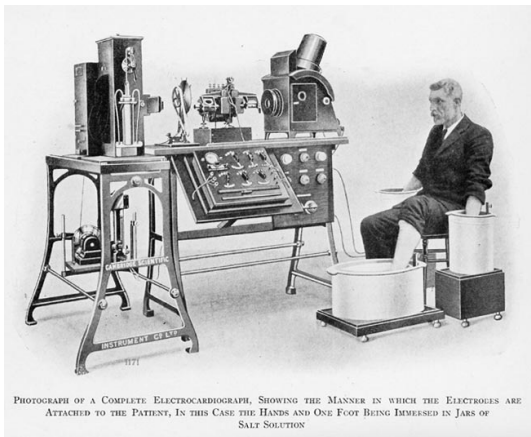
Комбинаторика и алгоритмы
для школьников



- Летняя школа — 2014 •
21 августа 2014

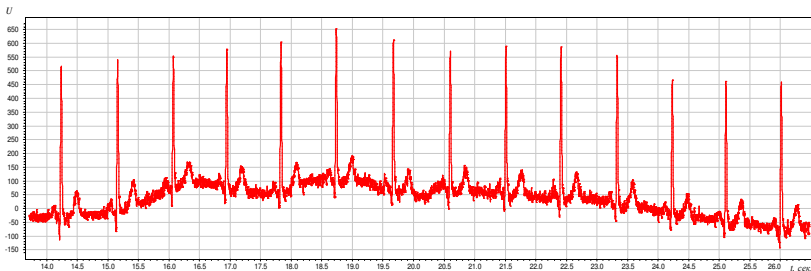
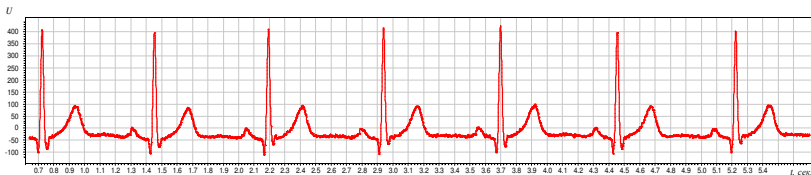
- 1 Анализ ЭКГ-сигналов и диагностика заболеваний**
 - Электрокардиография и электрокардиограммы
 - Метод В.М.Успенского
 - Задача диагностики как задача машинного обучения
- 2 Наш первый обучаемый алгоритм**
 - Линейная модель классификации
 - Отбор признаков
 - Оценивание качества классификации
- 3 Конкурсное задание**
 - Постановка задачи
 - Условия. Сроки. Подсказки.
 - Некоторые наши результаты

Электрокардиография



- 1872 — первые записи электрической активности сердца
- 1911 — коммерческий электрокардиограф (фото)
- 1924 — нобелевская премия по медицине, Виллем Эйнтховен

Примеры электрокардиограмм



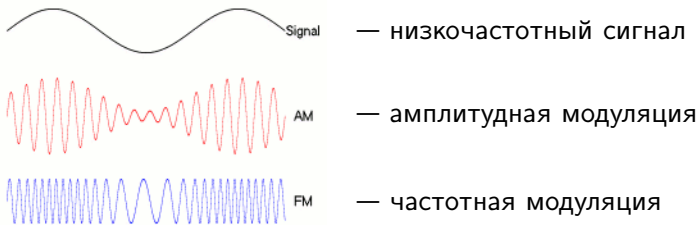
В основе диагностики заболеваний сердца — многочисленные наблюдения за особенностями PQRST-комплекса

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.

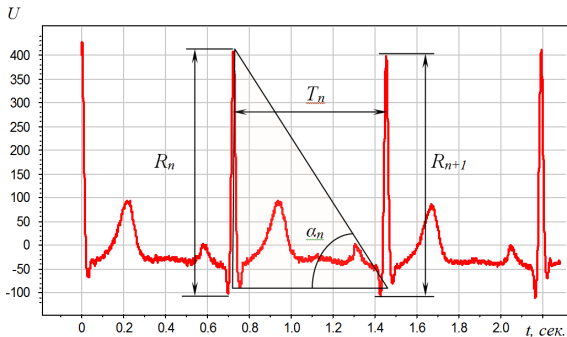


Демодуляция — процесс, обратный модуляции, преобразование модулированных колебаний высокой (несущей) частоты в исходный низкочастотный сигнал.

Что будет аналогом демодуляции в случае ЭКГ?

Информационный анализ электрокардиосигналов

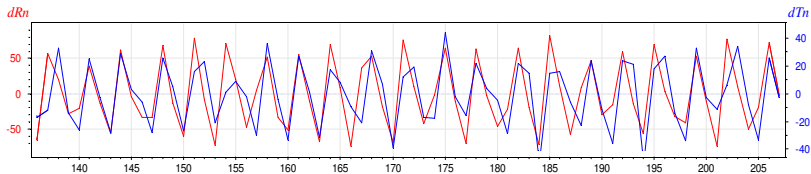
Открытие проф. Вячеслава Максимилиановича Успенского:
для диагностики болезней важны знаки приращений
амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.



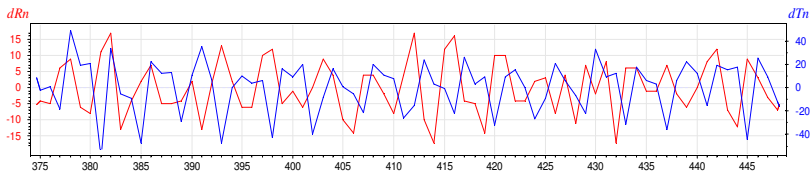
$$\alpha_n = \arctg \frac{R_n}{T_n}$$

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



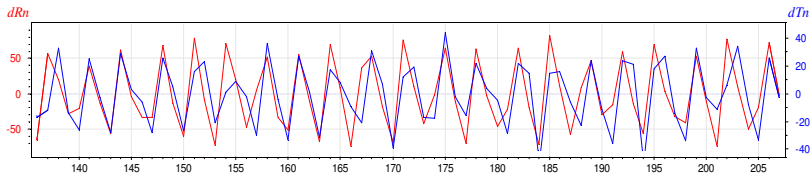
Больной (язвенная болезнь)



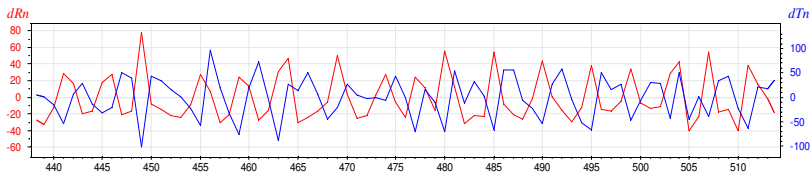
$$dRn = R_{n+1} - R_n, \quad dTn = T_{n+1} - T_n \quad \text{от номера кардиоцикла}$$

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



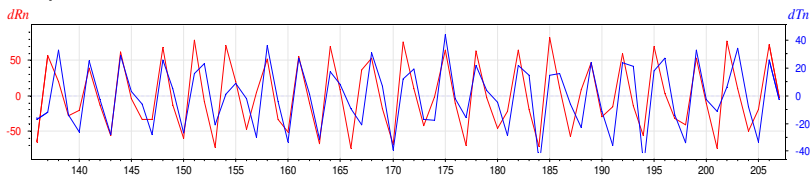
Больной (гипертония)



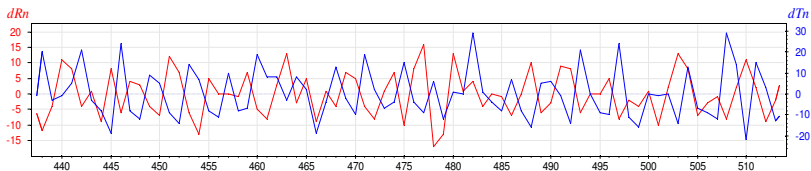
$$dR_n = R_{n+1} - R_n, \quad dT_n = T_{n+1} - T_n \quad \text{от номера кардиоцикла}$$

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



Больной (рак)



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

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 $\mathcal{A} = \{A, B, C, D, E, F\}$

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

Векторизация кодограммы ЭКГ-сигнала:

Вход: кодограмма S ;

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

Задача статистического (машинного) обучения с учителем

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

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

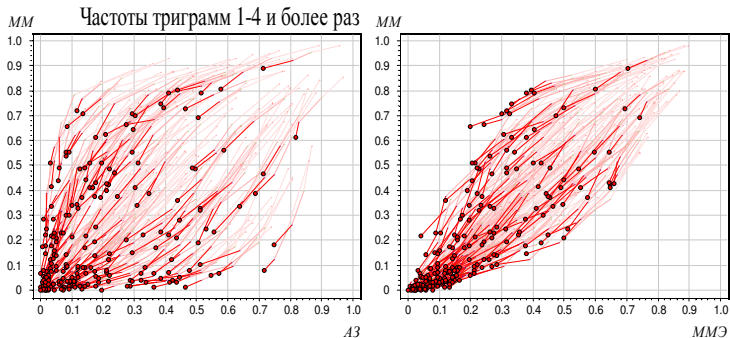
$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

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

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

Гипотеза существования информативных признаков-триграмм

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

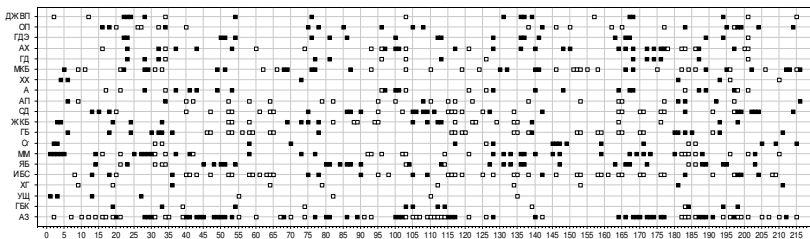


Слева: триграммы в осях «доля здоровых» — «доля больных».
Справа: триграммы в осях «доля больных» — «доля больных».

Гипотеза существования информативных признаков-триграмм

По оси абсцисс — номера триграмм 1..216

По оси ординат — болезни (АЗ — абсолютно здоровые)



□ — anomalously low frequency of trigram

■ — anomalously high frequency of trigram

Вывод 1. Для каждой болезни есть триграммы с anomalously high and anomalously low frequency of occurrence

Вывод 2. Diseases are well distinguished by sets of trigrams!

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:

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

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

Бинаризация 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:

$b_i^j = [x_i^j \geq A]$ (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:

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

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

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

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

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

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

... и разрешается фантазировать!

Отбор признаков

Гипотеза: если в линейный классификатор $SCORE(x, w)$ набрать кучу неинформативных признаков (не связанных с данным заболеванием), то получится $SCORE(x, w) + \text{шум}$.

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 $(x_i, y_i)_{i=1}^{\ell}$; $\mathcal{A} = \{A_1, \dots, A_N\}$, $\mathcal{K} = \{K_1, \dots, K_M\}$

Выход: веса w_j , параметры A, K

- 1 **для** всех значений параметров $A \in \mathcal{A}, K \in \mathcal{K}$
- 2 вычислить N_{yz}^j для всех $j = 1, \dots, n, y, z \in \{0, 1\}$;
- 3 вычислить веса w_j для всех признаков $j = 1, \dots, n$;
- 4 отсортировать признаки по убыванию $|w_j|$;
- 5 обнулить веса признаков w_{K+1}, \dots, w_n ;
- 6 оценить качество классификации $Q(A, K)$;
- 7 оставить A, K , при которых $Q(A, K) \rightarrow \max$;

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:

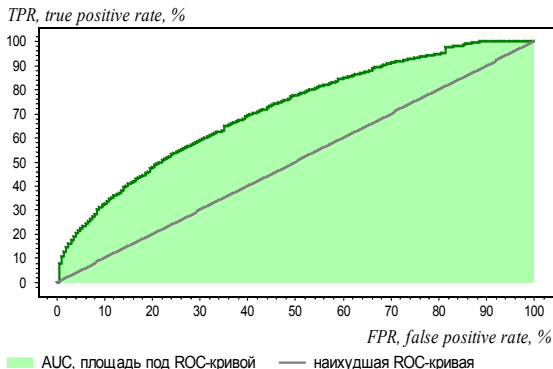
$$\text{специфичность} = \frac{\sum_{i=1}^{\ell} [y_i = 0][a(x_i) = 0]}{\sum_{i=1}^{\ell} [y_i = 0]}$$

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: $1 - \text{специфичность} = \text{FPR}$, false positive rate,
по оси Y: чувствительность = TPR, true positive rate



AUC — площадь под ROC-кривой

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

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

$$AUC = \frac{1}{l_0 l_1} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [y_i < y_j] [SCORE(x_i, w) < SCORE(x_j, w)]$$

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: $\{(FPR_i, TPR_i)\}_{i=0}^{\ell}$, AUC — площадь под ROC-кривой.

- 1 $\ell_0 := \sum_{i=1}^{\ell} [y_i = 0]$; $\ell_1 := \sum_{i=1}^{\ell} [y_i = 1]$;
- 2 упорядочить выборку по убыванию $SCORE(x_i, w)$;
- 3 $(FPR_0, TPR_0) := (0, 0)$; AUC := 0;
- 4 **для** $i := 1, \dots, \ell$
- 5 **если** $y_i = 0$ **то**
- 6 сместиться на один шаг вправо:
7 $FPR_i := FPR_{i-1} + \frac{1}{\ell_0}$; $TPR_i := TPR_{i-1}$;
7 $AUC := AUC + \frac{1}{\ell_0} TPR_i$;
- 8 **иначе**
- 9 сместиться на один шаг вверх:
- 10 $TPR_i := TPR_{i-1} + \frac{1}{\ell_1}$; $FPR_i := FPR_{i-1}$;

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

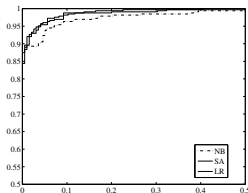
Задача хорошо решается многими методами. . .

- 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

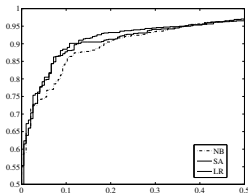
. . . если приложить усилия и попробовать

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

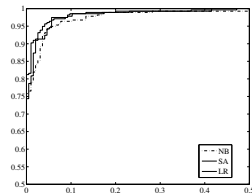
ROC-кривые для некоторых заболеваний



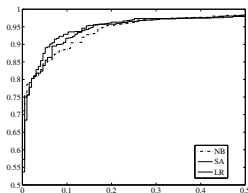
некроз ГБК



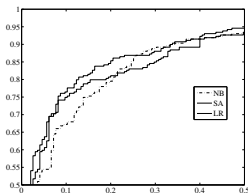
язвенная болезнь



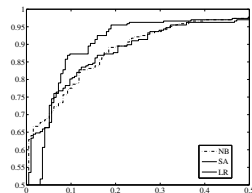
желчекаменная болезнь



сахарный диабет

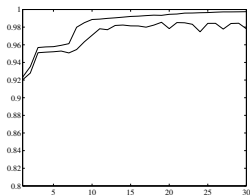


анемия железододефицитная

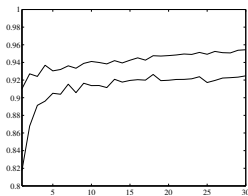


рак

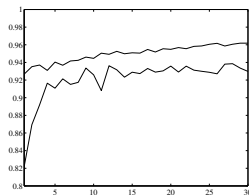
Зависимости AUC от числа используемых признаков K



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 :)