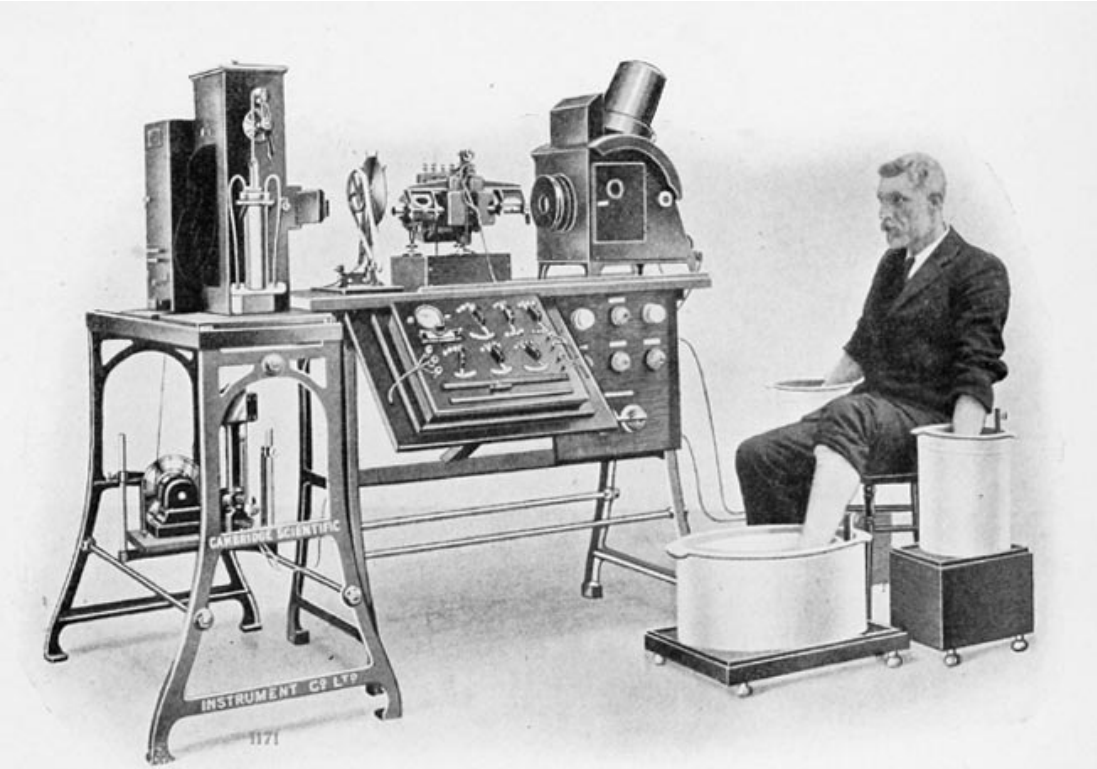
Семинар по машинному обучению
• • •
Задача диагностики многих заболеваний по одной электрокардиограмме

Воронцов Константин Вячеславович

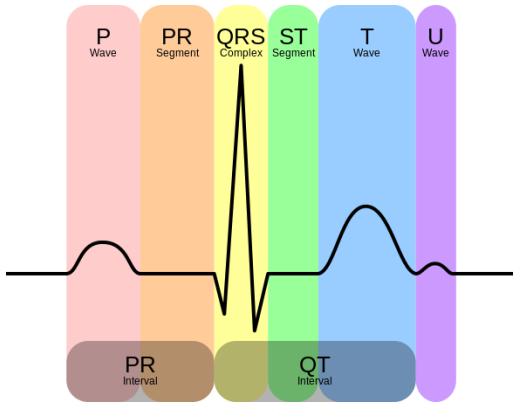
ВМК МГУ • 12 сентября 2014

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

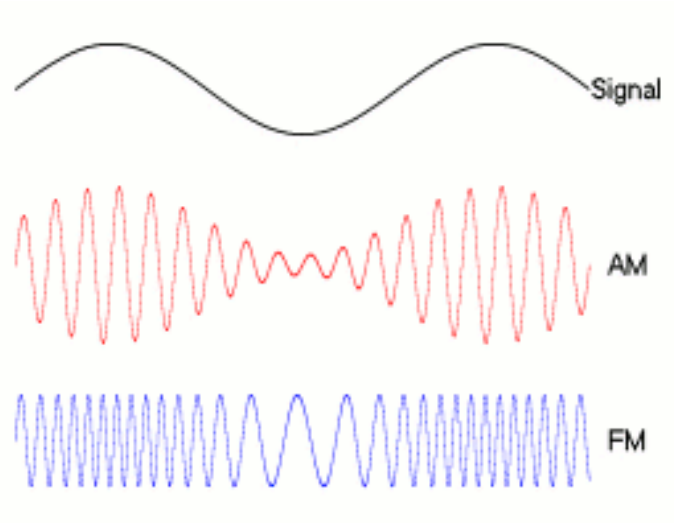
В основе диагностики заболеваний сердца — многочисленные наблюдения за особенностями PQRST-комплекса

Предпосылки метода В.М.Успенского

- ЭКГ-сигнал может нести информации о функционировании не только сердца, но и всех систем организма.
- В Китайской Традиционной Медицине давно успешно применяется *пульсовая диагностика*.
- Если информация о заболевании проявляться на любой его стадии, то возможна *ранняя* диагностика.
- 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

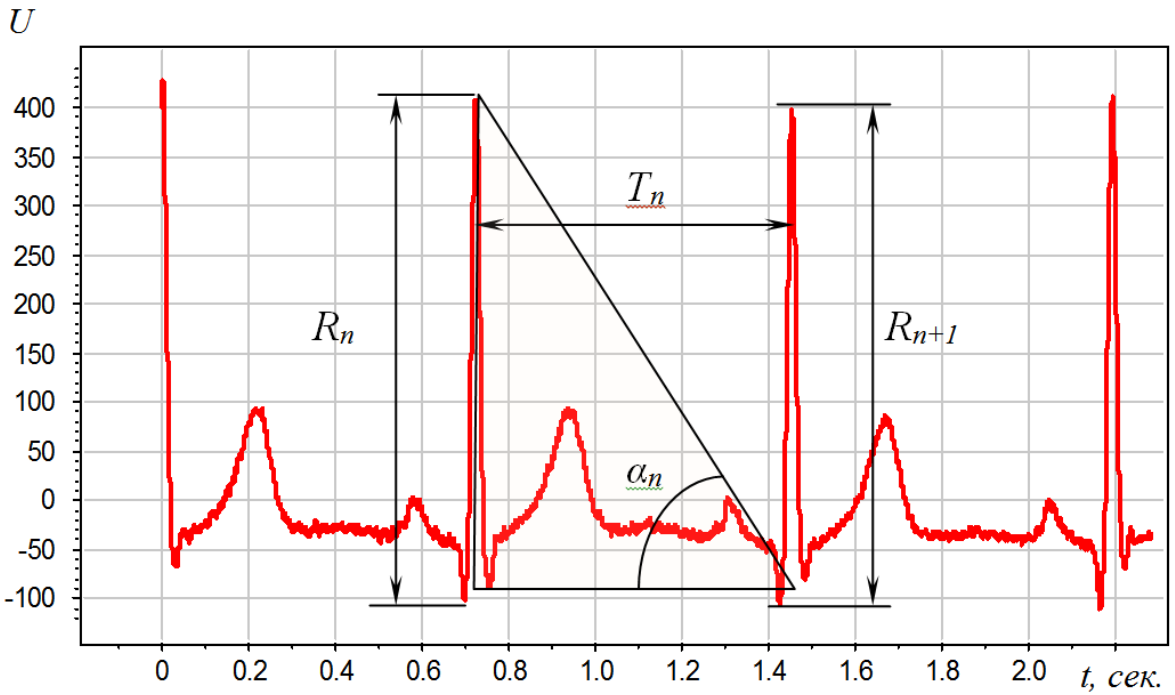
— частотная модуляция

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

Что будет аналогом демодуляции в случае ЭКГ?

Информационный анализ электрокардиосигналов

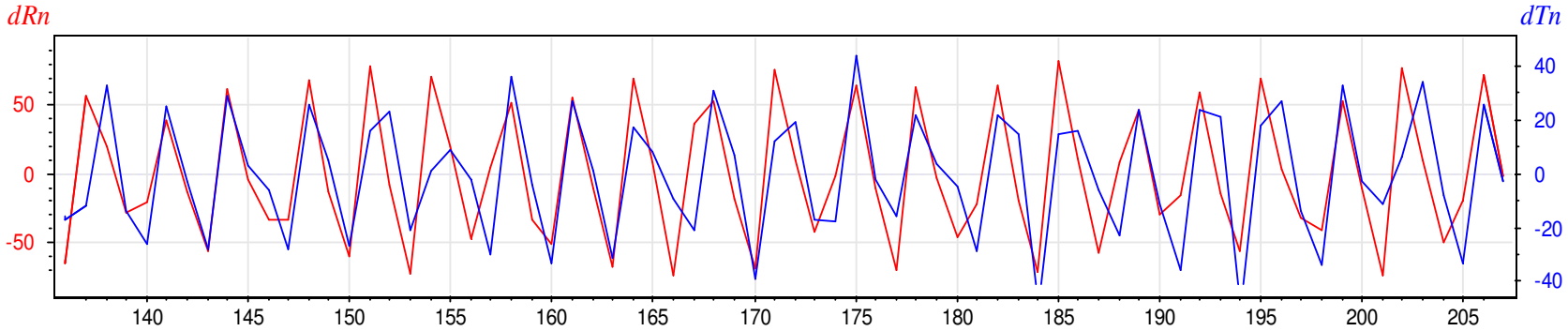
Открытие проф. Вячеслава Максимилиановича Успенского:
для диагностики болезней важны знаки приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.



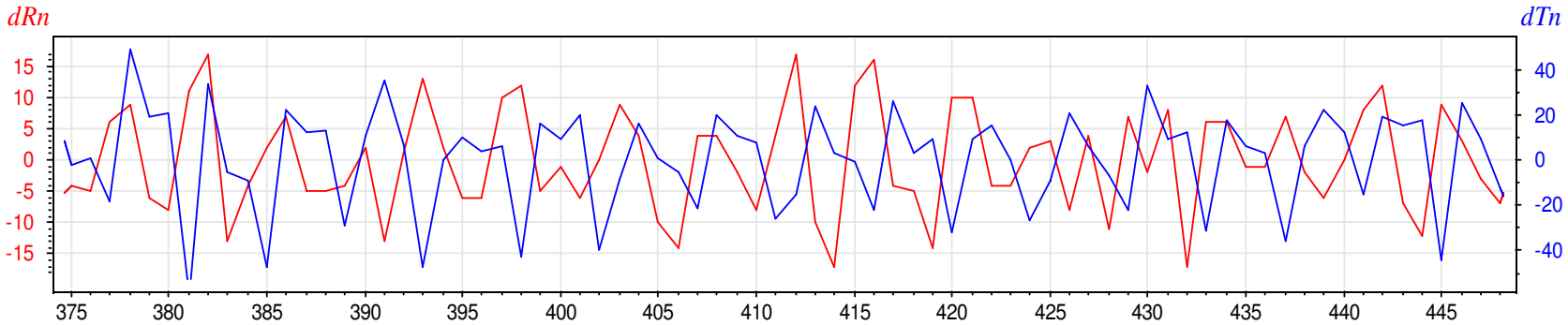
$$\alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$$

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



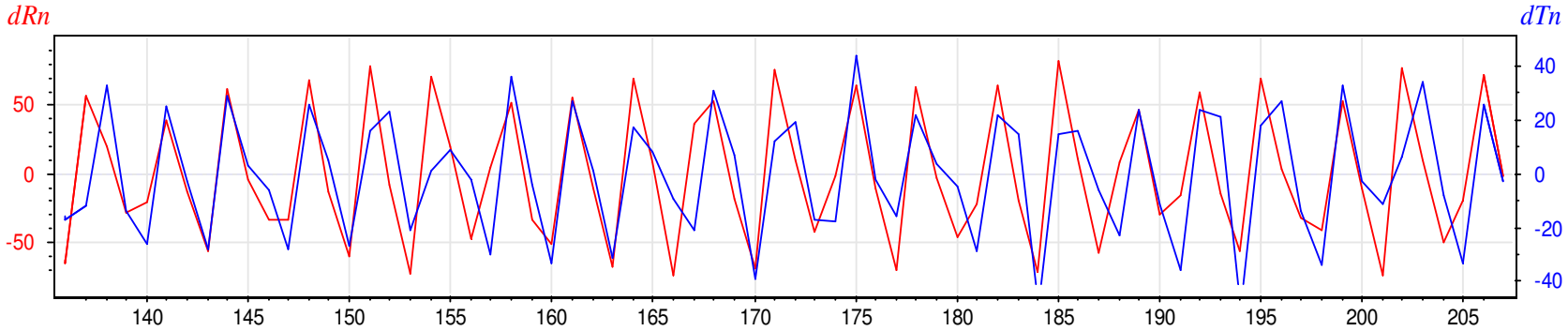
Больной (язвенная болезнь)



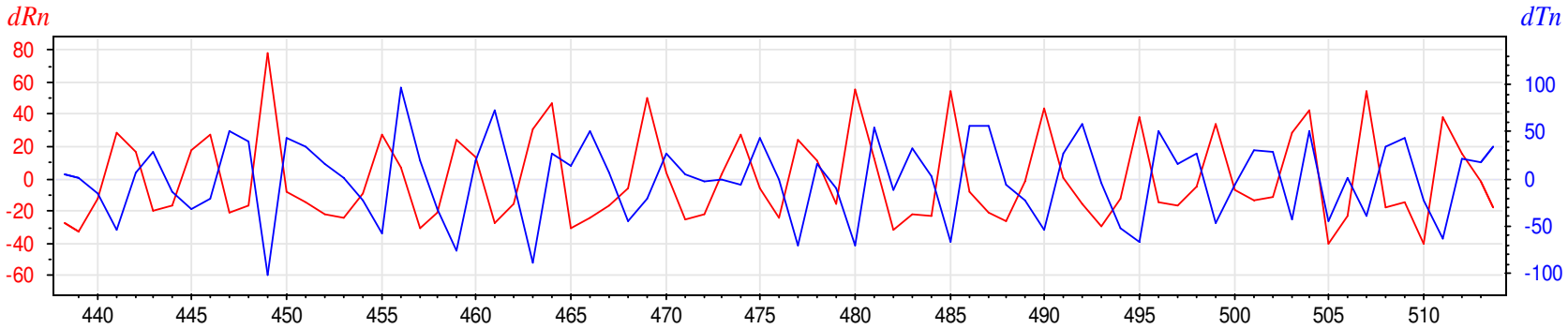
$dRn = R_{n+1} - R_n$, $dTn = T_{n+1} - T_n$ от номера кардиоцикла

Есть ли различия в знаках приращений у больных и здоровых?

Здоровый



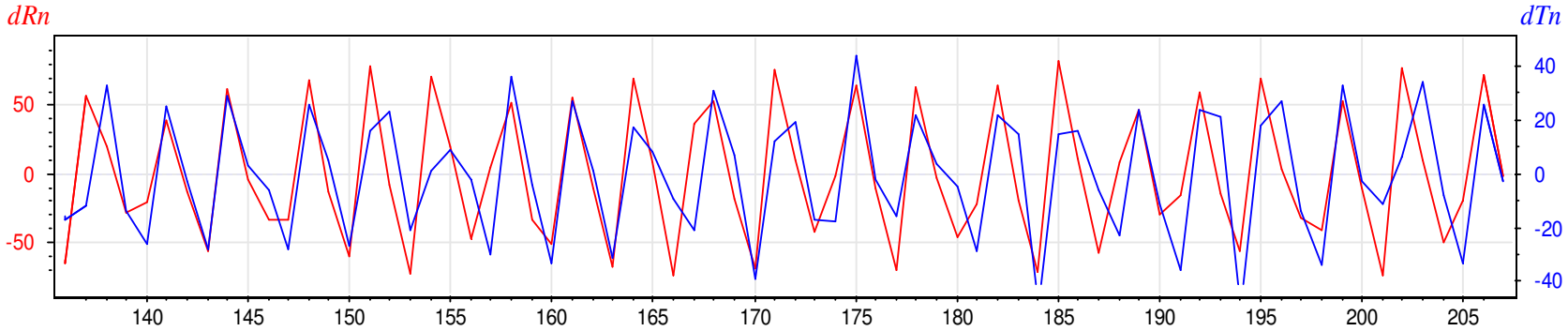
Больной (гипертония)



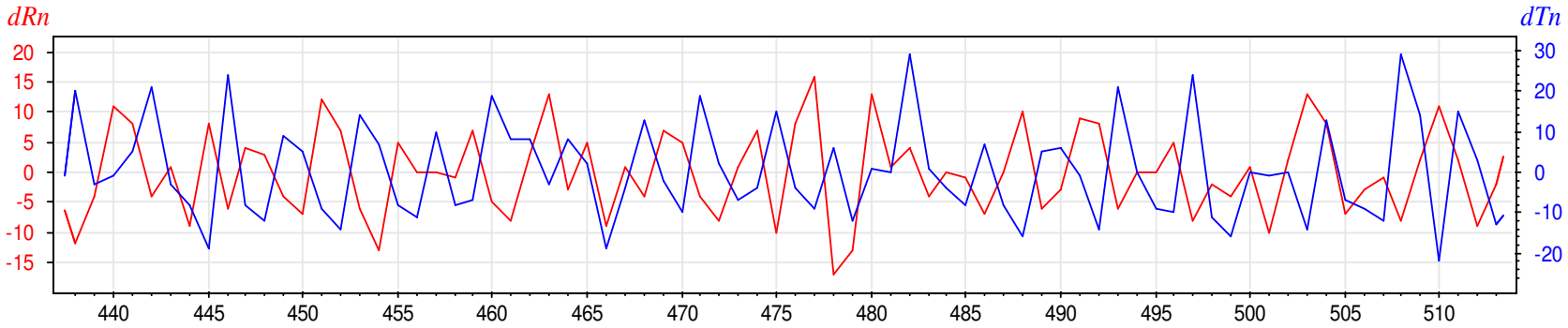
$dRn = R_{n+1} - R_n$, $dTn = T_{n+1} - T_n$ от номера кардиоцикла

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

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. Разработка модели классификации
2. Обучение (оптимизация) 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* $x = (s_n)_{n=1}^{N-1}$ — 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 x;
Выход: вектор частот $n = 6^3 = 216$ триграмм $(f_1(x), \ldots, f_n(x))$,
$f_j(x)$ — сколько раз j-я триграмма встретилась в x

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

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

Методы обучения линейных классификаторов

Линейная модель классификации:

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

Методы обучения весов w_j в линейных классификаторах

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

Простые эвристики для выбора весов

Число объектов класса y, для которых триграмма j частая

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

Число объектов класса y, для которых триграмма j редкая

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

Эвристика: вес триграммы j должен быть тем больше,
чем больше N_1^j и n_0^j,
чем меньше N_0^j и n_1^j.

Простые эвристики для выбора весов

Эвристика: вес триграммы j должен быть тем больше,
чем больше N_1^j и n_0^j,
чем меньше N_0^j и n_1^j.

Поэтому можно пробовать разные формулы для весов:

$$w_j = \frac{N_1^j}{N_0^j} \qquad w_j = \frac{N_1^j n_0^j}{N_0^j n_1^j}$$

$$w_j = \log \frac{N_1^j}{N_0^j} \qquad w_j = \log \frac{N_1^j n_0^j}{N_0^j n_1^j}$$

$$w_j = \sqrt{N_1^j} - \sqrt{N_0^j} \qquad w_j = \sqrt{N_1^j n_0^j} - \sqrt{N_0^j n_1^j}$$

... и разрешается фантазировать!

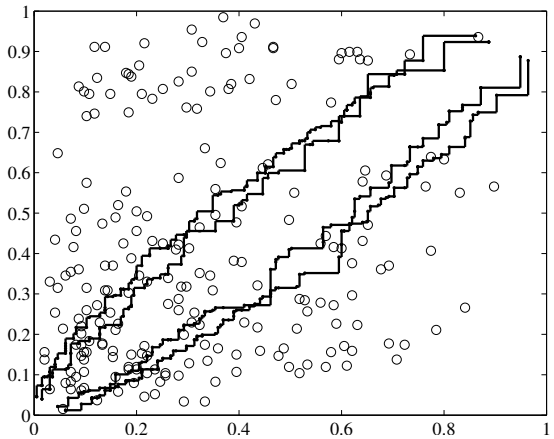
Перестановочный тест для поиска информативных триграмм

Точки на графиках — это триграммы, $j = 1, \ldots, 216$

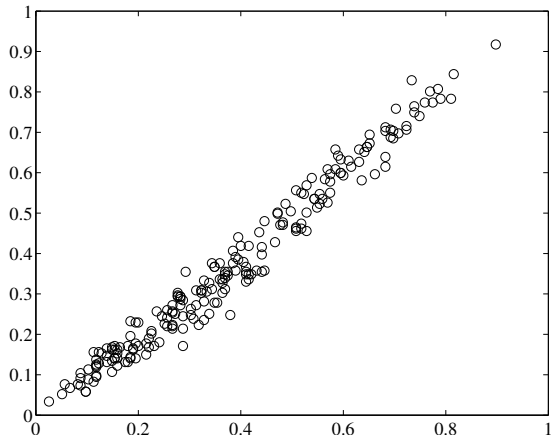
ось X: $\frac{1}{\ell_0} N_0^j$ — доля здоровых с частой триграммой j

ось Y: $\frac{1}{\ell_1} N_1^j$ — доля больных с частой триграммой j

Болезнь: некроз головки бедренной кости



истинные y_i



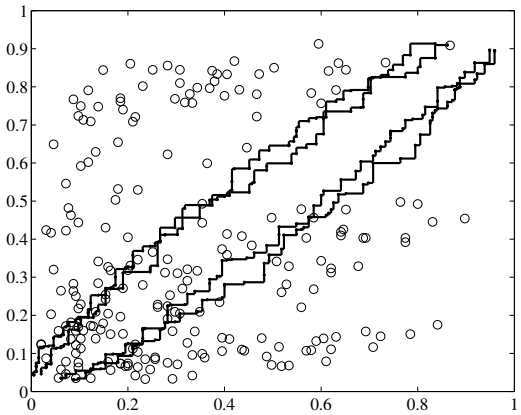
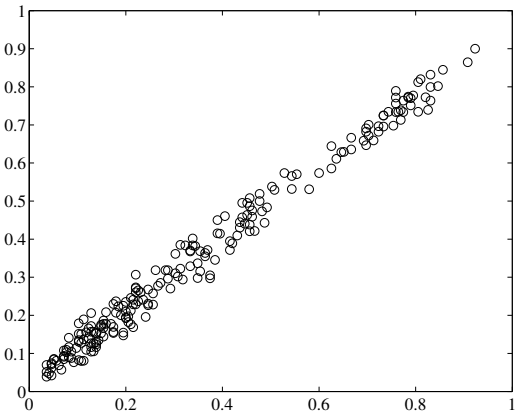
случайно перепутанные y_i

Значимые триграммы — вне 90% (99.8%) доверительной области, при 20 (1000) случайных перемешиваний меток y_i

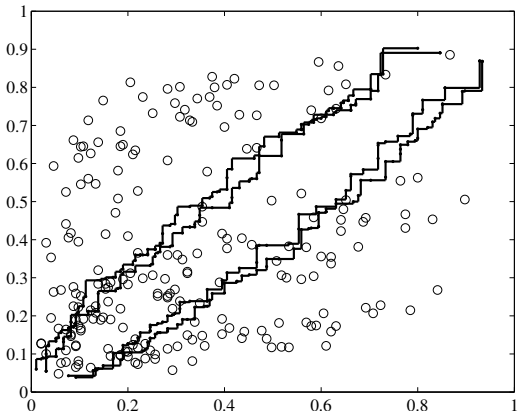
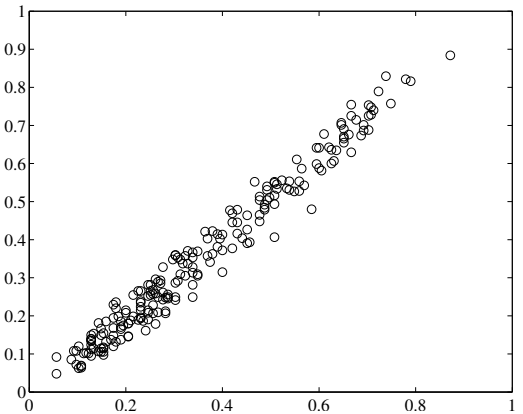
Перестановочный тест для поиска информативных триграмм

Для каждой болезни есть свои неслучайно частые триграммы

Болезнь: ишемия сердца

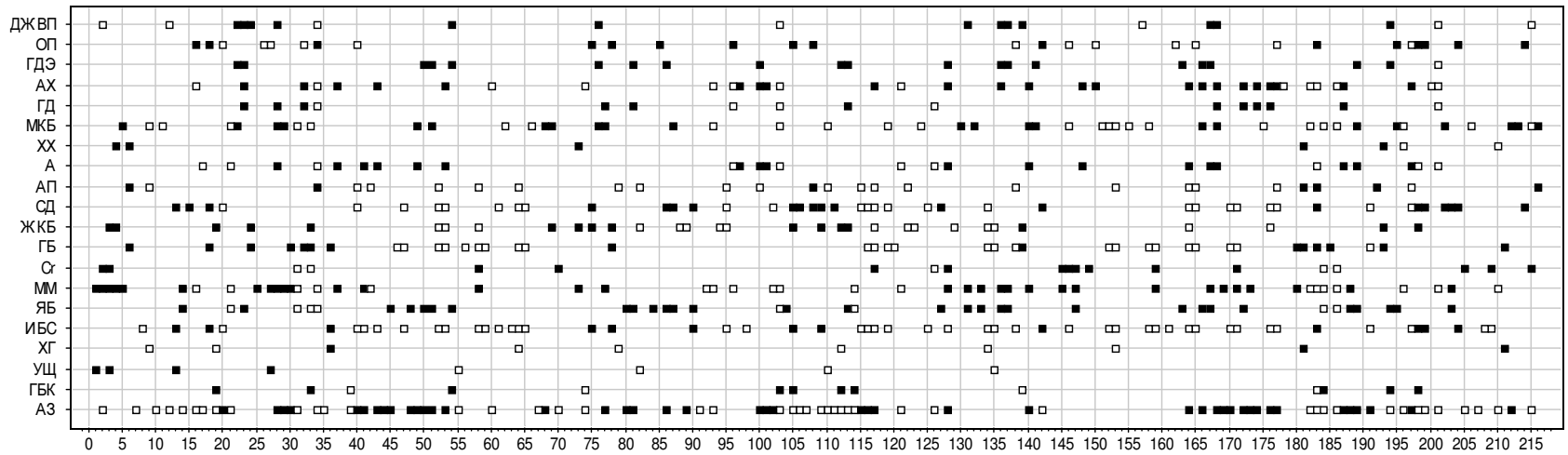
Болезнь: узловой зоб щитовидной железы

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!

Терминология диагностики

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{1}{\ell_1} \sum_{i:\, y_i=1} \big[a(x_i) = 1\big]$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{1}{\ell_0} \sum_{i:\, y_i=0} \big[a(x_i) = 0\big]$$

Area Under Curve — доля правильно упорядоченных пар:

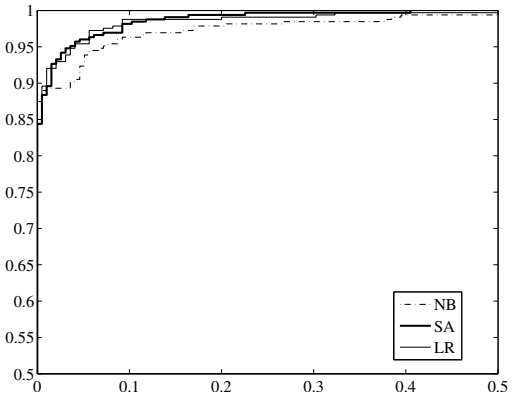
$$\text{AUC} = \frac{1}{\ell_0 \ell_1} \sum_{i:\, y_i=0} \sum_{k:\, y_k=1} \big[\langle x_i, w\rangle < \langle x_k, w\rangle\big]$$

Результаты кросс-валидации

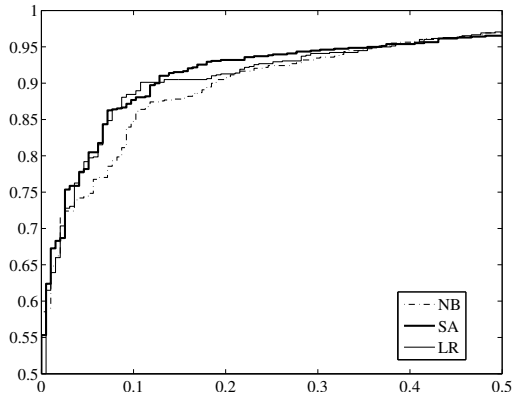
Обучающая выборка — для оптимизации параметров w_j
Тестовая выборка — для оценивания чувс., спец., AUC
40×10-fold cross-validation — для доверительного оценивания

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 щитов. 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

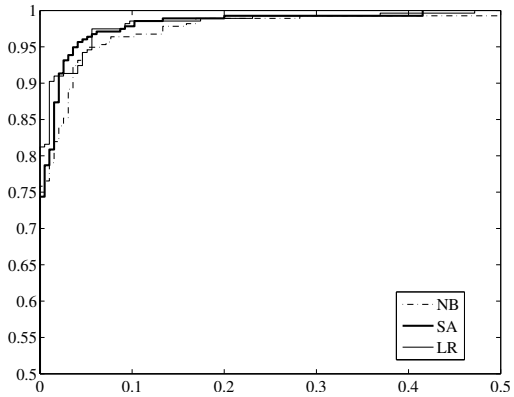
ROC-кривые в осях (1−специфичность) — чувствительность



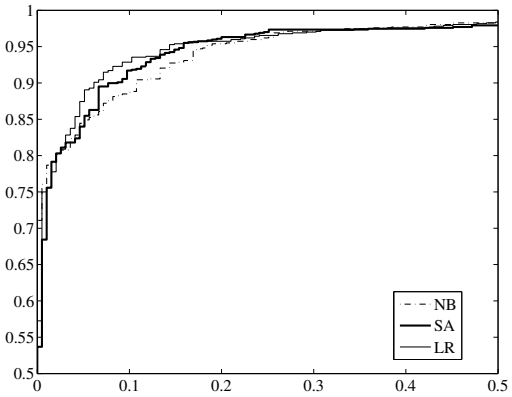
некроз ГБК



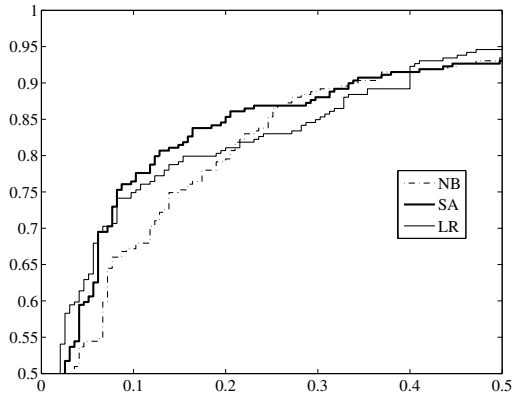
язвенная болезнь



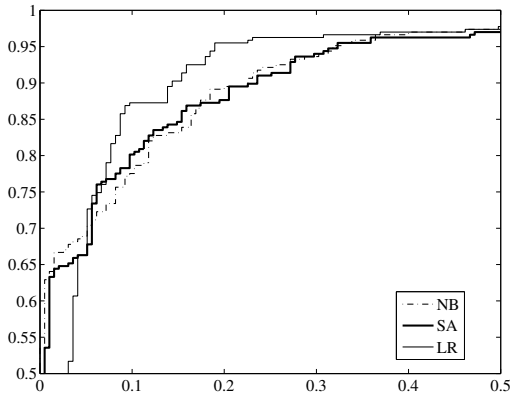
желчнокаменная болезнь



диабет



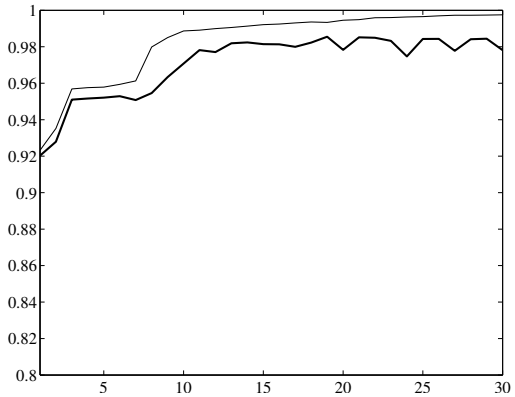
анемия



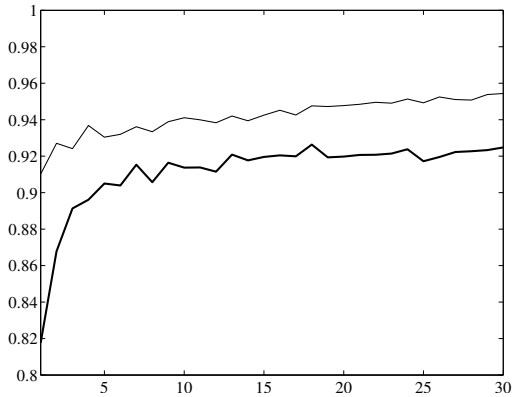
рак

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

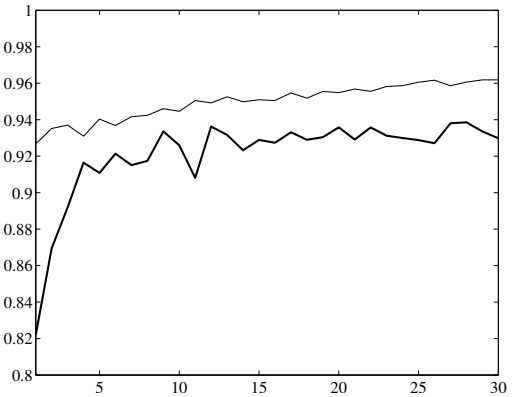
Зависимости AUC от числа используемых признаков K



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!

Две проблемы, мотивирующие нечёткое кодирование

1. **Проблема выбросов:**
   До 5% пар (R_n, T_n) в ЭКГ могут быть выбросами
2. **Проблема шума:**
   неопределённость $\operatorname{sign} dR_n$, $\operatorname{sign} dT_n$ при $dR_n \to 0$, $dT_n \to 0$

Вместо дискретизации $(T_n, R_n), (T_{n+1}, R_{n+1}) \to s_n$, $s_n \in \mathcal{A}$
оцениваем распределения $q_n(s)$ на $s \in \mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

s_n

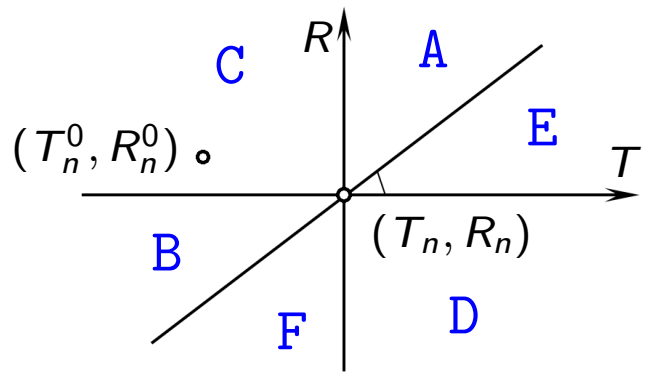
	B	F	A	B	D	F	D	E	E	C	A	B	C	C	F	E	A
A	10%	11%	48%	0%	15%	2%	0%	0%	0%	23%	49%	29%	3%	0%	1%	0%	59%
B	44%	0%	35%	58%	3%	7%	0%	12%	0%	0%	5%	52%	4%	27%	1%	12%	0%
C	28%	0%	13%	0%	0%	1%	11%	21%	0%	37%	1%	7%	83%	47%	2%	0%	0%
D	0%	0%	2%	1%	82%	0%	80%	0%	2%	19%	44%	6%	0%	0%	7%	0%	41%
E	5%	37%	0%	22%	0%	0%	9%	48%	98%	0%	0%	0%	10%	9%	0%	87%	0%
F	13%	52%	2%	19%	0%	90%	0%	19%	0%	21%	1%	6%	0%	17%	89%	1%	0%

$q_n(s)$

Модель измерений для нечёткого кодирования

R_n из распределения Лапласа, $\mathrm{E}R_n = R_n^0$, $\mathrm{D}R_n = \sigma_R^2$
T_n из распределения Лапласа, $\mathrm{E}T_n = T_n^0$, $\mathrm{D}T_n = \sigma_T^2$

Геометрическая интерпретация:
$q_n(a)$ — вероятность, что (T_n^0, R_n^0)
попадает в сектор $a \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

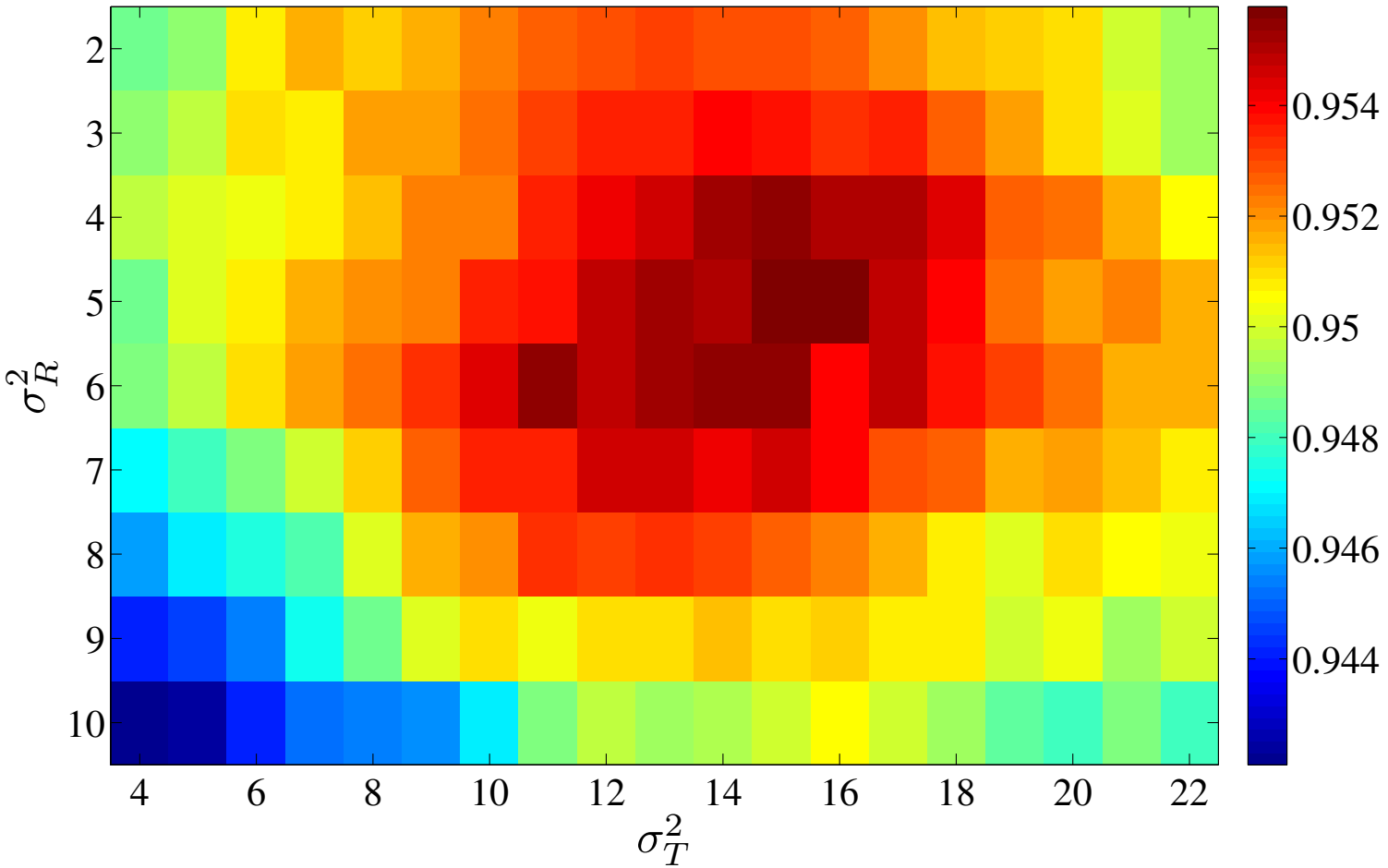


Нечёткая частота триграммы j, состоящей из символов abc:

$$f_j(x) = \frac{1}{N-3} \sum_{n=1}^{N-3} q_n(a)\, q_{n+1}(b)\, q_{n+2}(c).$$

Обработка выбросов:
если R_n — выброс, то $\mathrm{P}(R_{n-1} < R_n) = \mathrm{P}(R_n < R_{n+1}) = \frac{1}{2}$
если T_n — выброс, то $\mathrm{P}(T_{n-1} < T_n) = \mathrm{P}(T_n < T_{n+1}) = \frac{1}{2}$

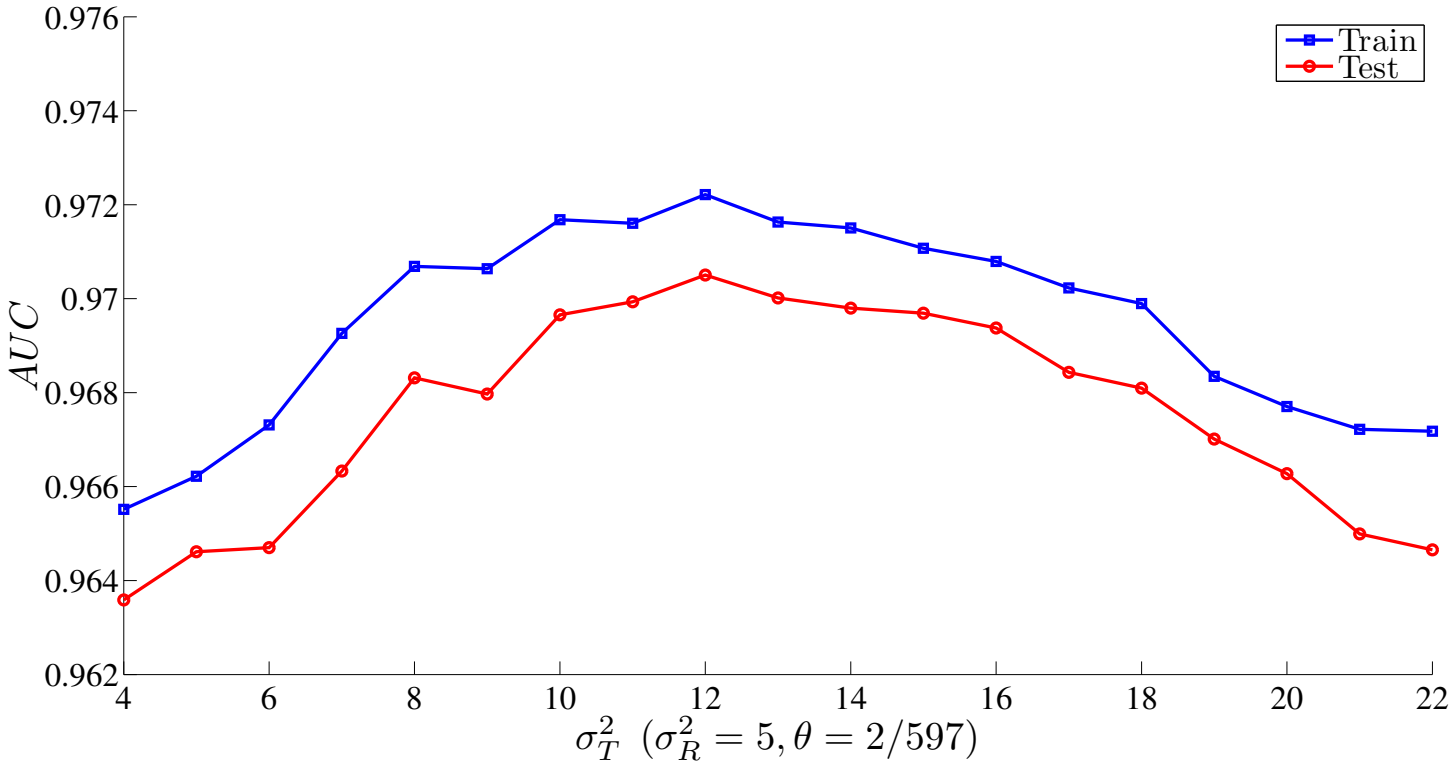
Оптимизация параметров в модели измерений



Найденный оптимум: $\sigma_T^2 = 15, \quad \sigma_R^2 = 5$

Кросс-валидация AUC

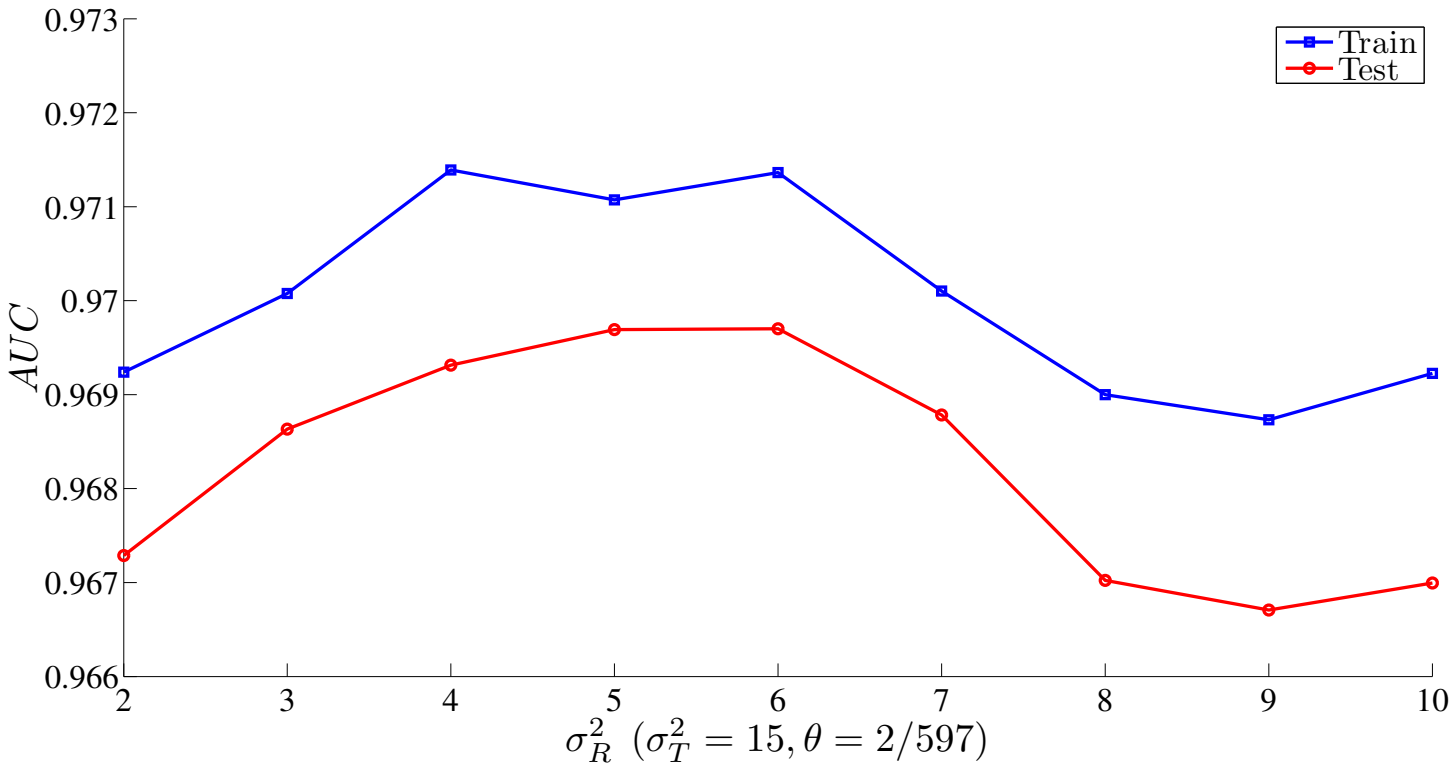
Болезнь: сахарный диабет



Вывод: нечёткое кодирование лучше дискретного (при $\sigma_T^2 = 0$)

Кросс-валидация AUC

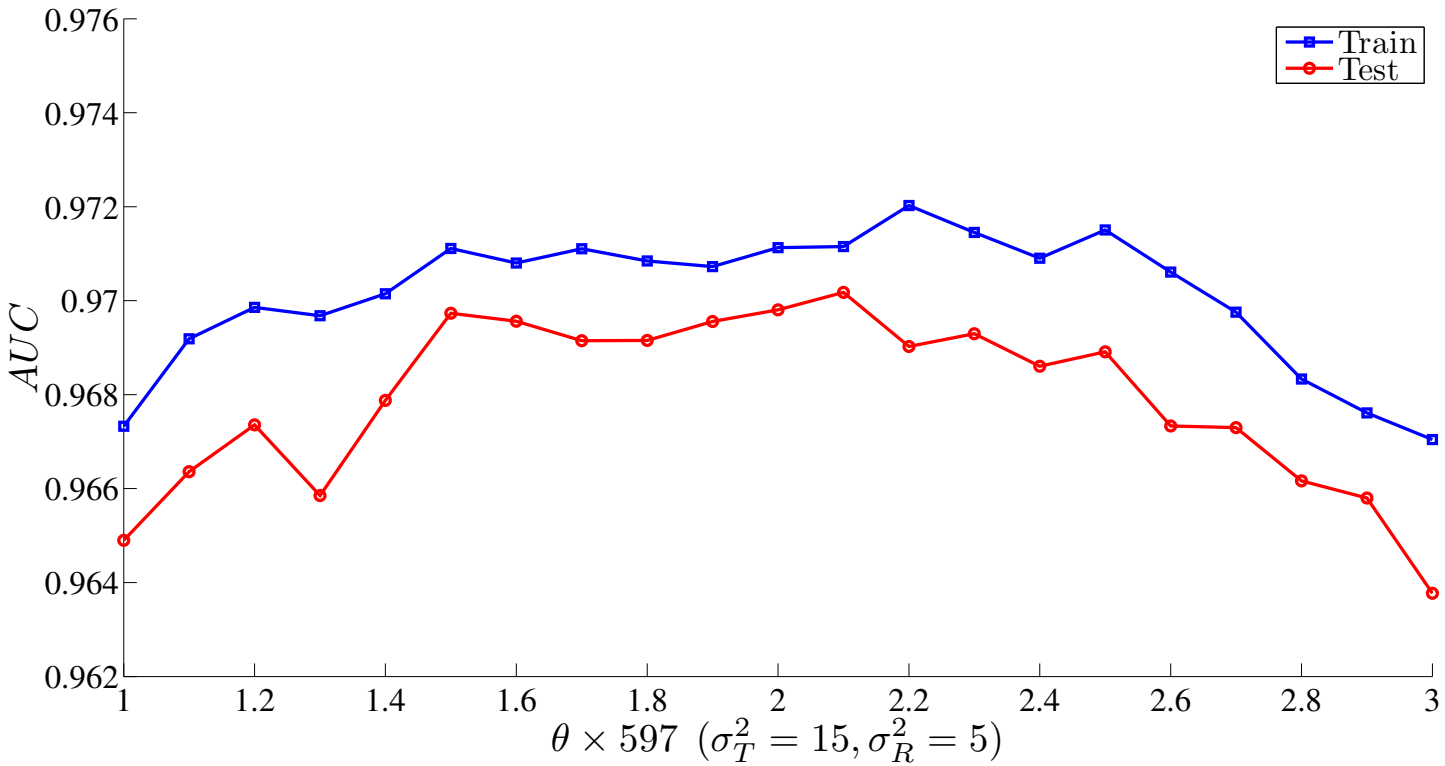
Болезнь: сахарный диабет



Вывод: нечёткое кодирование лучше дискретного (при $\sigma_R^2 = 0$)

Кросс-валидация AUC

Болезнь: сахарный диабет



Вывод: Триграммы, встречающиеся реже $\theta = 2$, незначимы.

Что ещё

Традиционная молодёжная летняя школа по оптимизации под руководством Б.Т.Поляка (ИПУ РАН)

Презентация, соревнование и решения участников:
www.MachineLearning.ru/wiki/images/0/0a/Voron-2014-06-26-school-VI.pdf

Данные для соревнования:
www.MachineLearning.ru/wiki/images/e/e1/School-VI-2014-task-3.rar

. . . легче зайти сюда и найти выступление «26 июня 2014»:
www.MachineLearning.ru/wiki/index.php?title=User:Vokov