Аддитивная регуляризация наивного байесовского классификатора

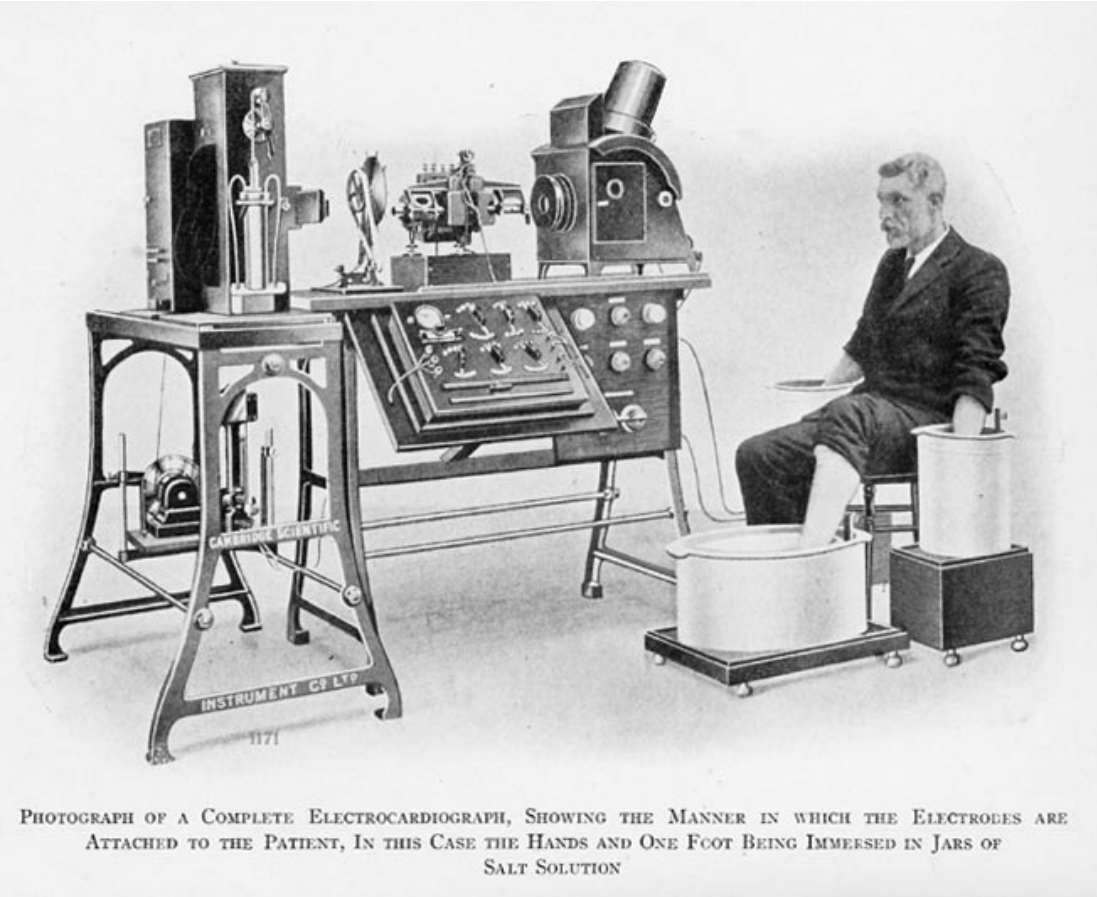
Воронцов Константин Вячеславович,
Целых В.Р., Шишковец С.С., Усков М.О.

ВЦ РАН ФИЦ ИУ РАН • ФУПМ МФТИ

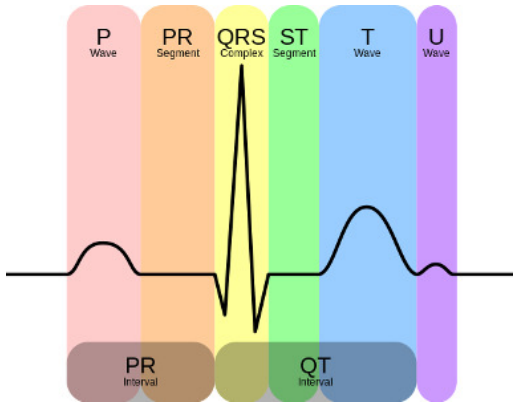
Девятая международная конференция
«Управление развитием крупномасштабных систем»
(MLSD'2016)
Москва • ИПУ РАН • 3–5 октября 2016

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

Теория информационной функции сердца [В.М.Успенский]

Возможна ли диагностика несердечных заболеваний по ЭКГ?

Предпосылки:

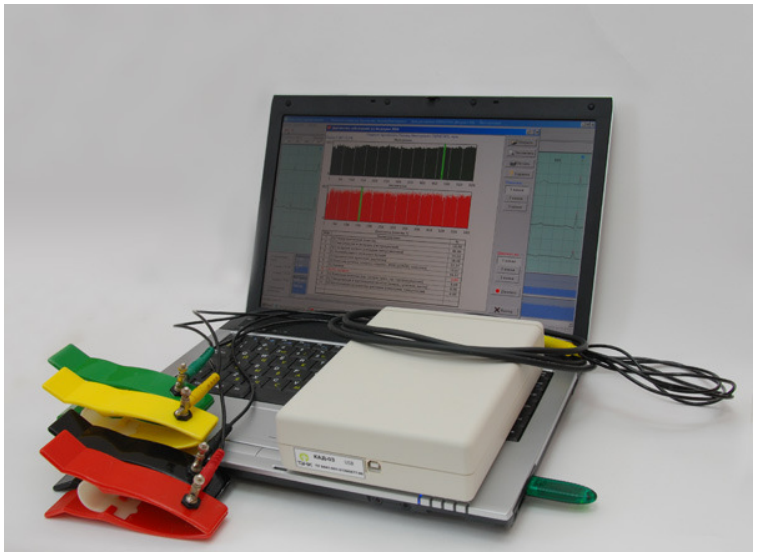
- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Цифровая электрокардиография высокого разрешения

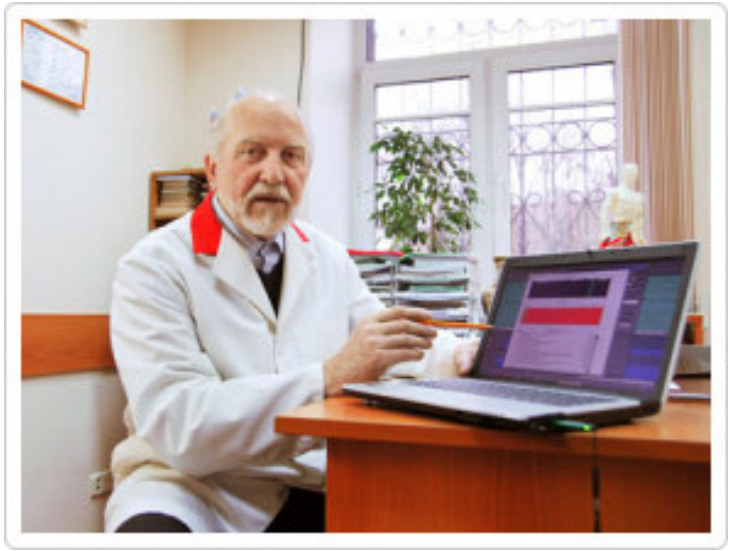
Предположения:

- 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

Диагностическая система «Скринфакс»

Цифровой электрокардиограф с улучшенной помехозащищённостью и расширенной полосой пропускания.





- более 15 лет исследований и накопления данных
- более 20 тысяч прецедентов (кардиограмма + диагнозы)
- 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:

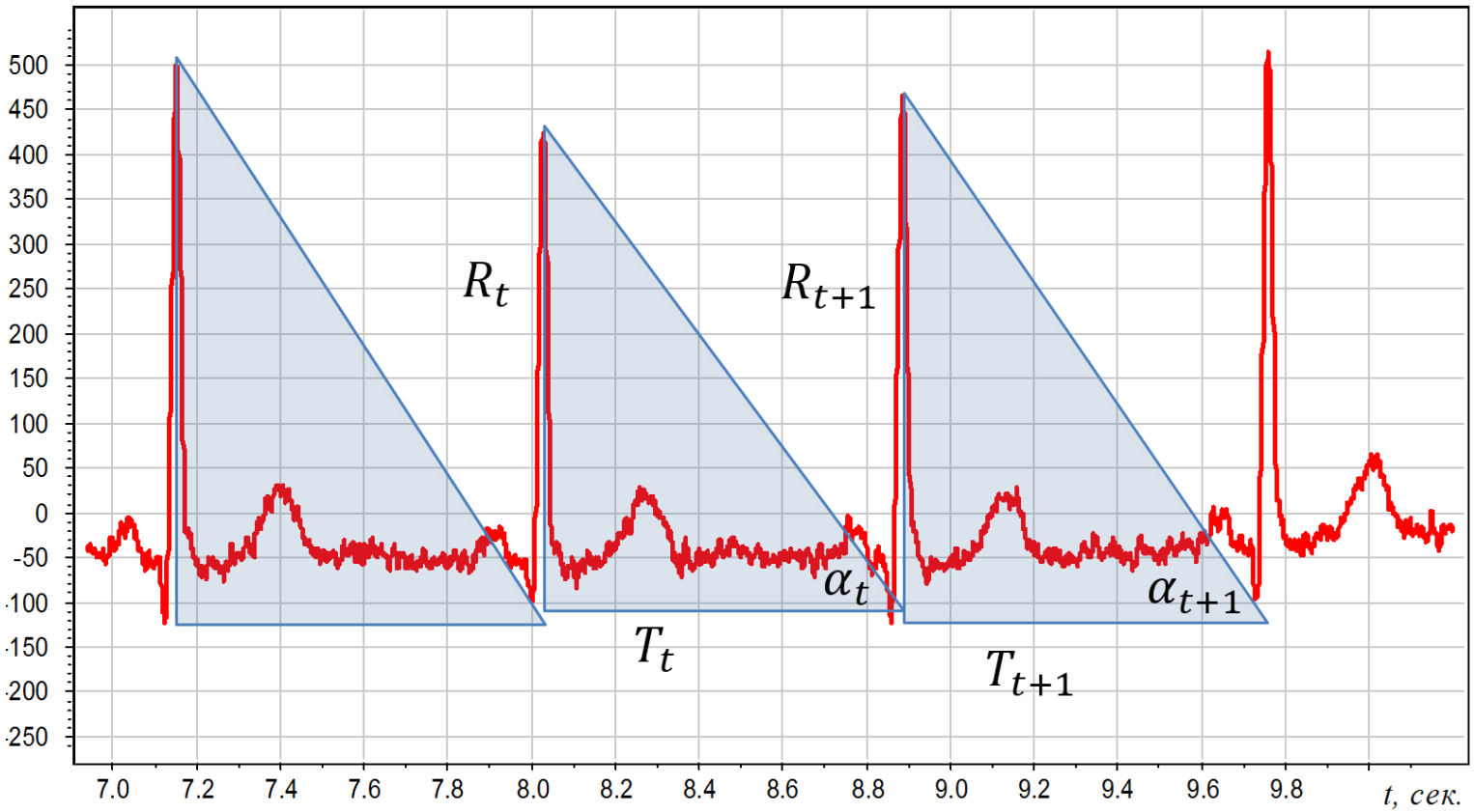
1. Формирование обучающих выборок здоровых и больных
2. Формировании модели классификации
3. Оптимизация модели классификации
4. Оценивание качества диагностики

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

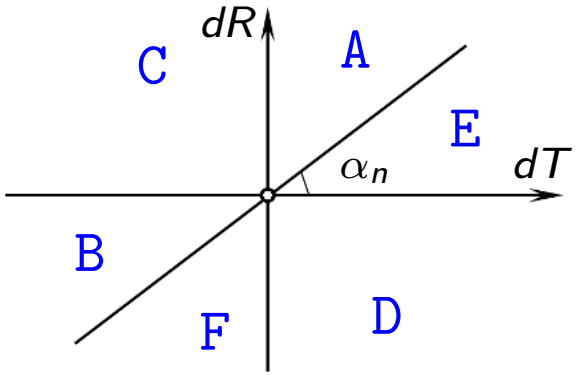


Дискретизация ЭКГ-сигнала

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

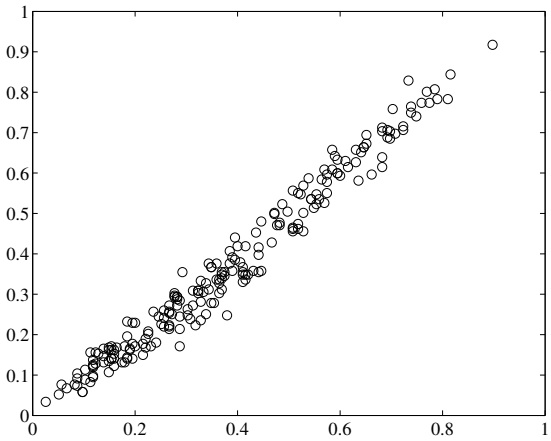
Существуют сочетания триграмм, специфичные для болезней

Точки на графиках соответствуют триграммам, $j = 1, \ldots, 216$

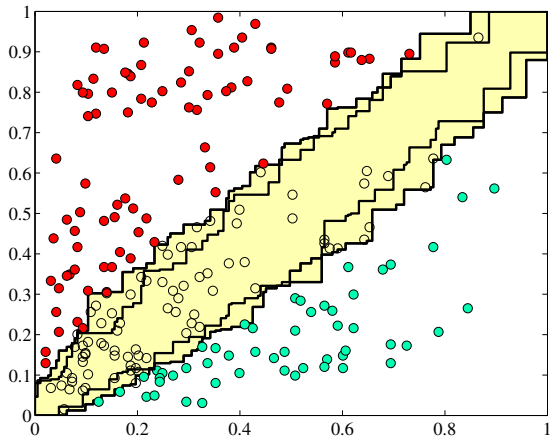
— ось X: доля здоровых x_i с частотой триграммы $x_i^j \geqslant 2$ из 600

— 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



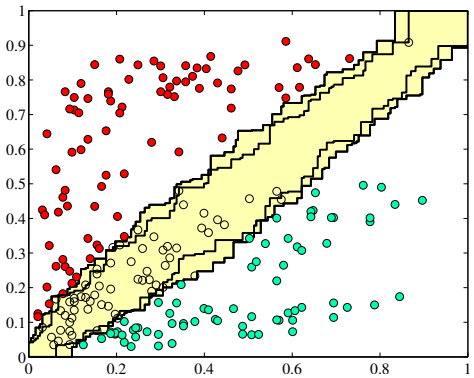
наблюдаемые y_i

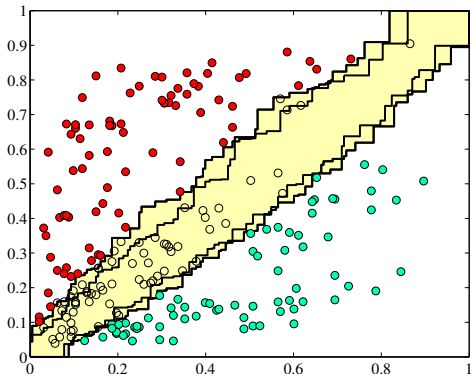
Слева: как распределятся точки, если объектам x_i назначить случайные (случайно перемешанные) метки классов y_i.

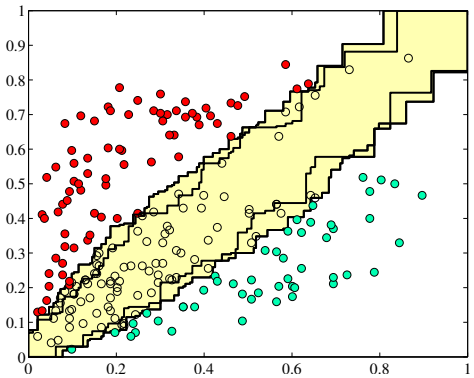
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

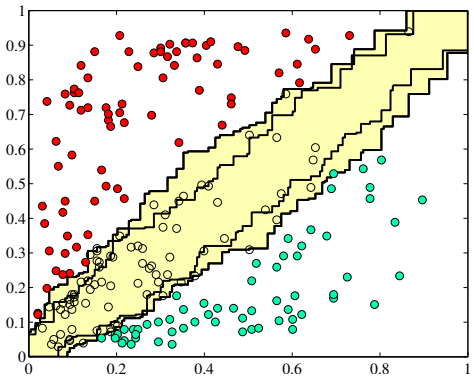
Существуют сочетания триграмм, специфичные для болезней

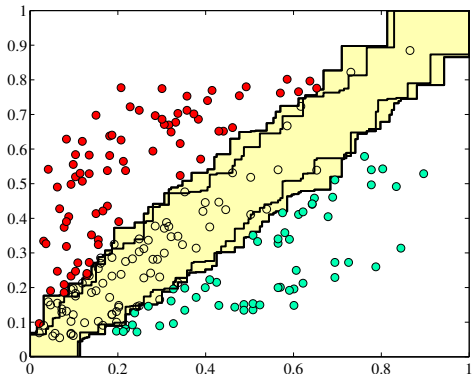
Для каждой болезни есть свои неслучайно частые триграммы

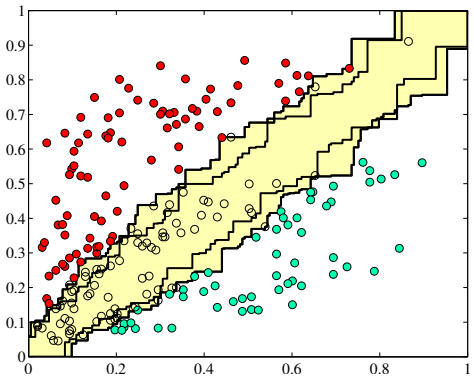

ишемия сердца


гипертония


рак


желчнокаменная болезнь


миома матки


язвенная болезнь

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 $\tilde{x}_1, \ldots, \tilde{x}_k$:

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

Результаты кросс-валидации

Обучающая выборка: оптимизация параметров модели
Тестовая выборка: Чувствительность, Специфичность, AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	С% 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

Байесовский классификатор

Пусть $\mathbb{X} \times \mathbb{Y}$ — в.п. с плотностью $p(x, y)$

Принцип максимума апостериорной вероятности:

$$a(x) = \arg\max_{y \in \mathbb{Y}} P(y|x) = \arg\max_{y \in \mathbb{Y}} P(y)p(x|y)$$

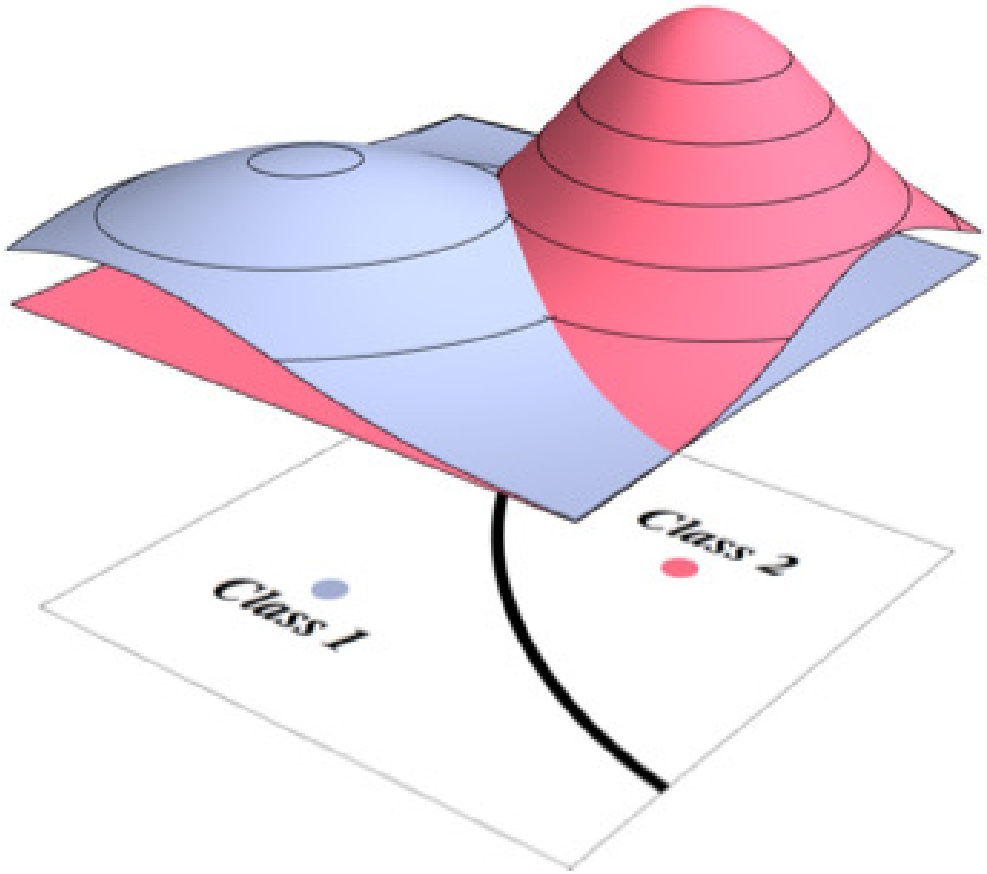
$\mathrm{P}(y|x)$ — *апостериорная вероятность* класса y,
$\mathrm{P}(y)$ — *априорная вероятность* класса y,
$p(x|y)$ — *модель плотности распределения* класса y.

Для двух классов, $\mathbb{Y} = \{-1, +1\}$:

$$a(x) = \operatorname{sign}\Big(P(+1)p(x|+1) - P(-1)p(x|-1)\Big) =$$
$$= \operatorname{sign}\left(\ln \frac{p(x|+1)}{p(x|-1)} + \ln \frac{P(+1)}{P(-1)}\right)$$

Байесовский классификатор для двух классов, $\mathbb{Y} = \{-1, +1\}$

$$a(x) = \text{sign}\left(\ln \frac{p(x|+1)}{p(x|-1)} + \ln \frac{P(+1)}{P(-1)}\right)$$



Задача оценивания плотностей классов по выборке

Пусть задана модель плотности с параметром θ_y:

$$p(x|y) = p(x|\theta_y)$$

Принцип максимума правдоподобия:

$$\ln \prod_{i=1}^{\ell} p(x_i, y_i) = \sum_{i=1}^{\ell} \ln p(x_i|y_i)P(y_i) \;\to\; \max_{\{\theta_y\}}.$$

Задача распадается на независимые подзадачи оценивания плотностей распределения классов $y \in \mathbb{Y}$:

$$\sum_{y\in\mathbb{Y}} \sum_{x_i\in X_y} \ln p(x_i|\theta_y) + \underbrace{|X_y| \ln P(y_i)}_{C=\text{const}} \;\to\; \max_{\{\theta_y\}},$$

где X_y — выборка объектов x_i класса $y_i = y$.

Наивный байесовский классификатор (Naïve Bayes, NB)

Пусть признаки статистически независимы:

$$p(x|y) = p(x^1|y) \cdots p(x^n|y), \qquad x = (x^1, \ldots, x^n)$$

Одномерная модель плотности с параметром θ_y^j:

$$p(x^j|y) = p(x^j|\theta_y^j)$$

Теперь задача максимизации правдоподобия распадается на независимые подзадачи ещё и по признакам j:

$$\mathcal{L} = \sum_{y \in \mathbb{Y}} \sum_{j=1}^{n} \sum_{x_i \in X_y} \ln p(x_i^j|\theta_y^j) \;\to\; \max_{\{\theta_y^j\}}.$$

В каких случаях она решается аналитически?

Экспоненциальное семейство плотностей

Пусть одномерные плотности $p(x^j|\theta_y^j)$ экспоненциальны:

$$p(x|\theta) = \exp\left(\frac{x\theta - c(\theta)}{\varphi} + h(x,\varphi)\right),$$

где θ и φ — числовые параметры распределения,
$c(\theta)$, $h(x,\varphi)$ — функциональные параметры.

Почему именно такое представление плотности?

1. многие полезные распределения экспоненциальны
2. максимизация правдоподобия выполняется аналитически
3. наивный байесовский классификатор линеен
4. оценивание параметров за $O(\ell n)$ — подходит для Big Data!

Примеры распределений из экспоненциального семейства

Нормальное (гауссовское) распределение, $x \in \mathbb{R}$:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) =$$
$$= \exp\left(\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right);$$

$$\theta = \mu, \qquad c(\theta) = \tfrac{1}{2}\theta^2, \qquad \varphi = \sigma^2.$$

Пуассоновское распределение, $x \in \{0,1,2,\dots\}$:

$$p(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} = \exp\left(\frac{x\ln(\mu) - \mu}{1} - \ln x!\right);$$

$$\theta = \ln(\mu), \qquad c(\theta) = e^{\theta}, \qquad \varphi = 1.$$

Примеры распределений из экспоненциального семейства

Биномиальное распределение, $x \in \{0, 1, \ldots, n\}$:

$$p(x|\mu, n) = C_n^x \mu^x (1-\mu)^{n-x} = \\ = \exp\left(x \ln \tfrac{\mu}{1-\mu} + n \ln(1-\mu) + \ln C_n^x\right);$$

$$\theta = \ln \tfrac{\mu}{1-\mu}, \qquad c(\theta) = n \ln(1 + e^\theta), \qquad \varphi = 1.$$

Распределение Бернулли, $x \in \{0, 1\}$:

$$p(x|\mu) = \mu^x (1-\mu)^{1-x} = \exp\left(x \ln \tfrac{\mu}{1-\mu} + \ln(1-\mu)\right);$$

$$\theta = \ln \tfrac{\mu}{1-\mu}, \qquad c(\theta) = \ln(1 + e^\theta), \qquad \varphi = 1.$$

Некоторые распределения из экспоненциального семейства

- нормальное (гауссовское)
- распределение Пуассона
- биномиальное и мультиномиальное
- геометрическое
- χ^2-распределение
- бета-распределение
- гамма-распределение
- распределение Дирихле
- распределение Лапласа с фиксированным матожиданием

Не экспоненциальные:

t-распределение Стьюдента, Коши, гипергеометрическое

Максимизация правдоподобия выполняется аналитически

Подставим экспоненциальную плотность

$$\ln p(x^j|\theta^j_y) = \frac{x^j\theta^j_y - c(\theta^j_y)}{\varphi^j_y} + h(x^j, \varphi^j_y)$$

в необходимое условие максимума правдоподобия:

$$\frac{\partial \mathcal{L}}{\partial \theta^j_y} = 0; \qquad \frac{\partial}{\partial \theta^j_y} \sum_{y\in\mathbb{Y}} \sum_{j=1}^{n} \sum_{x_i \in X_y} \ln p(x^j_i|\theta^j_y) = 0,$$

получим аналитическое решение, вычисляемое за $O(\ell n)$:

$$\boxed{\theta^j_y = [c']^{-1}\big(\langle x^j_i \rangle_y\big)} \quad \text{где} \quad \langle x^j_i \rangle_y = \frac{1}{|X_y|} \sum_{x_i \in X_y} x^j_i$$

Для оценивания параметров достаточно вычислить средние значения каждого признака x^j в каждом классе y

Наивный Байес линеен

Подставим экспоненциальные плотности в классификатор $a(x)$.

В случае двух классов, $\mathbb{Y} = \{-1, +1\}$:

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} x^j w_j - w_0\right), \qquad w_j = \sum_{y \in \mathbb{Y}} y \frac{\theta_y^j}{\varphi_y^j}$$

В случае произвольного числа классов:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \left(\sum_{j=1}^{n} x^j w_{yj} - w_{y0}\right), \qquad w_{yj} = \frac{\theta_y^j}{\varphi^j}$$

Пороги w_0, w_{y0} подбираются по внешним критериям, исходя из компромисса между чувствительностью и специфичностью

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Парадокс:

- NB неплохо решает задачи даже когда независимости нет
- пример — классификация символьных последовательностей

Почему? Гипотеза:

- любой линейный классификатор соответствует некоторому NB, если разрешить использовать смещённые оценки

Как смещать оценки и делать NB менее наивным:

- ввести регуляризатор отбора признаков
- ввести регуляризатор, приближающий NB к SVM
- но сохранить при этом $O(\ell)$

Наивный Байес с произвольным регуляризатором

Добавим регуляризатор $\mathcal{R}(w) \to \max$ с коэффициентом τ:

$$\mathcal{L} + \tau\mathcal{R} = \sum_{y\in\mathbb{Y}} \sum_{j=1}^{n} \sum_{x_i\in X_y} \left(\frac{x_i^j \theta_y^j - c(\theta_y^j)}{\varphi_y^j} \right) + \tau\mathcal{R}(w) \;\to\; \max_{\theta_y^j}$$

Теорема

Пусть плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству распределений и параметры разброса не зависят от класса, $\varphi_y^j \equiv \varphi^j$. Тогда точка максимума регуляризованного правдоподобия удовлетворяет системе уравнений

$$w_y^j = \frac{1}{\varphi^j}[c']^{-1}\left(\langle x_i^j \rangle_y + \frac{\tau}{|X_y|}\frac{\partial \mathcal{R}}{\partial w_y^j} \right).$$

Частные случаи регуляризаторов

1. L_0-регуляризатор с эффектом отбора признаков:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} \big[|w_j| > 0\big]$$

2. L_1-регуляризатор с эффектом отбора признаков:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} |w_j|$$

3. Декоррелирующий регуляризатор с эффектами отбора признаков и выделения диагностических эталонов классов:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} \sum_{y \in \mathbb{Y}} \sum_{z \in \mathbb{Y}} |w_{yj}| \cdot |w_{zj}|$$

Результаты кросс-валидации, AUC_n (%)

Гаусс, Пуассон: признаки — частоты триграмм x^j
Бернулли: признаки — встречаемости триграмм $[x^j \geqslant 2]$
RI (rule induction) — алгоритм, основанный на правилах

Сравнение NB нерегуляризованного, L_0 и L_1

		Гаусс			Пуассон			Бернулли		
	RI	NB	NB-L_0	NB-L_1	NB	NB-L_0	NB-L_1	NB	NB-L_0	NB-L_1
ЖКБ	74,8	89,1	$89{,}2_{97}$	$89{,}5_{83}$	88,6	$89{,}4_{7}$	$88{,}8_{73}$	90,0	$90{,}4_{30}$	$90{,}3_{31}$
ИБС	74,7	92,7	$92{,}9_{118}$	$92{,}8_{82}$	92,2	$92{,}5_{39}$	$92{,}2_{82}$	93,2	$93{,}8_{23}$	$93{,}4_{31}$
СД	63,4	90,0	$90{,}3_{114}$	$90{,}0_{175}$	89,4	$89{,}5_{7}$	$89{,}4_{63}$	90,7	$90{,}5_{27}$	$90{,}8_{29}$
УЩ	67,6	89,2	$89{,}4_{130}$	$89{,}4_{76}$	89,3	$89{,}6_{8}$	$89{,}3_{77}$	90,3	$90{,}2_{25}$	$90{,}7_{49}$
ЯБ	73,1	86,1	$86{,}2_{133}$	$86{,}1_{82}$	85,9	$87{,}3_{5}$	$85{,}9_{83}$	87,0	$87{,}7_{9}$	$87{,}0_{21}$

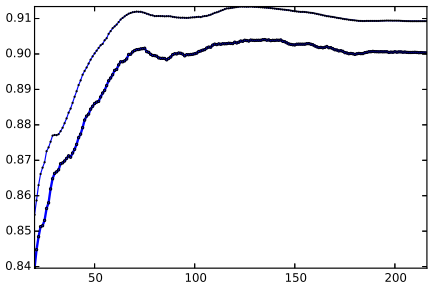
Вывод:

- Распределение Пуассона с L_0-регуляризацией сокращает число признаков почти без потери качества диагностики.

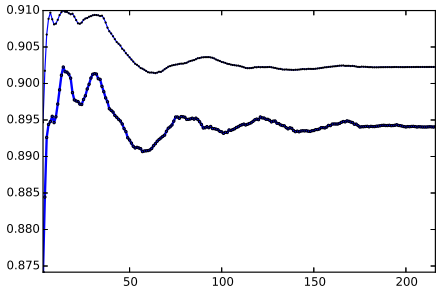
Зависимости AUC на обучении и тесте от числа признаков

Сахарный диабет:

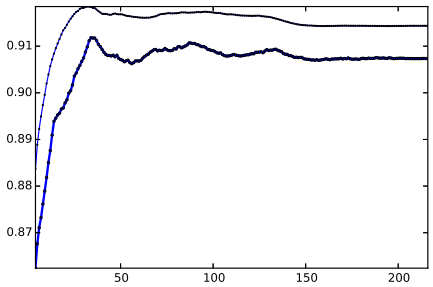
Гаусс



Пуассон

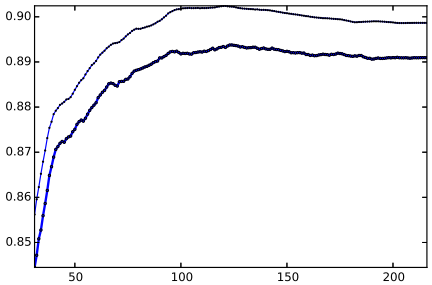


Бернулли

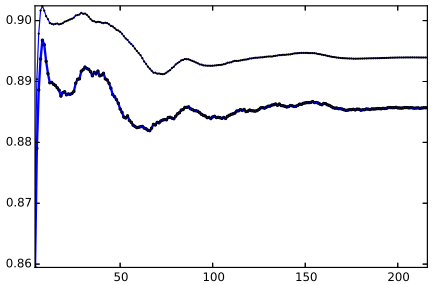


Желчнокаменная болезнь:

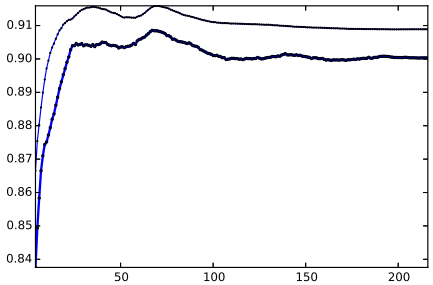
Гаусс



Пуассон



Бернулли



- Переобучение крайне мало́
- Оптимальное число признаков можно определять по обучению

Регуляризатор декоррелирования

Результаты кросс-валидации, AUC

Сравнение NB нерегуляризованного, L_0, L_1 а также сочетания L_0- и L_1-регуляризаторов с декоррелированием

Болезнь	NB	L_0	L_1	L_0-DC	L_1-DC
ВСД	83,40	82,02 (13)	84,01	82,45	83,85
ГБК	95,78	95,16 (4)	95,79	95,26	95,93
ГБ	94,64	94,54 (14)	94,79	94,66	95,80
ГД	92,80	91,96 (14)	92,98	91,78	92,95
ДЖВП	91,88	91,57 (7)	92,15	91,82	92,59
ЖКБ	96,82	96,04 (16)	97,04	96,70	97,98
ИБС	96,26	93,57 (17)	96,36	94,15	96,99
МКБ	92,90	92,72 (17)	93,08	92,79	93,86
ММ	91,20	90,99 (7)	91,30	91,29	91,93

- Декоррелирование немного уменьшает переобучение и повышает качество диагностики.

Выводы

- Точность диагностики превосходит многие методы
- Линейный наивный байесовский классификатор оказался неожиданно эффективным в этой задаче
- Регуляризация является способом ухода от гипотезы независимости и получения более интересных решений
- При этом время обучения остаётся линейным $O(\ell n)$

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov