Обучение ~~наивного~~ забавного Байеса

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

• Традиционная молодёжная летняя школа •
17 июня 2015

Содержание

Задача статистического (машинного) обучения с учителем

$\mathbb{X}$ — объекты; $\mathbb{Y}$ — ответы (классы, прогнозы);
$y^*: \mathbb{X} \to \mathbb{Y}$ — неизвестная зависимость.

Дано: X^ℓ — *обучающая выборка объектов* $x_i = (x_i^1, \ldots, x_i^n)$ с известными ответами $y_i = y^*(x)$, $i = 1, \ldots, \ell$:

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

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

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

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — $\langle$товар, магазин, дата$\rangle$; y — объём продаж;
  - x — $\langle$клиент, товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - x — характеристики недвижимости; y — цена;

Обучение регрессии — это оптимизация

Задача регрессии, $\mathbb{Y} = \mathbb{R}$

1. Выбираем модель регрессии, например, линейную:
$$a(x, w) = \langle x, w \rangle = \sum_{j=1}^{n} x^j w_j, \qquad x, w \in \mathbb{R}^n$$
2. Выбираем функцию потерь, например, квадратичную:
$$\mathscr{L}(a, y) = (a - y)^2$$
3. Минимизируем эмпирический риск — это МНК:
$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \big(a(x_i, w) - y_i\big)^2 \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$Q(a, X^k) = \frac{1}{k} \sum_{i=1}^{k} \big(a(\tilde{x}_i, w) - \tilde{y}_i\big)^2$$

Обучение классификации — это тоже оптимизация

Задача классификации с двумя классами, $\mathbb{Y} = \{-1, +1\}$

1. Выбираем модель классификации, например, линейную:
$$a(x, w) = \text{sign}\langle x, w\rangle$$
2. Выбираем функцию потерь, например, число ошибок:
$$\mathscr{L}(a, y) = \big[a(x_i, w)y_i < 0\big]$$
3. Минимизируем эмпирический риск:
$$Q(a, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[a(x_i, w)y_i < 0\big] \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$Q(a, X^k) = \frac{1}{k}\sum_{i=1}^{k}\big[a(\tilde{x}_i, w)\tilde{y}_i < 0\big]$$

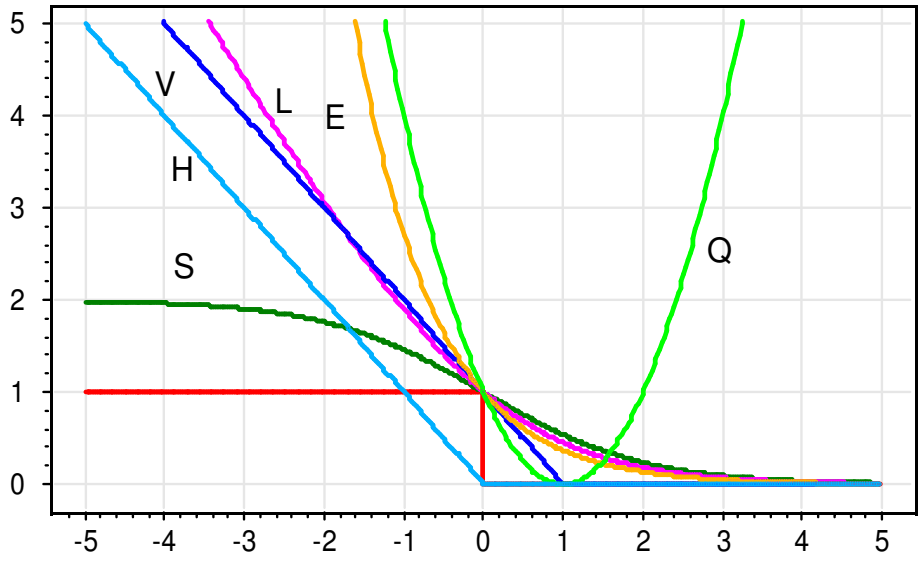
Минимизация аппроксимированного эмпирического риска

Задача классификации с двумя классами, $\mathbb{Y} = \{-1, +1\}$

1. Выбираем модель классификации, например, линейную:
$$a(x, w) = \text{sign}\langle x, w\rangle$$
2. Аппроксимируем пороговую функцию потерь непрерывной:
$$\big[M_i < 0\big] \leqslant \mathscr{L}\big(M_i\big), \qquad M_i = \langle x_i, w\rangle y_i \;\; — \;\; \text{отступ (margin)}$$
3. Минимизируем аппроксимированный эмпирический риск:
$$\tilde{Q}(a, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$Q(a, X^k) = \frac{1}{k}\sum_{i=1}^{k} \big[a(\tilde{x}_i, w)\tilde{y}_i < 0\big]$$

Непрерывные аппроксимации пороговой функции потерь

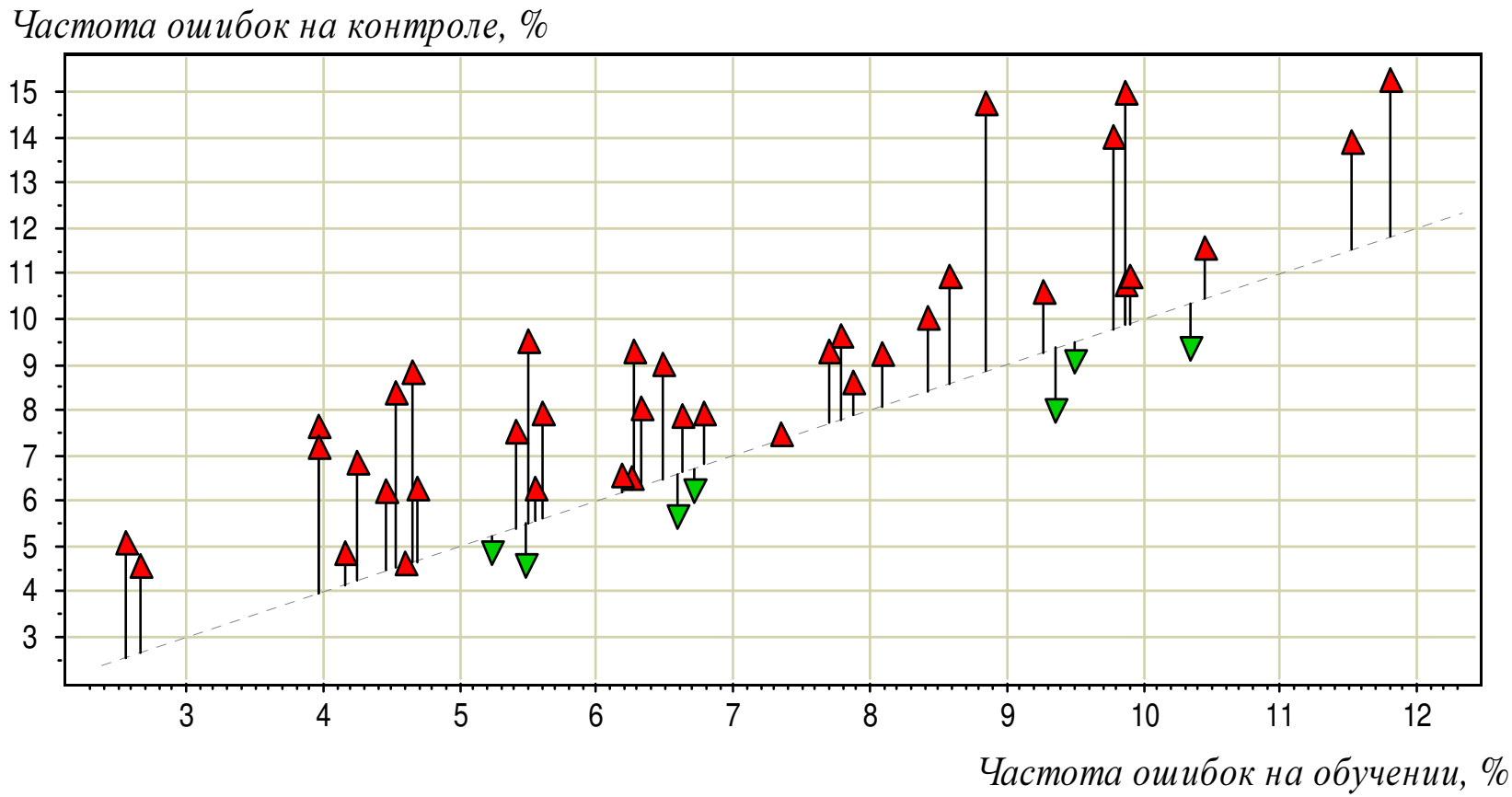
Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1 - M)_+$ — кусочно-линейная (SVM);
$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule);
$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (LR);
$Q(M) = (1 - M)^2$ — квадратичная (FLD);
$S(M) = 2(1 + e^{M})^{-1}$ — сигмоидная (ANN);
$E(M) = e^{-M}$ — экспоненциальная (AdaBoost);
$[M < 0]$ — пороговая функция потерь.

Проблема переобучения в прикладных задачах классификации

Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — различные алгоритмы.



Причины переобучения линейных моделей

1. слишком мало объектов; слишком много признаков;
2. линейная зависимость (мультиколлинеарность) признаков:
   $\exists v \in \mathbb{R}^n$: $\quad \forall x \;\; \langle x, v \rangle \approx 0$;
   тогда
   $\forall \gamma \in \mathbb{R} \quad a(x, w) = \text{sign}\langle x, w \rangle \approx \text{sign}\langle x, w + \gamma v \rangle$

Последствия:

- слишком большие веса $|w_j|$;
- неустойчивость классификаций $a(x, w)$;
- $Q(X^\ell) \ll Q(X^k)$;

Суть проблемы — задача некорректно поставлена.

Решение проблемы — регуляризация, ограничивающая w.

Часто используемые регуляризаторы

1. L_2-регуляризация (SVM, RLR, гребневая регрессия)

$$\tilde{Q}(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) + \mu \sum_{j=1}^{n} w_j^2 \to \min_w$$

2. L_1-регуляризация (LASSO)

$$\tilde{Q}(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) + \mu \sum_{j=1}^{n} |w_j| \to \min_w$$

3. L_0-регуляризация (AIC, BIC, VCdim, OBD)

$$\tilde{Q}(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) + \mu \sum_{j=1}^{n} [w_j \neq 0] \to \min_w$$

Наиболее известные линейные методы классификации

$a(x, w) = \text{sign}\big(\langle x, w\rangle - w_0\big)$ — линейный классификатор
$M_i(w, w_0) = y_i\big(\langle x_i, w\rangle - w_0\big)$ — *отступ* (margin) объекта x_i

1. Метод опорных векторов (SVM, Support Vector Machine):

$$\sum_{i=1}^{\ell}\big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \;\to\; \min_{w, w_0}.$$

2. Логистическая регрессия (LR, Logistic Regression):

$$\sum_{i=1}^{\ell}\ln\big(1 + \exp(-M_i(w, w_0))\big) \;\to\; \min_{w, w_0}.$$

3. Регуляризованная логистическая регрессия (RLR):

$$\sum_{i=1}^{\ell}\ln\big(1 + \exp(-M_i(w, w_0))\big) + \frac{\mu}{2}\|w\|^2 \;\to\; \min_{w, w_0}.$$

Негладкий регуляризатор приводит к отбору признаков

LASSO — least absolute shrinkage and selection operator

$$\sum_{i=1}^{\ell} \text{Loss}_i(w) + \mu \sum_{j=1}^{n} |w_j| \;\to\; \min_w .$$

Замена переменных: $u_j = \frac{1}{2}\big(|w_j| + w_j\big)$, $v_j = \frac{1}{2}\big(|w_j| - w_j\big)$.
Тогда $w_j = u_j - v_j$ и $|w_j| = u_j + v_j$;

$$\begin{cases} \sum_{i=1}^{\ell} \text{Loss}_i(u - v) + \mu \sum_{j=1}^{n} (u_j + v_j) \to \min_{u,v} \\ u_j \geqslant 0, \quad v_j \geqslant 0, \quad j = 1, \ldots, n; \end{cases}$$

чем больше μ, тем больше ограничений-неравенств активны,
но если $u_j = v_j = 0$, то $w_j = 0$ и j-й признак не учитывается.

1-norm SVM (LASSO SVM)

LASSO — least absolute shrinkage and selection operator

$$\sum_{i=1}^{\ell}\big(1 - M_i(w, w_0)\big)_+ + \mu \sum_{j=1}^{n} |w_j| \;\to\; \min_{w, w_0}.$$

- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊖ LASSO начинает отбрасывать значимые признаки, когда ещё не все шумовые отброшены
- ⊖ Нет *эффекта группировки* (grouping effect): значимые зависимые признаки должны отбираться вместе и иметь примерно равные веса w_j

---

Bradley P., Mangasarian O. Feature selection via concave minimization and support vector machines // ICML 1998.

Doubly Regularized SVM (Elastic Net SVM)

$$C\sum_{i=1}^{\ell}\big(1 - M_i(w, w_0)\big)_+ + \mu\sum_{j=1}^{n}|w_j| + \frac{1}{2}\sum_{j=1}^{n}w_j^2 \;\to\; \min_{w,w_0}.$$

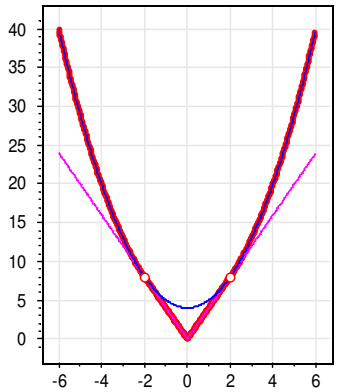
- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊕ Есть эффект группировки
- ⊖ Шумовые признаки также группируются вместе, и группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine // Statistica Sinica, 2006. №16, Pp. 589–615.

Support Features Machine (SFM)

$$C\sum_{i=1}^{\ell}\big(1 - M_i(w, w_0)\big)_+ + \sum_{j=1}^{n} R_\mu(w_j) \;\to\; \min_{w, w_0}.$$

$$R_\mu(w_j) = \begin{cases} 2\mu|w_j|, & |w_j| \leqslant \mu; \\ \mu^2 + w_j^2, & |w_j| \geqslant \mu; \end{cases}$$



- Отбор признаков с параметром *селективности* μ
- Есть эффект группировки
- Значимые зависимые признаки ($|w_j| > \mu$) группируются и входят в решение совместно (как в Elastic Net),
- Шумовые признаки ($|w_j| < \mu$) подавляются независимо (как в LASSO)

Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities // Multiple Classifier Systems. LNCS, Springer-Verlag, 2010. Pp. 165–174.

Relevance Features Machine (RFM)

$$C\sum_{i=1}^{\ell}\big(1 - M_i(w, w_0)\big)_+ + \sum_{j=1}^{n}\ln\big(w_j^2 + \frac{1}{\mu}\big) \;\to\; \min_{w, w_0}.$$

- Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- Есть эффект группировки
- Лучше отбирает набор значимых признаков, когда они лишь совместно обеспечивают хорошее решение

Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines // 19th International Conference on Pattern Recognition, Vol 1-6, 2008, Pp. 2336–2339.

Байесовский классификатор

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

Принцип максимума апостериорной вероятности:

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

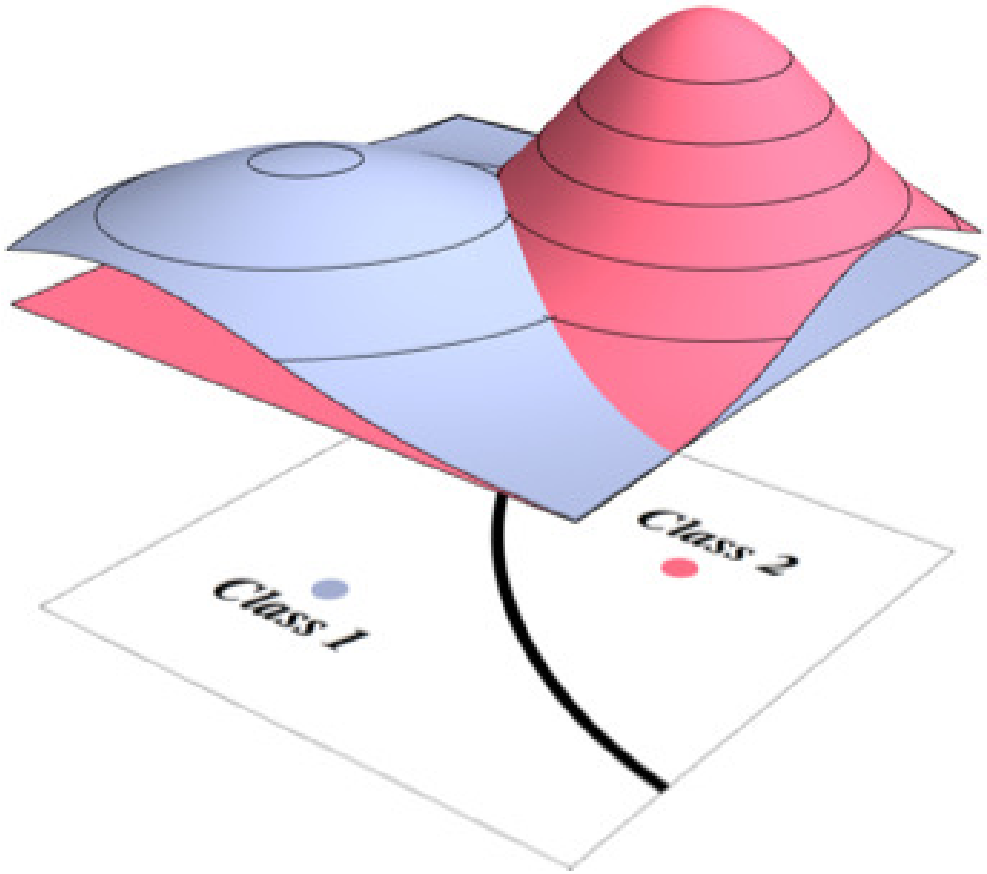
$\mathrm{P}(y|x)$ — *апостериорная вероятность* класса y,
$\mathrm{P}(y)$ — *априорная вероятность* класса y,
$p(x|y)$ — *модель плотности распределения* класса y.

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 распалась на независимые подзадачи по классам 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. оценки параметров вычисляются за $O(\ell)$
4. при этом наивный байесовский классификатор линеен

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 $O(\ell)$:

$$\frac{1}{\varphi^j_y} \sum_{x_i\in X_y} \left(x^j_i - c'(\theta^j_y)\right) = 0;$$

$$c'(\theta^j_y) = \frac{1}{|X_y|} \sum_{x_i\in X_y} x^j_i; \qquad \boxed{\theta^j_y = [c']^{-1}\left(\frac{1}{|X_y|} \sum_{x_i\in X_y} x^j_i\right)}$$

Наивный Байес линеен

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

$$a(x) = \text{sign}\left(\ln \frac{p(x|+1)}{p(x|-1)} + \underbrace{\ln \frac{P(+1)}{P(-1)}}_{-w_0}\right) =$$

$$= \text{sign}\left(\sum_{j=1}^{n} \ln p(x^j|\theta^j_+) - \ln p(x^j|\theta^j_-) - w_0\right) =$$

$$= \text{sign}\left(\sum_{j=1}^{n} \frac{x^j\theta^j_+ - c(\theta^j_+)}{\varphi^j_+} - \frac{x^j\theta^j_- - c(\theta^j_-)}{\varphi^j_-} - w_0\right) =$$

$$= \text{sign}\left(\sum_{j=1}^{n} x^j\left(\frac{\theta^j_+}{\varphi^j_+} - \frac{\theta^j_-}{\varphi^j_-}\right) - w_0\right) =$$

$$= \text{sign}\left(\sum_{j=1}^{n} x^j w_j - w_0\right) \qquad \boxed{w_j = \sum_{y\in\mathbb{Y}} y\frac{\theta^j_y}{\varphi^j_y}}$$

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Парадокс:

- NB неплохо решает задачи даже когда независимости нет

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Парадокс:

- NB неплохо решает задачи даже когда независимости нет

Почему? Гипотеза:

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

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Парадокс:

- NB неплохо решает задачи даже когда независимости нет

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 $O(\ell)$

Порассуждаем...

Недостаток NB:

- ограничение независимости признаков

Парадокс:

- NB неплохо решает задачи даже когда независимости нет

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 $O(\ell)$ — **подходит для Big Data!**

Наивный Байес с произвольным регуляризатором

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

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

Необходимое условие максимума по θ_y^j:

$$\frac{\partial}{\partial \theta_y^j}(\mathcal{L} + \tau\mathcal{R}) = 0; \qquad \sum_{x_i\in X_y} \frac{x_i^j - c'(\theta_y^j)}{\varphi_y^j} + \frac{\tau y}{\varphi_y^j} \frac{\partial \mathcal{R}}{\partial w_j} = 0$$

$$c'(\theta_y^j) = \frac{1}{|X_y|} \left(\sum_{x_i\in X_y} x_i^j + \tau y \frac{\partial \mathcal{R}}{\partial w_j} \right)$$

$$w_j = \sum_{y\in\mathbb{Y}} \frac{y}{\varphi_y^j} [c']^{-1} \left(\frac{1}{|X_y|} \left(\sum_{x_i\in X_y} x_i^j + \tau y \frac{\partial \mathcal{R}}{\partial w_j} \right) \right)$$

Частные случаи

1. L_2-регуляризатор $\mathcal{R}(w) = -\frac{1}{2}\sum_{j=1}^{n} w_j^2$:

$$w_j = \sum_{y\in\mathbb{Y}} \frac{y}{\varphi_y^j}[c']^{-1}\left(\frac{1}{|X_y|}\left(\sum_{x_i\in X_y} x_i^j - \tau y w_j\right)\right)$$

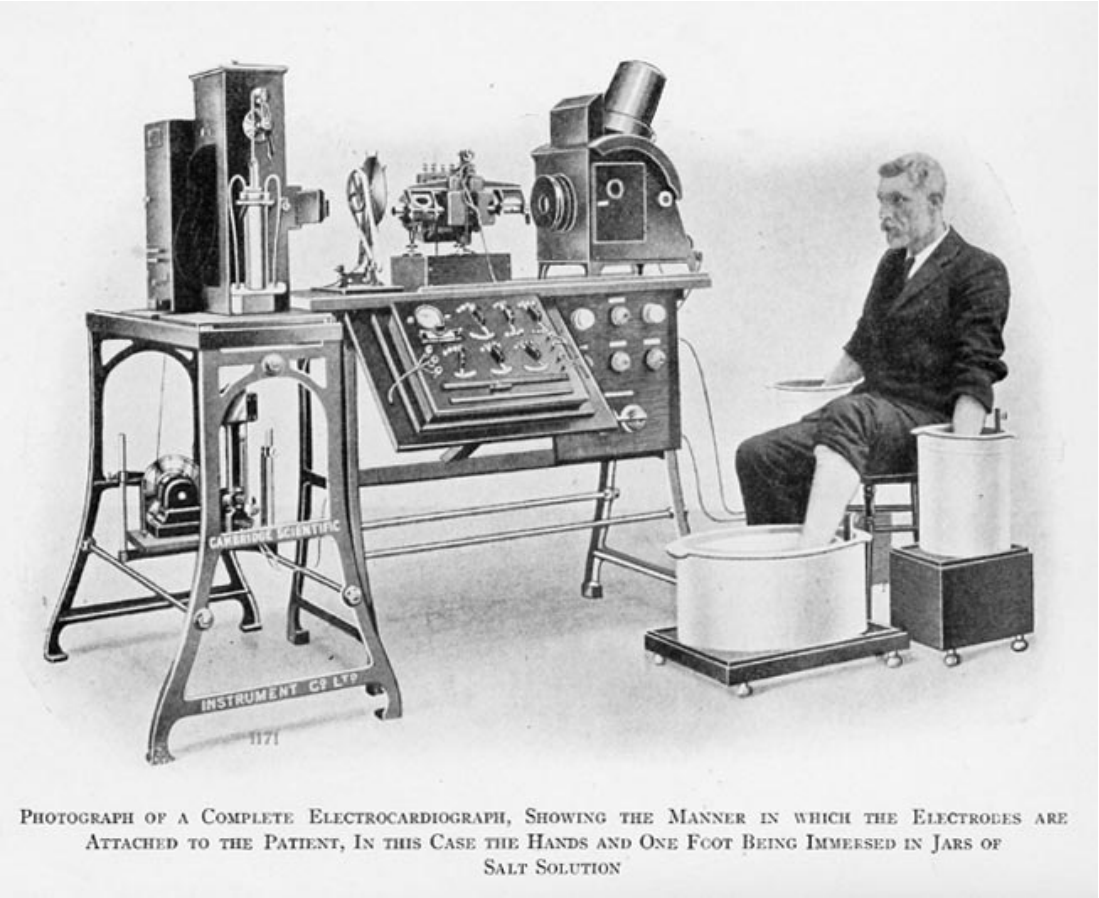
Нелинейное уравнение относительно w_j решаем численно, можно сразу для сетки значений τ.

2. Аппроксимированный эмпирический риск:
$\mathcal{R}(w) = -\sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big)$:

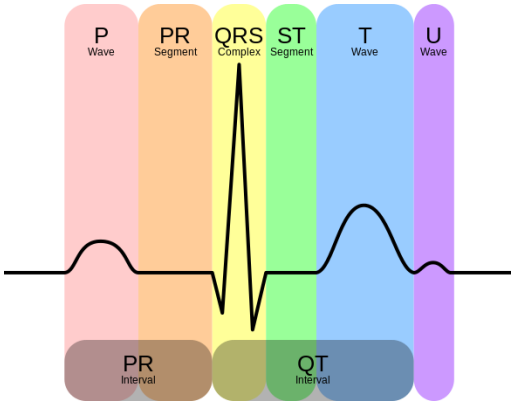
$$w_j = \sum_{y\in\mathbb{Y}} \frac{y}{\varphi_y^j}[c']^{-1}\left(\frac{1}{|X_y|}\left(\sum_{x_i\in X_y} x_i^j - \tau y \sum_{i=1}^{\ell} \mathscr{L}'\big(\langle x_i, w\rangle y_i\big) y_i x_i^j\right)\right)$$

Итерационный процесс, каждая итерация занимает $O(\ell)$.

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]

Предпосылки:

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

Предположения:

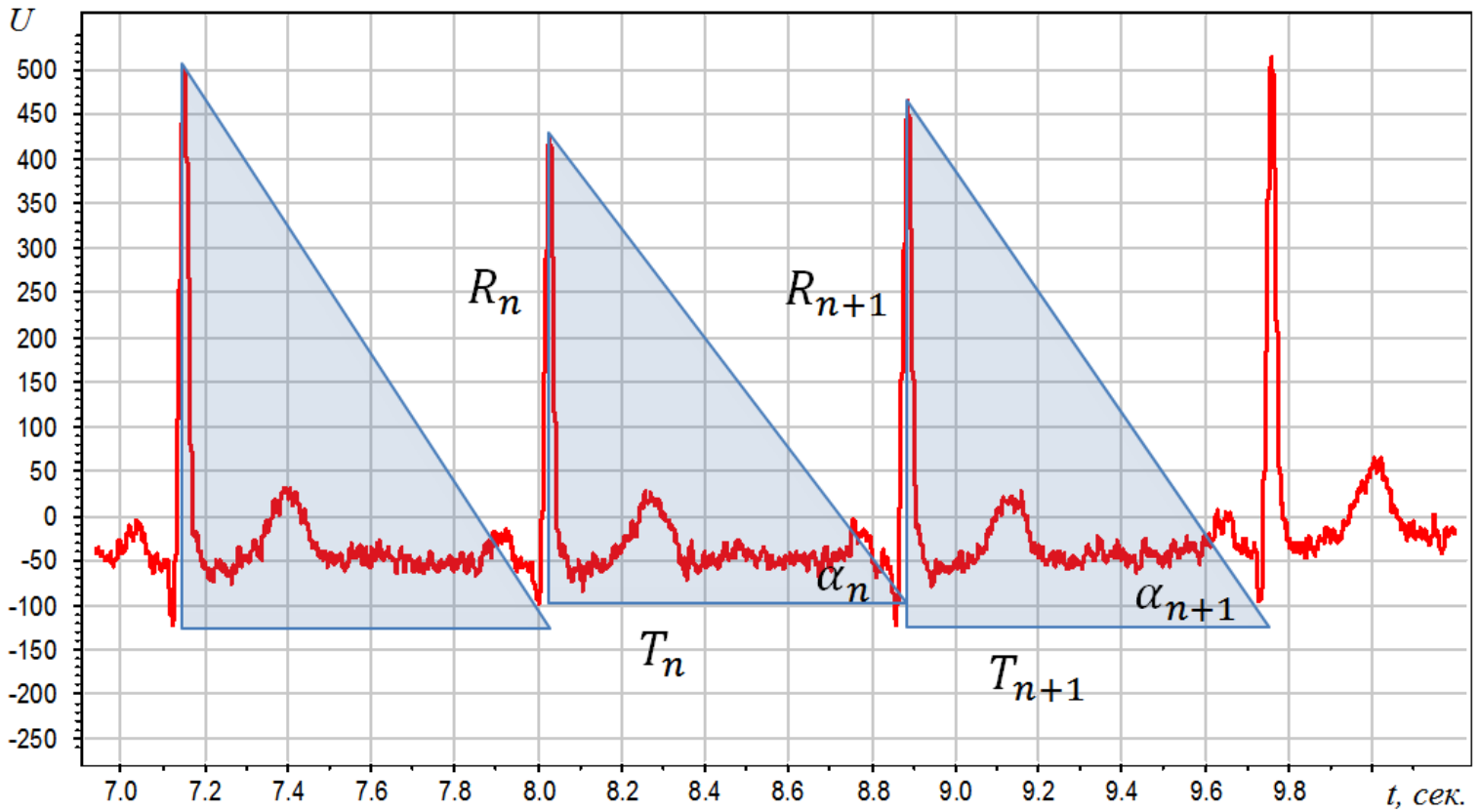
- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Каждое заболевание по-своему «модулирует» ЭКГ-сигнал
- Для диагностики важны *знаки приращений интервалов и амплитуд последовательных кардиоциклов*
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика

Приращения интервалов и амплитуд кардиоциклов

приращение амплитуд: $dR_n = R_{n+1} - R_n$

приращение интервалов: $dT_n = T_{n+1} - T_n$

приращение углов: $d\alpha_n = \alpha_{n+1} - \alpha_n, \qquad \alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$



Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^{N}$

Правила кодирования:

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

Выход: *кодограмма* $x = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

Векторизация кодограммы ЭКГ-сигнала

Вход: *кодограмма* $x = (s_1, \ldots, s_{N-1})$ как текстовая строка

Выход: частоты триграмм x^j — 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

Диагностическая система «Скринфакс» (2-е поколение)

- более 15 лет эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 40 заболеваний
- из них более 20 имеют отобранные эталонные выборки

Объём исходных данных (по заболеваниям)

абсолютно здоровые	АЗ	193
аденома простаты	ДГПЖ	260
аднексит хронический	АХ	276
анемия железодефицитная	ЖДА	260
асептический некроз головки бедренной кости	НГБК	324
вегетососудистая дистония	ВСД	694
гипертоническая болезнь	ГБ	1894
дискинезия желчевыводящих путей	ДЖВП	717
желчнокаменная болезнь	ЖКБ	278
ишемическая болезнь сердца	ИБС	1265
миома матки	ММ	781
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
сахарный диабет (СД1 и СД2)	СД	871
узловой (диффузный) зоб щитовидный железы	УЩ	748
холецистит хронический	ХХ	340
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
язвенная болезнь	ЯБ	785

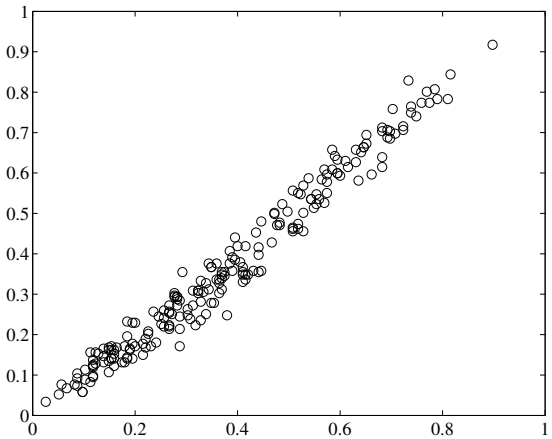
Нулевая гипотеза: частота триграммы не зависит от класса

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

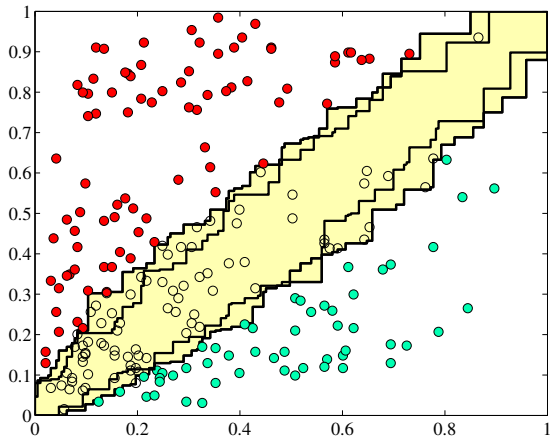
— ось X: доля здоровых i с частотой триграммы $x_i^j \geqslant 2$ из 600

— ось Y: доля больных i с частотой триграммы $x_i^j \geqslant 2$ из 600

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i

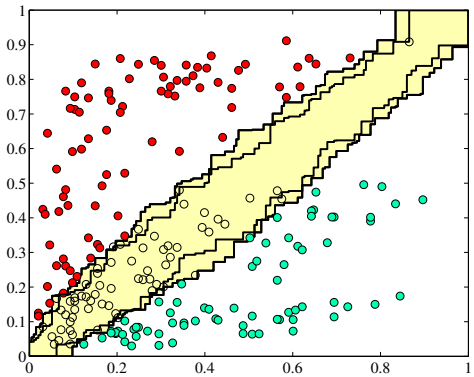


наблюдаемые y_i

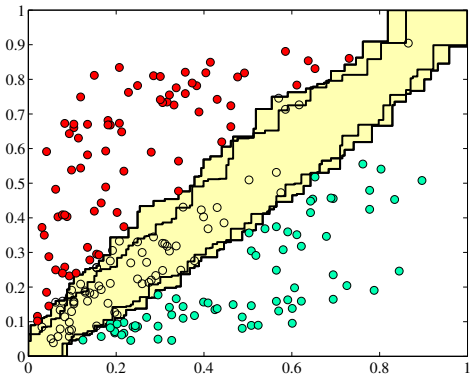
Нулевая гипотеза отвергается для большинства триграмм (красные и зелёные точки вне жёлтой области), при уровнях значимости 10% и 0.2% (20 и 1000 перемешиваний)

Результаты перестановочного теста для различных болезней

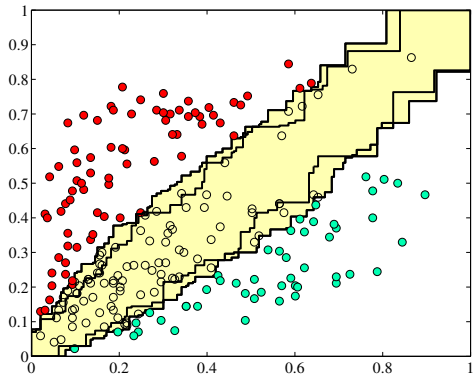
Для каждой болезни есть свои неслучайно частые триграммы



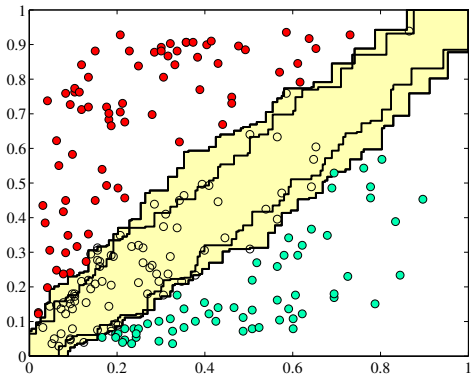
ишемия сердца



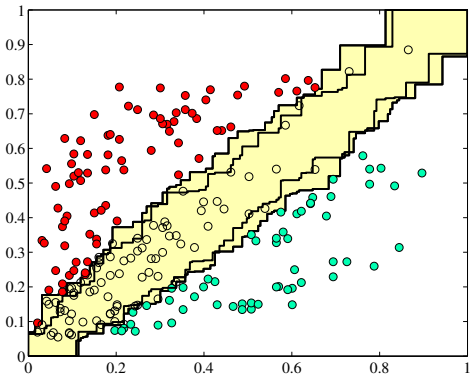
гипертония



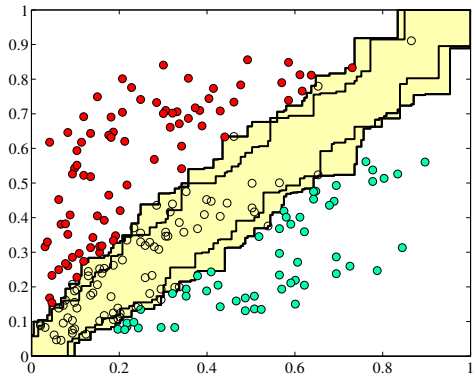
рак



желчнокаменная болезнь



миома матки

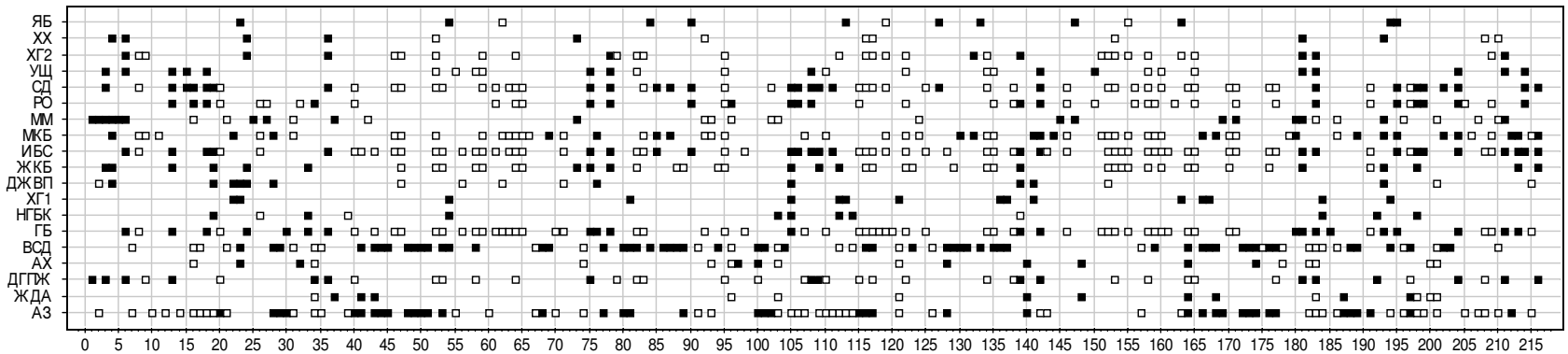


язвенная болезнь

Болезни отличаются наборами информативных триграмм

ось X: — номера триграмм 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 отличаются *диагностическими эталонами* — наборами информативных триграмм

Модель классификации

x_i — обучающая выборка кодограмм, $i = 1, \ldots, \ell$
y_i — диагноз: 0 = здоровый, 1 = больной
x_i^j — частота триграммы j в кодограмме i

Предположения:
1) для каждой болезни есть свой набор частых триграмм
2) если триграмма часто встречается, то не важно, сколько раз

Линейная модель классификации:

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

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

Стандартные критерии качества диагностики

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{1}{\ell_1} \sum_{i:\, y_i=1} \big[\langle x_i, w\rangle \geqslant w_0\big]$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{1}{\ell_0} \sum_{i:\, y_i=0} \big[\langle x_i, w\rangle < w_0\big]$$

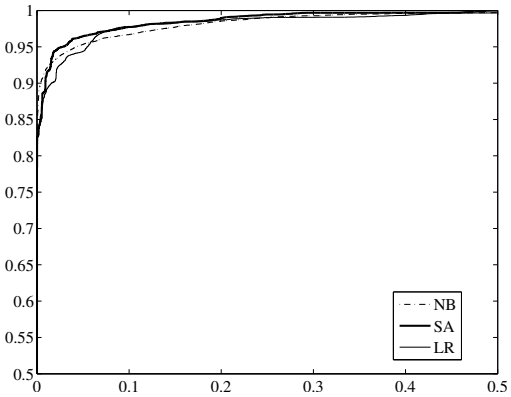
Area Under Curve — доля правильно упорядоченных пар:

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

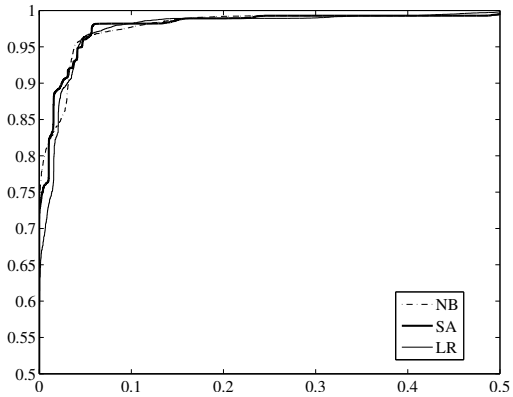
Результаты кросс-валидации

	Логистическая регрессия			Синдромный алгоритм		
	AUC, %	С (Ч=95%)	С=Ч, %	AUC, %	С (Ч=95%)	С=Ч, %
НГБК	99.26 ± 0.36	95.8 ± 1.6	95.2 ± 0.8	99.23 ± 0.05	97.4 ± 1.0	95.8 ± 0.9
ЖКБ	99.00 ± 0.25	94.9 ± 2.2	95.1 ± 1.1	98.90 ± 0.02	95.3 ± 0.5	95.5 ± 0.5
ИБС	98.21 ± 0.16	90.4 ± 1.3	93.1 ± 0.8	97.84 ± 0.03	91.8 ± 0.4	93.3 ± 0.0
ХГ1	97.64 ± 0.13	87.9 ± 1.9	91.9 ± 1.1	97.84 ± 0.09	89.4 ± 1.3	93.0 ± 0.8
СД	97.08 ± 0.14	84.1 ± 1.8	91.9 ± 0.9	96.66 ± 0.05	84.0 ± 0.9	91.2 ± 0.6
ГБ	96.91 ± 0.18	84.5 ± 2.7	91.8 ± 0.7	96.60 ± 0.05	81.6 ± 1.8	91.5 ± 0.4
РО	96.77 ± 0.17	82.7 ± 2.9	90.6 ± 0.9	95.81 ± 0.14	80.2 ± 3.0	90.5 ± 0.8
ДГПЖ	96.62 ± 0.40	77.2 ± 4.7	91.0 ± 1.2	96.59 ± 0.10	79.8 ± 3.7	91.2 ± 0.7
УЩ	95.75 ± 0.14	75.0 ± 2.7	90.2 ± 0.6	95.17 ± 0.10	66.7 ± 2.2	90.4 ± 0.6
ХГ2	95.22 ± 0.18	72.0 ± 2.0	88.4 ± 0.8	94.77 ± 0.11	71.7 ± 2.8	88.8 ± 1.0
ДЖВП	95.18 ± 0.15	73.4 ± 1.9	88.9 ± 0.9	95.14 ± 0.08	70.9 ± 2.2	89.1 ± 1.0
МКБ	95.11 ± 0.28	71.9 ± 3.5	88.6 ± 1.0	95.17 ± 0.07	69.0 ± 4.2	89.0 ± 0.3
ХХ	95.07 ± 0.21	73.4 ± 2.8	88.8 ± 1.3	95.51 ± 0.10	76.3 ± 1.9	90.1 ± 0.5
ЯБ	94.69 ± 0.40	66.2 ± 3.2	88.6 ± 1.4	94.67 ± 0.05	64.3 ± 2.5	89.6 ± 0.5
ММ	93.52 ± 0.30	60.5 ± 2.8	87.1 ± 1.1	93.37 ± 0.10	59.0 ± 2.1	87.6 ± 1.0
АХ	92.42 ± 0.48	62.7 ± 6.3	85.5 ± 2.0	91.90 ± 0.29	49.0 ± 3.4	85.6 ± 1.0
ЖДА	90.04 ± 0.60	54.4 ± 7.2	81.2 ± 1.8	89.27 ± 0.28	35.9 ± 6.1	83.0 ± 1.2
ВСД	87.62 ± 0.67	42.2 ± 5.0	79.9 ± 1.1	86.35 ± 0.24	39.5 ± 4.5	77.9 ± 1.0

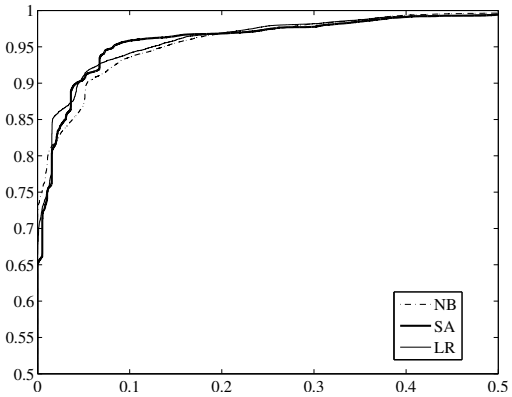
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



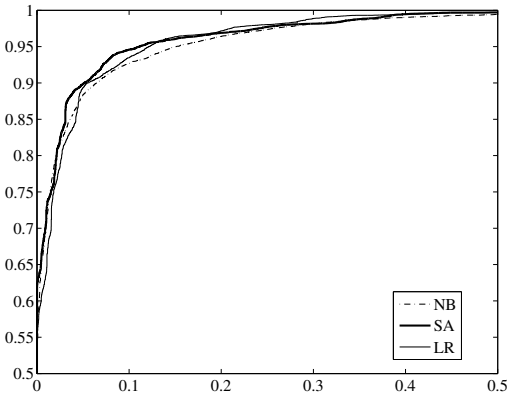
асептический некроз ГБК



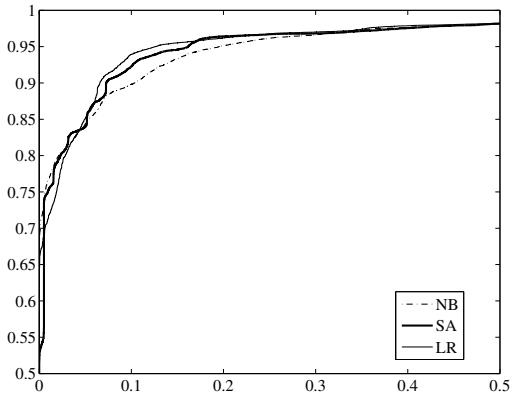
желчнокаменная болезнь



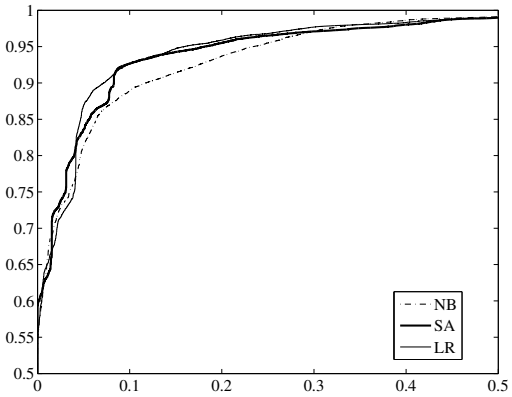
ишемическая болезнь



хронический гастрит 1



сахарный диабет

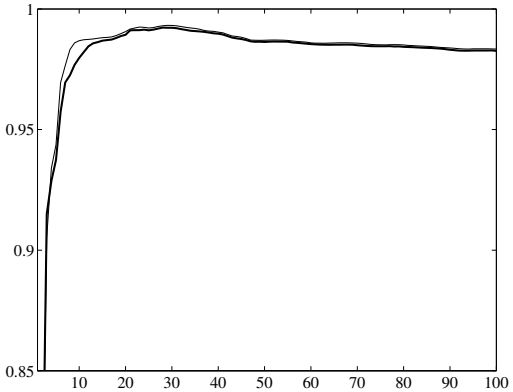


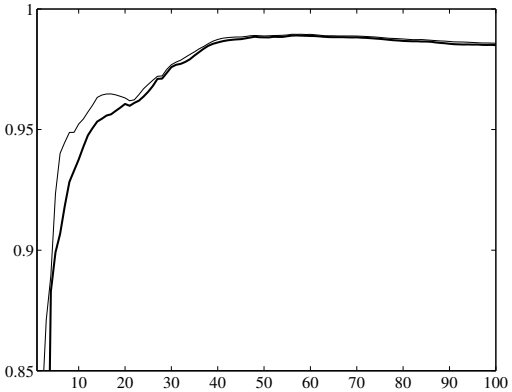
гипертония

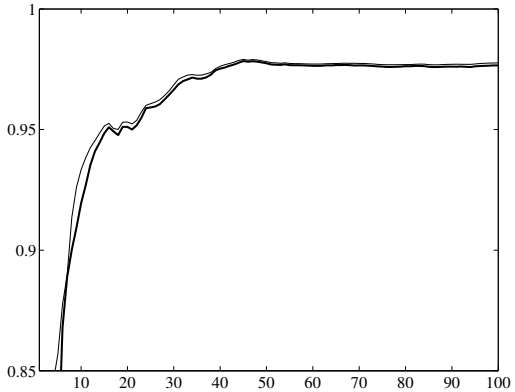
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

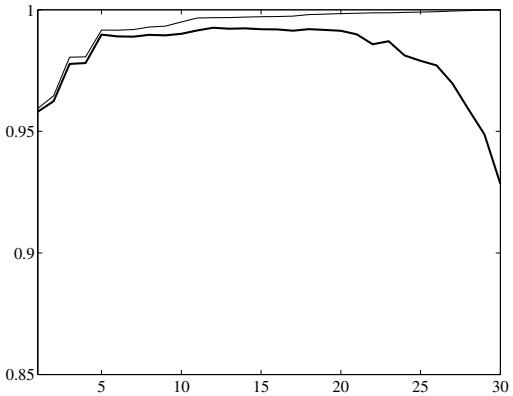
Синдромный алгоритм (наивный Байес на K признаках):

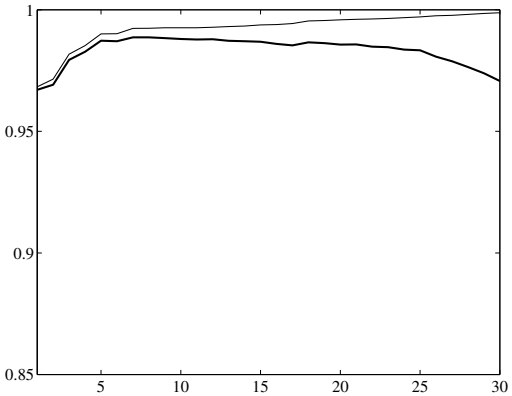


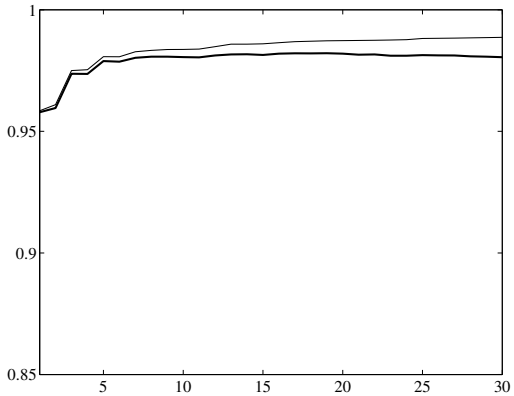




Логистическая регрессия (K — число главных компонент):







асептический некроз ГБК желчнокаменная болезнь ишемическая болезнь

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

ДНК конкурсной задачи

Дано:
матрица «объекты–признаки» по двум болезням
первый столбец — метки классов (0–здоровый, 1,2–больной),
остальные столбцы — 216 признаков,
строки — объекты (97 здоровых, по 200 больных)

Найти:
оценки объектов тестовой выборки, 496 объектов,
216 столбцов-признаков в том же порядке

Критерий: число ошибок классификации.

Данные — на странице
www.MachineLearning.ru/wiki?title=User:Vokov
www.MachineLearning.ru/wiki/images/b/b2/School-VII-2015-contest-Vorontsov.rar

Подсказки

Есть проблема:

- болезни плохо отличаются друг от друга

Что можно использовать для решения задачи:

- обучать три двухклассовых классификатора
- простые эвристики для сортировки и отбора признаков
- регуляризации для отбора признаков
- регуляризации наивного байесовского классификатора
- Python scikit-learn, Matlab, и др.

Чем ещё решали эту задачу:

- логистическая регрессия на главных компонентах
- бустинг над деревьями решений
- нейронная сеть
- тематические модели

Переоценка ценностей

В машинном обучении не всегда и не столь важно,

- какова скорость сходимости,
- есть ли вообще сходимость,
- насколько точно вычисляется решение,
- сколько времени уходит на поиск решения...

Новые вопросы выходят на первый план:

- как выбрать правильную модель зависимости,
- как учесть знания экспертов о предметной области,
- как синтезировать признаки по сырым данным,
- как отобрать из них информативные признаки,
- как избежать переобучения...

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

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