Методы статистического обучения. Задача диагностики заболеваний по электрокардиограмме

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

• Традиционная молодёжная летняя школа •
26 июня 2014

Содержание

Задача статистического (машинного) обучения с учителем

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

Дано: $x_i = (x_i^1, \ldots, x_i^n)$ — *обучающие объекты* 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, в данном случае 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:
$$Q(a, X^k) = \frac{1}{k}\sum_{i=1}^{k}\big[a(\tilde{x}_i, w)\tilde{y}_i < 0\big]$$

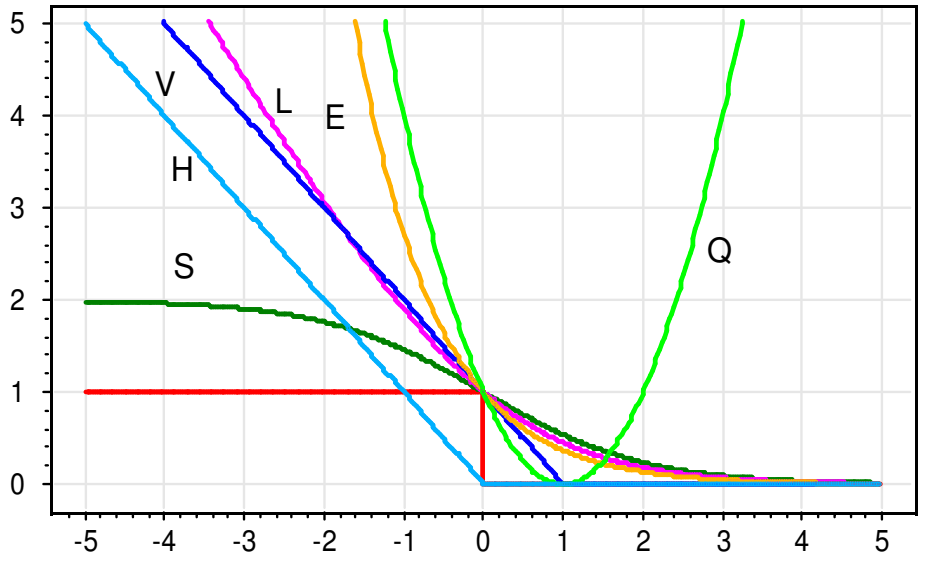
Минимизация аппроксимированного эмпирического риска

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

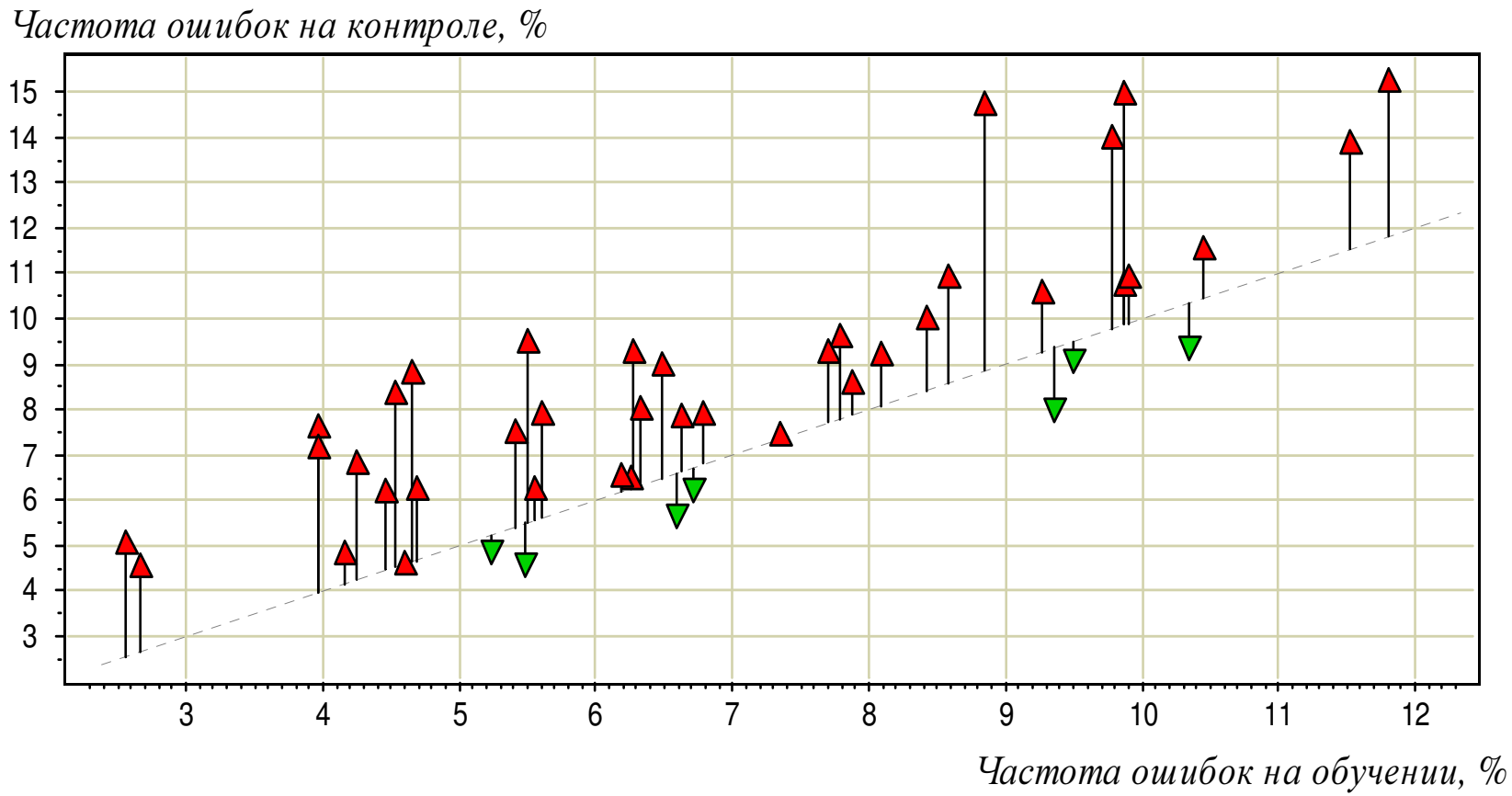
Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1 - M)_+$ — 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:
   пусть построен классификатор: $a(x, w) = \text{sign}\langle x, w\rangle$;
   мультиколлинеарность: $\exists v \in \mathbb{R}^n$: $\forall x \;\; \langle x, v\rangle \approx 0$;
   тогда $\forall \gamma \in \mathbb{R} \quad a(x, w) \approx \text{sign}\langle x, w + \gamma v\rangle$

Последствия:

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

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)

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

Метод опорных векторов (SVM, Support Vector Machine)

Задача классификации: $X = \mathbb{R}^n$, $Y = \{-1, +1\}$,
по обучающей выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$
найти параметры $w \in \mathbb{R}^n$, $w_0 \in \mathbb{R}$ алгоритма классификации

$$a(x, w) = \operatorname{sign}\big(\langle x, w\rangle - w_0\big).$$

Минимизация аппроксимированного регуляризованного эмпирического риска:

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

где $M_i(w, w_0) = y_i\big(\langle x_i, w\rangle - w_0\big)$ — *отступ* (margin) объекта x_i.

Почему именно такая функция потерь? и такой регуляризатор?

Оптимальная разделяющая гиперплоскость

Линейный классификатор:

$$a(x, w) = \operatorname{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}.$$

Пусть выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ линейно разделима:

$$\exists w, w_0 : \quad M_i(w, w_0) = y_i\big(\langle w, x_i\rangle - w_0\big) > 0, \quad i = 1, \ldots, \ell$$

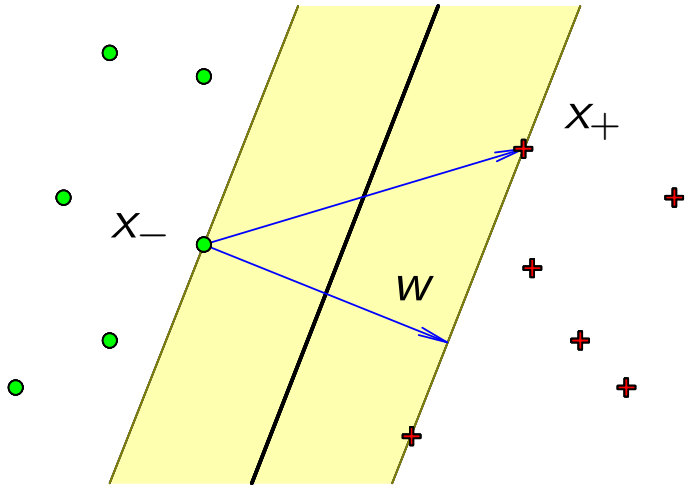
Нормировка: $\min\limits_{i=1,\ldots,\ell} M_i(w, w_0) = 1.$

Разделяющая полоса:

$$\big\{x:\ -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\big\}.$$

Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max.$$



Обоснование кусочно-линейной функции потерь

Линейно разделимая выборка

$$\begin{cases} \|w\|^2 \to \min\limits_{w,w_0}; \\ M_i(w,w_0) \geqslant 1, \quad i = 1,\ldots,\ell. \end{cases}$$

Переход к линейно неразделимой выборке (эвристика)

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1 - \xi_i, \quad i = 1,\ldots,\ell; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,\ell. \end{cases}$$

Эквивалентная задача безусловной минимизации:

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

Понятие «опорного вектора»

Двойственная задача:

$$\begin{cases} -\sum\limits_{i=1}^{\ell}\lambda_i + \frac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle \;\to\; \min\limits_{\lambda}; \\ \sum\limits_{i=1}^{\ell}\lambda_i y_i = 0; \qquad 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell. \end{cases}$$

Типизация объектов:

1. $\lambda_i = 0;\ \ \eta_i = C;\ \ \xi_i = 0;\ \ M_i \geqslant 1.$
   — периферийные (неинформативные) объекты.

2. $0 < \lambda_i < C;\ \ 0 < \eta_i < C;\ \ \xi_i = 0;\ \ M_i = 1.$
   — опорные граничные объекты.

3. $\lambda_i = C;\ \ \eta_i = 0;\ \ \xi_i > 0;\ \ M_i < 1.$
   — опорные-нарушители.

Отбор опорных объектов и отбор признаков

Важное свойство, которое есть у SVM:
решение зависит только от опорных объектов ($\lambda_i > 0$):

$$w = \sum_{i=1}^{\ell} \lambda_i y_i x_i; \qquad a(x) = \operatorname{sign}\Big(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle - w_0 \Big).$$

Отбор опорных объектов возникает по причине негладкости функции потерь $\mathscr{L}(M_i)$ в точке $M_i = 0$.

Важное свойство, которого нет у SVM:
нет отбора признаков, решение зависит от всех признаков.

Отбор признаков возникает по причине негладкости регуляризатора в точке $w = 0$.

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 меньше 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)

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

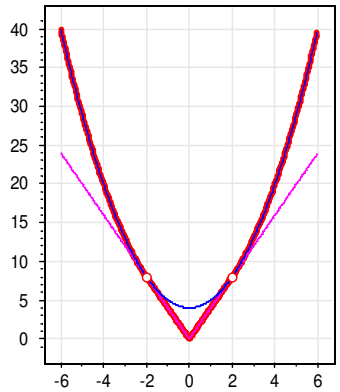
- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊕ Есть эффект группировки
- ⊖ Шумовые признаки также группируются вместе, и группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

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)

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

- 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.

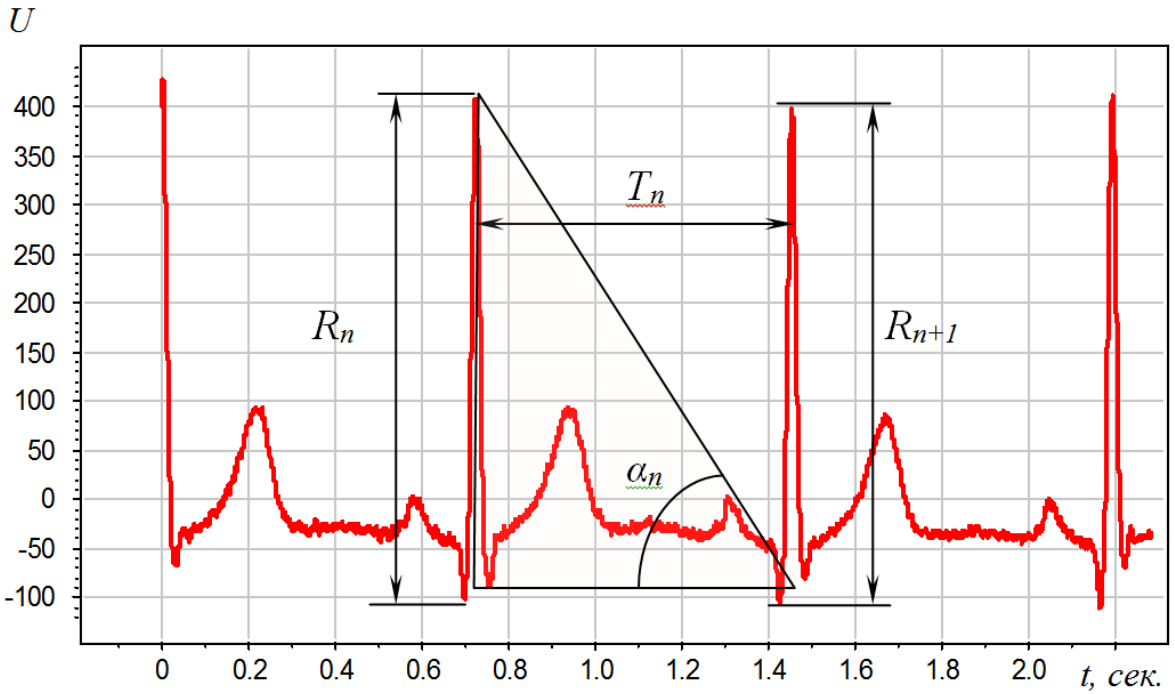
Резюме по методам аппроксимации и регуляризации

- Непрерывная аппроксимация пороговой функции потерь упрощает оптимизацию, увеличивает зазор между классами, тем самым повышает обобщающую способность.
- Регуляризаторы устраняют неустойчивость и переобучение.
- Негладкие функции потерь $\rightarrow$ отбор опорных объектов.
- Негладкие регуляризаторы $\rightarrow$ отбор признаков.
- Методы отбора признаков: LASSO, Elastic Net, методы Александра Татарчука: SFM, RFM.

Открытые проблемы:

- масшабируемые онлайновые методы отбора признаков
- универсально лучшие методы отбора признаков

Информационный анализ электрокардиосигналов



Открытие д.м.н. проф. В.М.Успенского:
для ранней диагностики многих заболеваний по ЭКГ достаточно использовать только знаки приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.

Диагностическая система «Скринфакс» (2-е поколение)

- более 10 лет эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 50 заболеваний
- из них более 20 имеют отобранные эталонные выборки

Технология информационного анализа ЭКГ по В.М.Успенскому

1. вычисление амплитуд, интервалов и углов по кардиограмме длиной 600 кардиоциклов
2. вычисление *кодограммы* — 599-символьной строки в 6-буквенном алфавите
3. вычисление $6^3 = 216$ признаков — частот триграмм
4. формирование эталонных выборок:
   1) абсолютно здоровых,
   2) больных (для каждого заболевания отдельная выборка)
5. поиск *диагностических эталонов* — наборов триграмм, часто совместно встречающихся у больных данным заболеванием, и редко встречающихся у здоровых
6. обучение алгоритма классификации
7. оценивание качества классификации
8. применение алгоритма классификации для диагностики

Дискретизация и векторизация ЭКГ-сигнала

Дискретизация ЭКГ-сигнала:

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

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

Векторизация кодограммы ЭКГ-сигнала:

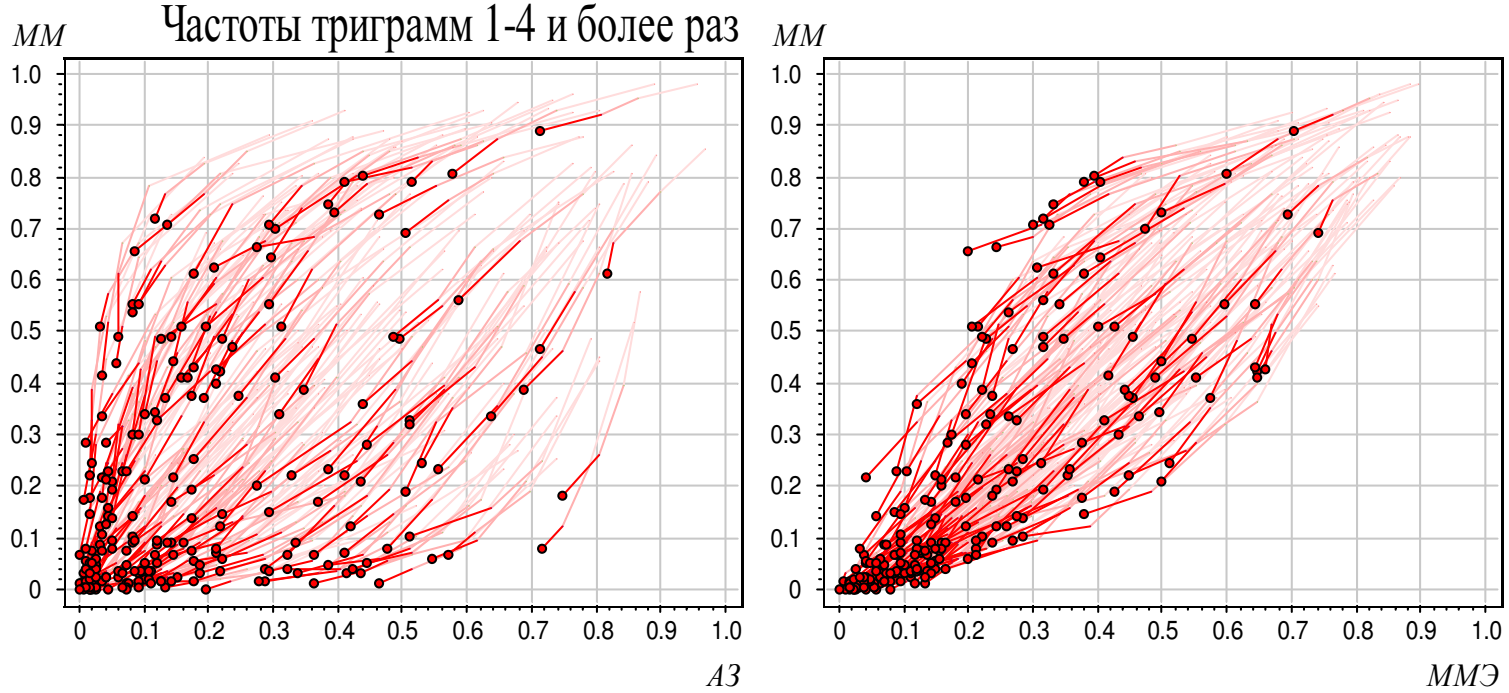
Вход: кодограмма S;
Выход: вектор частот триграмм w, размерности $|\mathcal{A}|^3 = 216$

Векторизация кодограммы ЭКГ-сигнала

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Отбор информативных признаков-триграмм

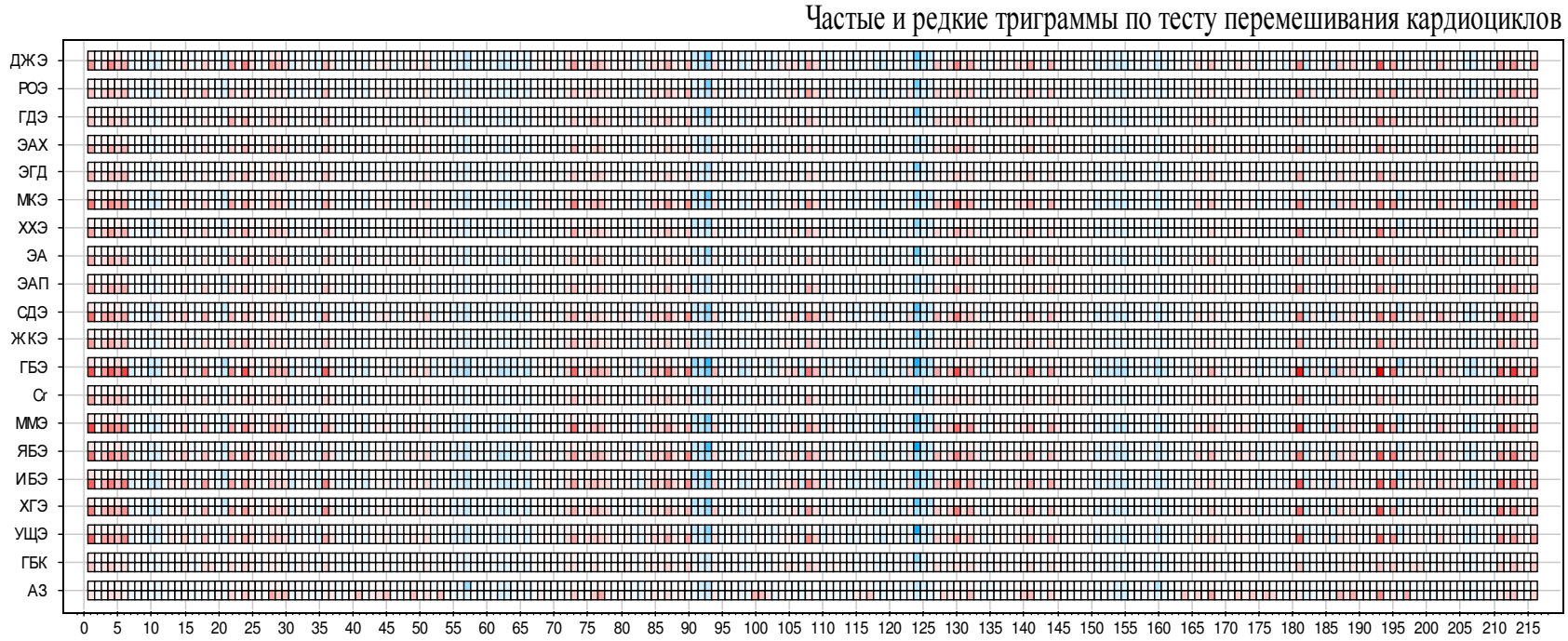
Слева: триграммы в осях «доля здоровых» — «доля больных».
Справа: триграммы в осях «доля больных» — «доля больных».



Вывод: болезнь имеет *диагностический эталон* — множество триграмм, часто встречающихся в кодограммах больных, и редко встречающихся в кодограммах здоровых людей.

Неслучайность триграмм. Перестановочные тесты

Нулевая гипотеза:
наблюдаемая частота триграммы реализовалась в результате случайной независимой выборки кардиоциклов

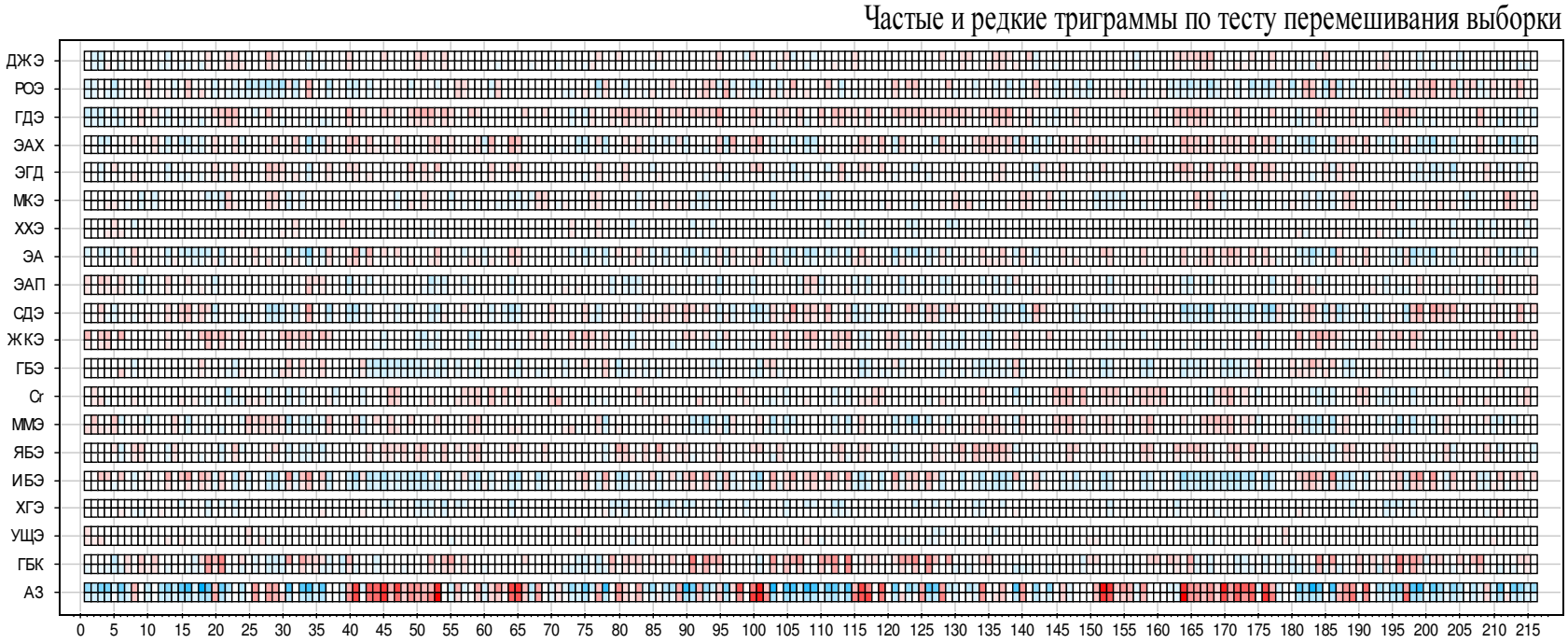


красным — неслучайно частые триграммы

синим — неслучайно редкие триграммы

Неслучайность триграмм. Перестановочные тесты

Нулевая гипотеза:
наблюдаемая частота триграммы реализовалась в результате случайной независимой выборки людей



красным — неслучайно частые триграммы

синим — неслучайно редкие триграммы

Неслучайность триграмм. Перестановочные тесты

Совмещение двух предыдущих тестов.

красным — неслучайно частые триграммы

синим — неслучайно редкие триграммы

Результаты

Болезнь	ГБК	УЩ	ХГ	ИБ	ЯБ	ММ	Cr
Число записей	327	750	698	1262	779	779	267
Лучший метод	SA1	RLR	LR0	LR1	SA3	LR1	RLR
AUC (контроль)	99	96	95	98	95	93	94
Чувствительность$_1$, %	95	95	95	95	95	95	95
Специфичность$_1$, %	96	83	72	91	63	59	81
Чувствительность$_2$, %	96	90	88	94	88	87	87
Специфичность$_2$, %	95	90	88	94	88	87	87

Болезнь	ГБ	ЖК	СД	АП	ЭА	ХХ	МК
Число записей	1891	277	868	257	259	336	649
Лучший метод	LR1	LR0	LR1	LR1	LR0	SA3	SA3
AUC (контроль), %	97	99	97	97	90	95	95
Чувствительность$_1$, %	95	95	95	95	95	95	95
Специфичность$_1$, %	85	94	86	72	47	75	69
Чувствительность$_2$, %	91	95	92	91	81	90	89
Специфичность$_2$, %	91	94	92	91	81	90	89

Задача

Дано:
матрица «объекты–признаки» по одной болезни
(некроз 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

Терминология диагностики

Положительный диагноз — алгоритм предсказывает болезнь
(хотя, казалось бы, что тут положительного...)

Чувствительность

— доля больных с верным положительным диагнозом.
Доля ошибок 2-го рода $= 1 -$ чувствительность.

Специфичность

— доля здоровых с верным отрицательным диагнозом.
Доля ошибок 1-го рода $= 1 -$ специфичность.

Чувствительность и специфичность надо максимизировать.

⊕ Они не зависят от соотношения мощностей классов.

⊕ Хорошо подходят для несбалансированных выборок.

Определение ROC-кривой

Модель классификации: $a(x_i, w, w_0) = \text{sign}\big(f(x_i, w) - w_0\big)$.

ROC — «receiver operating characteristic»,
каждая точка кривой соответствует некоторому w_0.

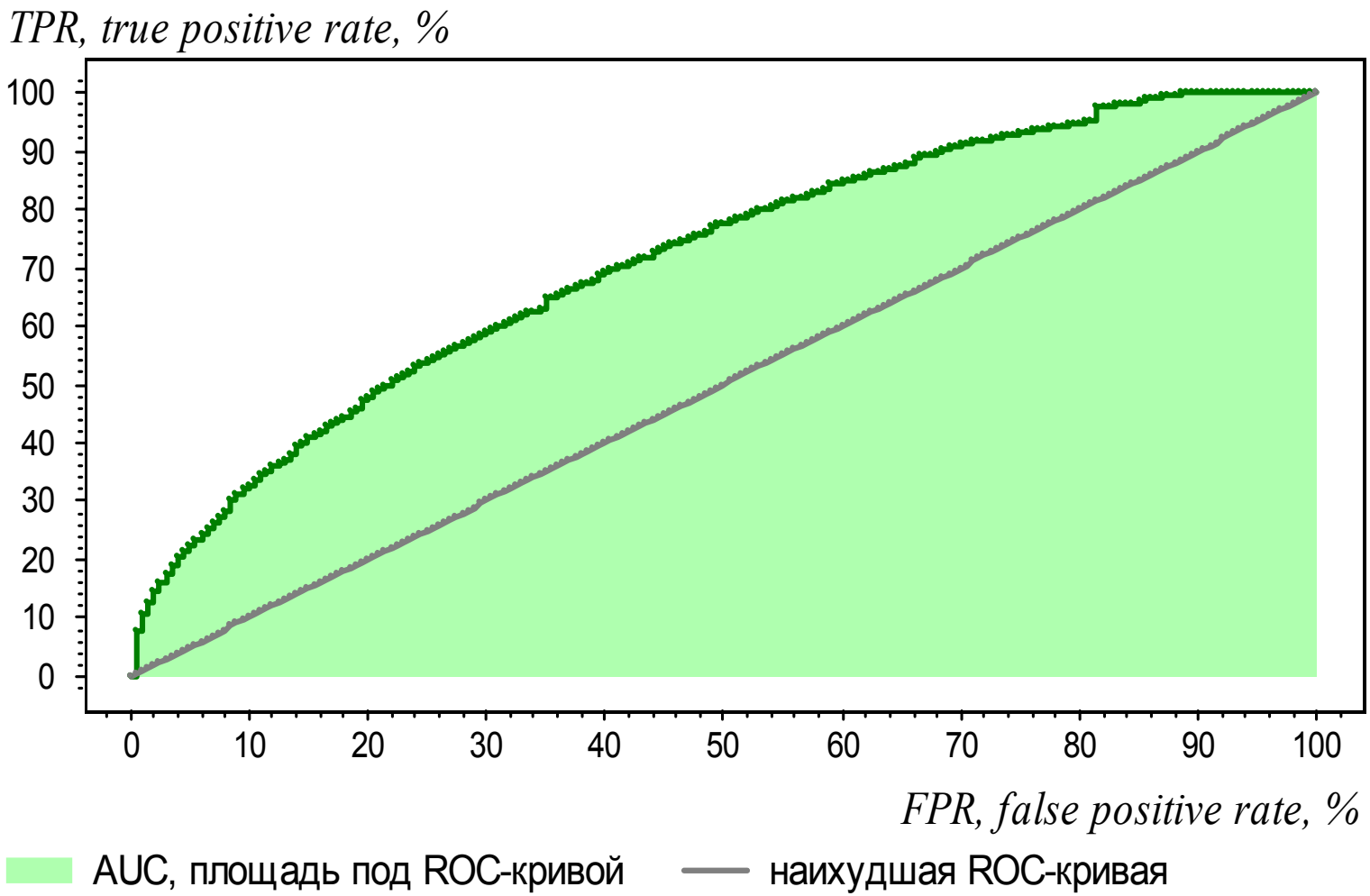
- по оси X: доля *ложно-положительных классификаций*
  FPR — false positive rate, *специфичность* $= 1 - \text{FPR}$:

$$\text{FPR} = \frac{\sum_{i=1}^{\ell}[y_i = -1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell}[y_i = -1]};$$

- по оси Y: доля *верно-положительных классификаций*
  TPR — true positive rate, *чувствительность* $= \text{TPR}$:

$$\text{TPR} = \frac{\sum_{i=1}^{\ell}[y_i = +1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell}[y_i = +1]};$$

Пример ROC-кривой



Вычисление AUC — площади под ROC-кривой

Модель классификации: $a(x_i, w, w_0) = \operatorname{sign}\big(f(x_i, w) - w_0\big)$,
$f(x, w)$ — дискриминантная функция.

AUC — это доля правильно упорядоченных пар (x_i, x_j):

$$\text{AUC} = \frac{1}{\ell_- \ell_+} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \big[y_i < y_j\big] \big[f(x_i, w) < f(x_j, w)\big] \to \max_w .$$

Кстати, явная максимизация аппроксимированного AUC:

$$Q(w) = \sum_{i,j:\ y_i < y_j} \mathscr{L}\big(\underbrace{f(x_j, w) - f(x_i, w)}_{M_{ij}(w)}\big) \to \min_w,$$

$\mathscr{L}(M)$ — убывающая функция отступа,
$M_{ij}(w)$ — новое понятие отступа для пар объектов.

И, кстати, это используют в задачах ранжирования.

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

Вход: выборка X^ℓ; дискриминантная функция $f(x, w)$;

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

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

Подсказки

Чем простым можно решать эту задачу:

- простые эвристики для отбора признаков
- нелинейные монотонные преобразования признаков
- наивный байесовский классификатор
- метод ближайшего соседа с жадным добавлением признаков
- готовые линейные классификаторы: SVM, LR, RLR,...

Чем ещё решали эту задачу:

- поиск синдромных закономерностей
- деревья решений
- бустинг над деревьями решений
- нейронная сеть

Переоценка ценностей

В задачах машинного обучения 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 :)

Результаты соревнования (хронология)

0.957234685	27 июня 17:52	кусни-Соболь
0.974485111	27 июня 19:11	кусни-Соболь
0.979994753	28 июня 13:38	кусни-Мясников
0.991276400	28 июня 15:38	Резков-2
0.984651712	28 июня 15:38	Резков-3
0.977174341	28 июня 15:38	Резков-4
0.960514233	28 июня 17:04	штопор-Аникин
0.989374262	28 июня 17:07	штопор-Старченко
0.984782894	28 июня 17:15	кусни-Бояров
0.950741178	28 июня 17:34	ми3-АлексАркадий
0.970680834	28 июня 17:34	ми3-Коноваленко-1
0.990948445	28 июня 17:34	ми3-Коноваленко-2

Результаты соревнования (лидерборд)

0.991276400	28 июня 15:38	Резков-2
0.990948445	28 июня 17:34	ми3-Коноваленко-2
0.989374262	28 июня 17:07	штопор-Старченко
0.984782894	28 июня 17:15	Кусни-Бояров
0.984651712	28 июня 15:38	Резков-3
0.979994753	28 июня 13:38	кусни-Мясников
0.977174341	28 июня 15:38	Резков-4
0.974485111	27 июня 19:11	кусни-Соболь
0.970680834	28 июня 17:34	ми3-Коноваленко-1
0.960514233	28 июня 17:04	штопор-Аникин
0.957234685	27 июня 17:52	кусни-Соболь
0.950741178	28 июня 17:34	ми3-АлексАркадий

Лидер соревнования — Илья Резков

- Диагностическими признаками класса (больных или здоровых) считаются триграммы, встречающиеся более n раз более чем в m кодограммах обучающей выборки данного класса, где $m = 0.8 \cdot m_{\max}$, $m_{\max}$ — максимальное число объектов, у которых есть триграмма, встретившаяся более n раз.
- В алгоритме Резков-2 $n = 2$, в Резков-3 $n = 3$, в Резков-4 $n = 4$.
- Признаки здоровья исключаются из признаков болезни.
- Для классификация болезни суммируются триграммы, попавшие в признаки болезни (результат записывается в sum_1).
- Результирующее число, определяющее принадлежность болезни

$$sum_{\text{бол.}} = \begin{cases} sum_1, & \text{если } sum_1 > S \\ 0.01 \cdot sum_1, & \text{если } sum_1 \leqslant S \end{cases}$$

- В алгоритме Резков-2 $S = 11$, в Резков-3 $S = 6$, в Резков-4 $S = 1$.
- Параметры n, 0.8, S подбираются перебором.

Лаборатория «МиМиМизации», Иван Коноваленко-2

- Признаки $X \in \mathbb{R}^{216}$ и класс $C \in \{0, 1\}$ рассматриваются как пара зависимых случайных величин. Выборка есть набор $x_{i,c,k}$, где i — номер признака, c — класс, $k \in \{1, 2, \ldots, n_c\}$, n_c — размер выборки для класса c. Строим нормальные ядерные оценки маргинальных плотностей распределения каждого признака с шириной ядра h:

$$\hat{f}_{X_i}(t|C=c) = \frac{1}{n_c} \sum_{k=1}^{n_c} \frac{1}{\sqrt{2\pi}h} e^{-\frac{(t-x_{i,c,k})^2}{2h^2}}.$$

- Строим наивный байесовский классификатор, а именно рассчитываем для новой реализации x^* вероятности классов по теореме Байеса, считая признаки независимыми:

$$\hat{\mathbb{P}}(C=c|X=x^*) = \frac{\mathbb{P}(C=c)\prod_{i=1}^{216}\hat{f}_{X_i}(x_i^*|C=c)}{\sum_{l=0}^{1}\mathbb{P}(C=l)\prod_{i=1}^{216}\hat{f}_{X_i}(x_i^*|C=l)}.$$

- Итоговая оценка класса берётся по порогу, b — свободный параметр:

$$\hat{C}(x^*) = \left[\hat{\mathbb{P}}(C=1|X=x^*) > b\right].$$

- Параметр h оптимизирует AUC по кроссвалидации leave-one-out на обучающей выборке.

Лаборатория «Штопор», А.Е.Старченко

- Использован метод LASSO SVM (1-norm SVM)
- Параметр регуляризации $\mu = 0.15$
- Функция эмпирического риска в этом случае выпуклая
- Поэтому применён пакет оптимизации CVX [1,2]
- Обучение проводилось по всей обучающей выборке один раз

[1] CVX Research, Inc. CVX: Matlab software for disciplined convex programming, version 2.0. http://cvxr.com/cvx, April 2011.

[2] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar), V. Blondel, S. Boyd, and H. Kimura, editors, pages 95–110, Lecture Notes in Control and Information Sciences, Springer, 2008. http://stanford.edu/~boyd/graph_dcp.html.

Лаборатория «Кусни», Андрей Бояров

1. $\bar{\mu}_j^y$ — среднее значение j-го признака по объектам класса y;
2. Отбор подмножества признаков $J \subset \{1, \ldots, n\}$, значимо отличающих больных от здоровых:
   - признаки, отличающие больных: $\bar{\mu}_j^1 > a$ и $\bar{\mu}_j^0 < b$ для всех $j \in J$;
   - признаки, отличающие здоровых: $\bar{\mu}_j^0 > a$ и $\bar{\mu}_j^1 < b$ для всех $j \in J$;
3. В качестве алгоритма классификации использовался SVM, параметр регуляризации $C = 1$, в качестве ядра использовалось RBF-ядро;
4. С помощью 10-fold cross validation были определены параметры: $a = 5.5$, $b = 0.7$;

В итоге было отобрано только 6 признаков!

Лаборатория «Кусни», Дмитрий Мясников

Решение основано на кусочно-линейной аппроксимации функции потерь $\mathscr{L}(M) = (1 - M)_+$ и RFM-регуляризаторе $\tilde{Q}$. Ответы $y \in [0; 1]$ для обучающей выборки приведены к $[-1; 1]$, а сама выборка разбита на две равные части: для поиска решения $X^{(1)}$ и оптимизации параметров $X^{(2)}$.

Алгоритм решения

1. Найти $w_0^k, w^k = \arg\min_{w_0, w} \tilde{Q}(X^{(1)}, y, w, w_0, C^{k-1}, \mu^{k-1})$
2. Найти параметры $C^k, \mu^k = \arg\max_{C,\mu} AUC(X^{(2)}, w^k)$
3. Если $|AUC^k - AUC^{k-1}| < \varepsilon_{AUC}$, завершить выполнение.
   Иначе увеличить k на 1 и повторить шаги 1, 3.

Для оптимизации использован алгоритм Нелдера-Мида.
Градиент $\tilde{Q}(w, w_0)$ можно получить аналитически,

поэтому на шаге 1 решение уточнено с помощью метода L-BFGS.

Лаборатория «Кусни», Виталий Соболь

Отбор признаков

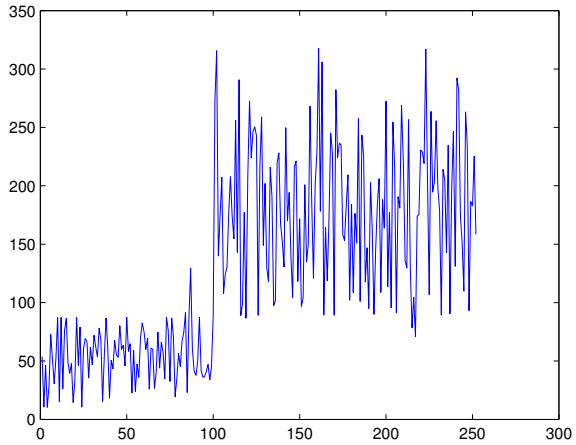
n_i^j — частота j-й триграммы в i-й кодограмме;
n_+^j — медиана частот j-й триграммы по классу +1 (больных);
n_-^j — медиана частот j-й триграммы по классу -1.
В качестве значимых берём первые 20 признаков (триграмм), упорядоченных по убыванию модуля $|n_+^j - n_-^j|$.
Выбор медианы обусловлен ее устойчивостью к выбросам.

Обучение модели

Метод опорных векторов SVM:

$$\sum_{i=1}^{\ell} \big(1 - y_i(\langle x_i, w\rangle - w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w_0, w},$$

$w \in \mathbb{R}^{20}$, $C = 0.25$,
на обучающей выборке AUC $= 0.9941$.



Лаборатория «Кусни», Андрей Бояров (вдогонку, 2014-07-02)

1. Применение анализа главных компонент (PCA) для выделения признаков, значимо отличающих больных от здоровых:
   - $(U_1, U_2, \ldots, U_n)$ — главные компоненты для тренировочных данных x_i, $i = 1, \ldots, \ell$;
   - выбирается m первых главных компонент, на которые строятся проекции тренировочных и тестовых данных соответственно: $\mathbf{Z}$ и $\tilde{\mathbf{Z}}$ — $\ell \times m$ матрицы;
   - алгоритм классификации обучается на $\mathbf{Z}$ и применяется к $\tilde{\mathbf{Z}}$;
2. В качестве алгоритма классификации использовался SVM, параметр регуляризации $C = 1$,
   в качестве ядра использовалось RBF-ядро;
3. С помощью метода 10-fold cross validation было определено количество главных компонент: $m = 30$;

Результат на тесте: AUC = 0.99193231

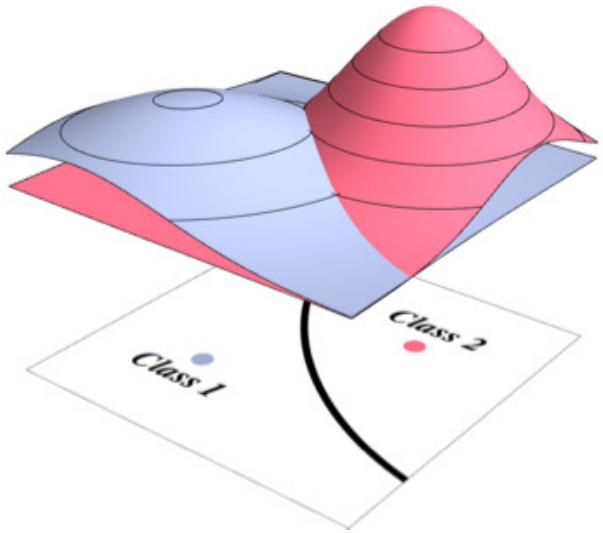
Байесовский классификатор

Принцип максимума апостериорной вероятности:

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

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

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



Наивный байесовский классификатор

Наивно предположим, что признаки статистически независимы:

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

Если признаки бинарные, $x^j \in \{0, 1\}$, то МП-оценка:

$$p(x^j = v|y) = \frac{\sum_{i=1}^{\ell} [y_i = y][x_i^j = v]}{\sum_{i=1}^{\ell} [y_i = y]}$$

Тогда наивный байесовский классификатор является линейным:

$$a(x) = \text{sign}\left(\log \frac{p(x|+1)}{p(x|-1)} + w_0\right) = \text{sign}\left(\sum_{j=1}^{n} x^j w_j + w_0\right),$$

$$w_j = \log \frac{p(x^j = 1|+1)}{p(x^j = 1|-1)} \frac{p(x^j = 0|-1)}{p(x^j = 0|+1)}$$

Преобразование и отбор признаков

Исходные целочисленные признаки:

n_i^j — сколько раз j-я триграмма встретилась в i-й кодограмме

Бинаризованные признаки:

$x_i^j = [n_i^j \geqslant 2]$

2 нашли перебором по обучающей выборке.

Отбор признаков:

w_j — веса признаков в наивном байесовском классификаторе

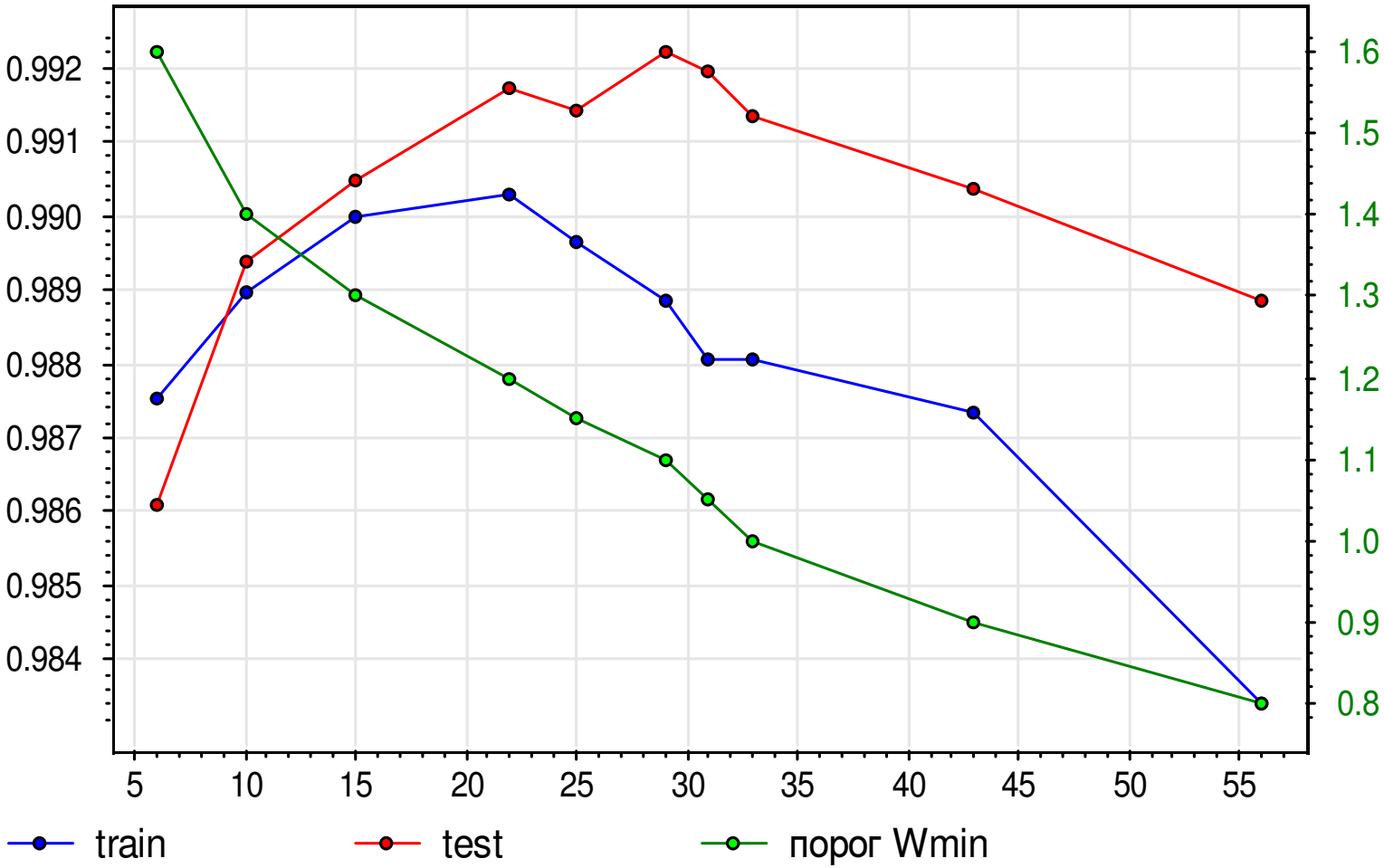
если $|w_j| < w_{\min}$, то $w_j := 0$

порог $w_{\min} = 1.2$ нашли перебором по обучающей выборке

в результате осталось 22 признака

Оптимизация порога $w_{\min}$

Зависимость AUC на обучении и тесте от числа признаков



Выводы

- Простые задачи хорошо решаются простыми методами :)
- Наивный байесовский классификатор не переобучается
- Гипотеза: его эффективная сложность равна не числу признаков, а 1 (есть над чем подумать теоретикам)
- Качество на тесте получилось даже лучше, чем на обучении (возможно, из-за конкретного разбиения малой выборки)
- В общепринятой методике $t \times q$-fold cross-validation выборку много раз разбивают на обучение и контроль, результат усредняют и оценивают доверительные интервалы

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov