Совместный выбор объектов и признаков при построении моделей в задачах банковского скоринга

Адуенко Александр

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель:
к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

13 июня 2013 года

Цели исследования

Цель исследования: создать метод совместного выбора признаков, объектов и моделей при построении моделей банковского кредитного скоринга.
Задача: построить алгоритм классификации объектов, который

- Отбирает информативные признаки
- Фильтрует выбросы
- Определяет число моделей, описывающих данные
- Определяет параметры этих моделей.

Обозначения

Регрессионная модель — отображение

$$f : \mathcal{X} \times \mathcal{W} \to \mathcal{Y}.$$

Функция регрессии — сужение функции

$$f\,|_{\mathbf{w}\in\mathcal{W}} : \mathcal{X} \to \mathcal{Y}.$$

Данные $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}.$$

Матрица плана $\mathbf{X} \in \mathbb{R}^{m\times n}$

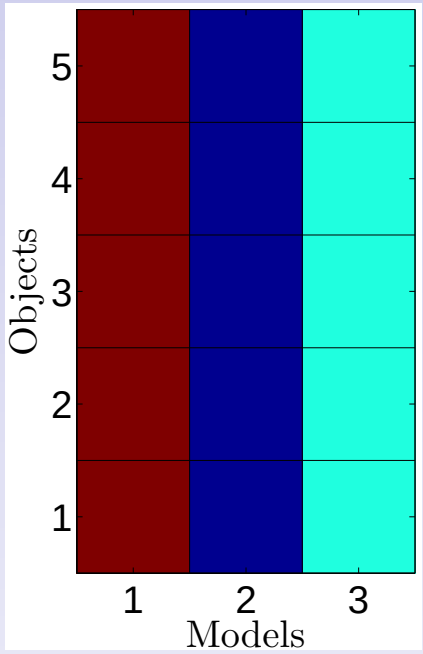
$$\mathbf{X} = (\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_m^\mathsf{T}).$$

Смеси моделей и многоуровневые модели

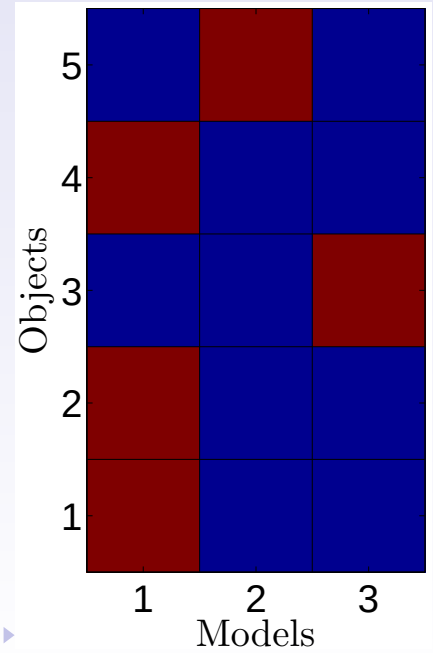
Смесь регрессионных моделей — регрессионная модель вида

$$f = \sum_{k=1}^{K} \pi_k f_k(\mathbf{w}_k), \text{ где}$$

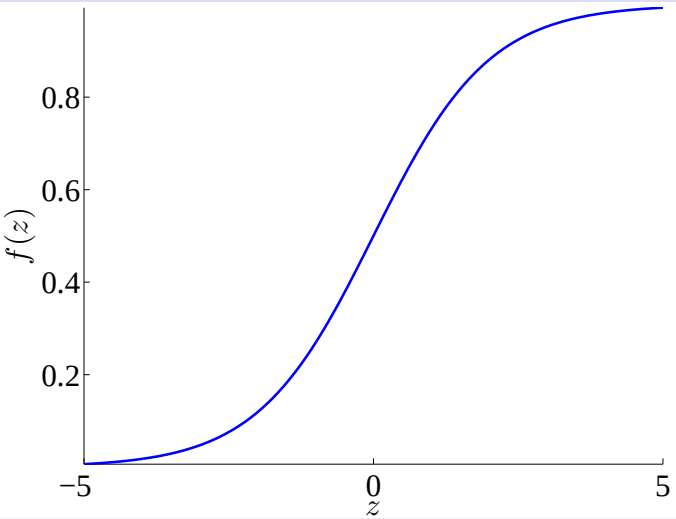
$$\sum_{k=1}^{K} \pi_k = 1, \ \pi_k \geq 0.$$



Многоуровневая регрессионная модель — набор регрессионных моделей f_k, $k = 1, \ldots, K$ такой, что при разбиении множества индексов объектов $\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{I}_k$ для всех объектов с индексами из $\mathcal{I}_k$ используется модель f_k.



Гипотеза порождения данных и модель



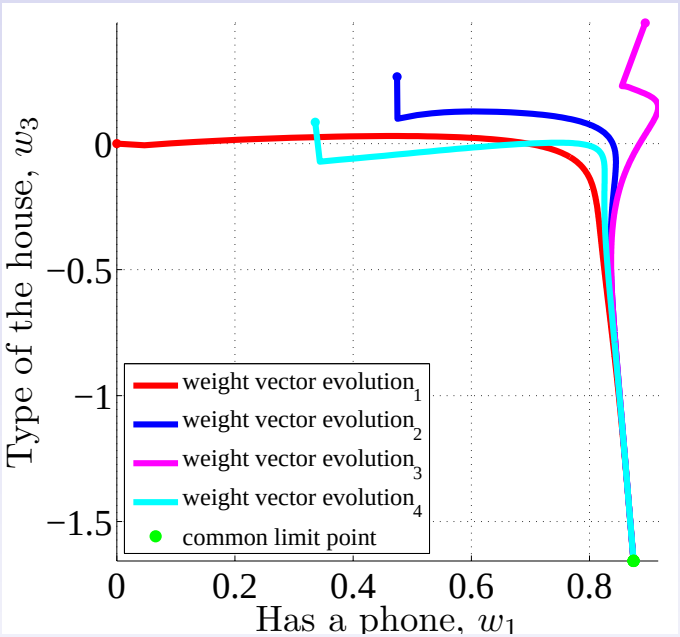
Базовые предположения

$$Y \sim Be(p),$$
$$p = f(\mathbf{x}^\mathsf{T}\mathbf{w}),$$
$$f(z) = \frac{1}{1+\exp(-z)}.$$

Следствие

$$\mathsf{E}(Y_i) = p(\mathbf{x}) = f,$$
$$\mathsf{D}(Y_i) = p(\mathbf{x})(1-p(\mathbf{x})) = f(1-f).$$

Оценка параметров модели



$$L(\mathbf{w}) = \prod_{i=1}^{m} f_i^{y_i}(1 - f_i)^{1-y_i}.$$

Функция штрафа

$$l(\mathbf{w}) = -\ln L(\mathbf{w}) =$$

$$= -\sum_{i=1}^{m} (y_i \ln f_i + (1 - y_i) \ln (1 - f_i)).$$

Итеративная оценка параметров

$$\mathbf{w}_j = \mathbf{w}_{j-1} - \mathbf{H}^{-1}(\mathbf{w}_{j-1})\nabla l(\mathbf{w}_{j-1}).$$

Оценка распределения параметров модели

Формула Ньютона-Рафсона для логистической регрессии

$$\mathbf{w}_j = \mathbf{w}_{j-1} - \left(\mathbf{X}^\mathsf{T}\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}(\mathbf{f}-\mathbf{y}),$$
$$\nabla l(\mathbf{w}) = \mathbf{X}^\mathsf{T}(\mathbf{f}-\mathbf{y}),\ \mathbf{H} = \mathbf{X}^\mathsf{T}\mathbf{R}\mathbf{X},$$
$$\mathbf{R} = \operatorname{diag}(\{f_i(1-f_i)\}_{i=1}^m) = \begin{pmatrix} \mathsf{D}(Y_1) & & 0 \\ & \ddots & \\ 0 & & \mathsf{D}(Y_m) \end{pmatrix}.$$

Теорема

Гессиан функции ошибки положительно определен, а потому минимум единственный.

Доказательство:

$$\mathbf{u}^\mathsf{T}\mathbf{H}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{R}\mathbf{X}\mathbf{u} = (\mathbf{X}\mathbf{u})^\mathsf{T}\mathbf{R}(\mathbf{X}\mathbf{u}) > 0.$$

Оценка параметров модели

Априорное распределение параметров

$\mathbf{w} \in p(\mathbf{w}|f, \alpha)$.

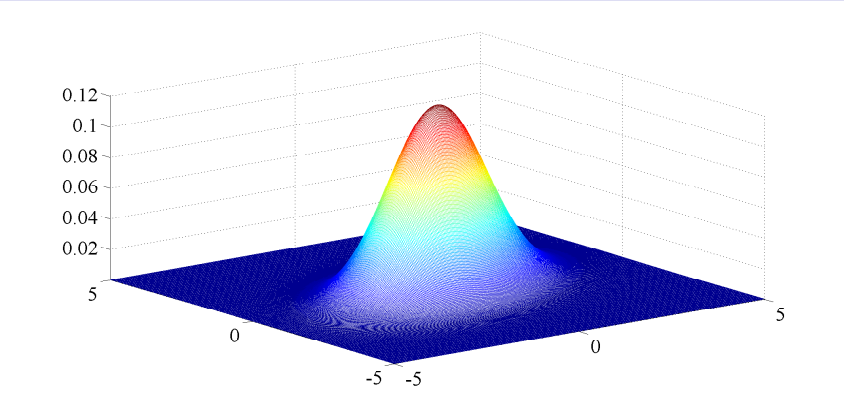
Апостериорное распределение параметров

$$p(\hat{\mathbf{w}}|D, f, \alpha) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, f)p(\mathbf{w}|f, \alpha)}{p(\mathbf{y}|\mathbf{X}, f, \alpha)}.$$

$$\frac{p(\mathbf{w}|D, f, \alpha)}{p(\hat{\mathbf{w}}|D, f, \alpha)} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, f)}{p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}, f)} \frac{p(\mathbf{w}|f, \alpha)}{p(\hat{\mathbf{w}}|f, \alpha)} \approx$$

$$\approx e^{-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^{\mathsf{T}}\mathbf{H}(\mathbf{w}-\hat{\mathbf{w}})} \Longrightarrow \hat{\mathbf{w}} \sim N(\mathbf{w}_0, \mathbf{H}^{-1}).$$

Модификация метода Белсли



$$\mathbf{H} = \mathbf{X}^{\mathsf{T}}\mathbf{R}\mathbf{X} = \tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}},$$
$$\tilde{\mathbf{X}} = \tilde{\mathbf{R}}\mathbf{X},\ \tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}.$$
$$\mathbf{var}(\mathbf{w}) = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{\mathsf{T}}.$$
$$\mathrm{var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2} =$$
$$= (q_{i1} + q_{i2} + \ldots + q_{in}) \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2},$$

Алгоритм отбора признаков, основанный на модификации метода Белсли

$\mathcal{A}$ – текущее множество признаков в модели.

Add

Для каждого $j \in \mathcal{J} \setminus \mathcal{A}$

$$\hat{\mathbf{w}}_j = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|+1}} L(\mathbf{w}|\mathbf{X}(\mathcal{A}_j),\ \mathbf{y}),$$

$$l(j) = l(\hat{\mathbf{w}}_j|\mathbf{X}(\mathcal{A} \cup \{j\}),\ \mathbf{y}),$$

$$j^* = \arg\min_j l_j.$$

$$l_0 = l(\hat{\mathbf{w}}_0|\mathbf{X}(\mathcal{A}),\ \mathbf{y}), \text{ где}$$

$$\hat{\mathbf{w}}_0 = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{A}),\ \mathbf{y}).$$

Если $l_{j^*} - l_0 \leq Z_1 < 0$, то берем признак в модель.

Del

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right],$$

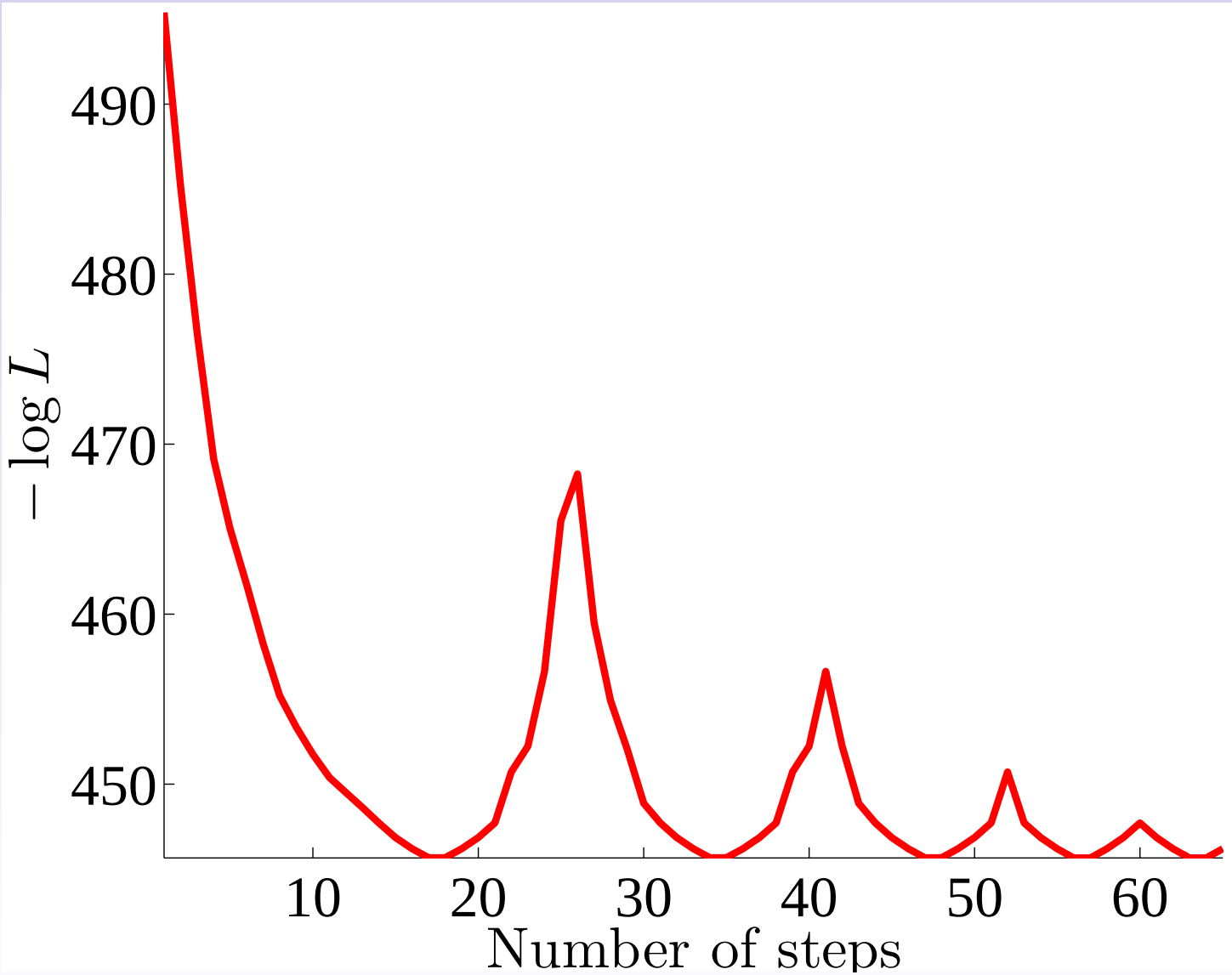
$$j^* = \arg\max_{j \in \mathcal{A}} \sum_{g=t-i^*+1}^{t} q_g^j,$$

$$\hat{\mathbf{w}}_{j^*} = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|-1}} L(\mathbf{w}|\mathbf{X}(\mathcal{A}_{j^*}),\ \mathbf{y}),$$

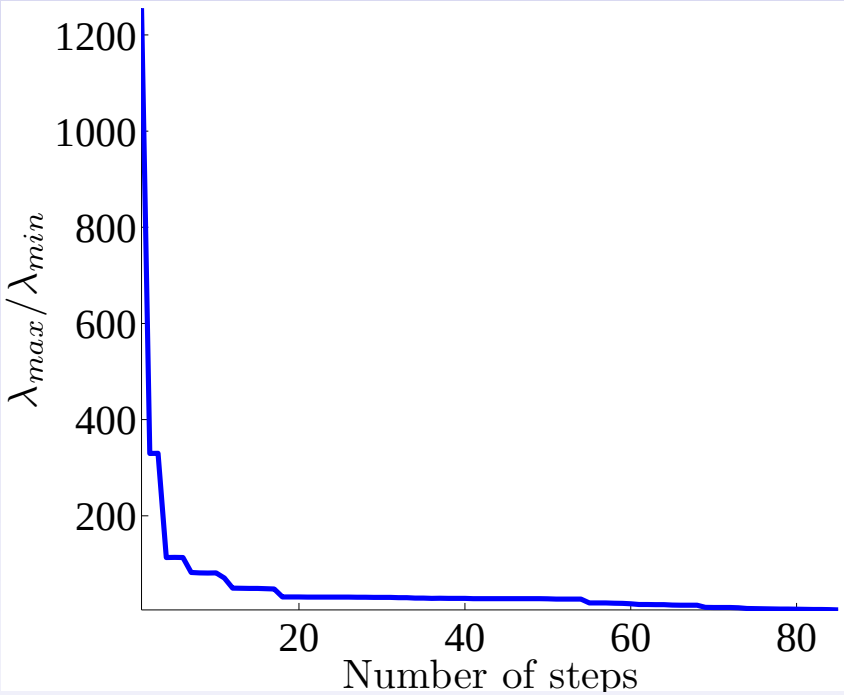
$$l_{j^*} = l(\hat{\mathbf{w}}_{j^*}|\mathbf{X}(\mathcal{A} \setminus \{j^*\}),\ \mathbf{y}).$$

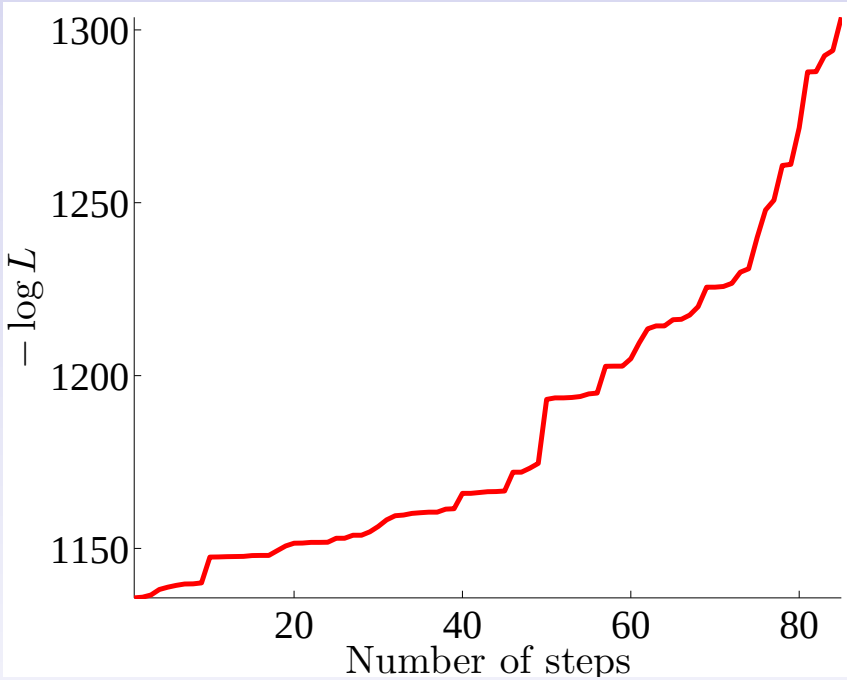
Если $l_{j^*} - l_0 \leq Z_2$, то признак j^* удаляется из модели.

Последовательный выбор признаков



Устойчивость и функция правдоподобия модели





Отбор объектов

Определение специфичности объекта

$\mathrm{Sp}(\mathbf{x}_i) = (\Delta_i \mathbf{w})^\intercal \mathbf{H} (\Delta_i \mathbf{w})$,
где $\Delta_i \mathbf{w} = \hat{\mathbf{w}}_i - \hat{\mathbf{w}}$,
$\Delta_i \mathbf{w} \sim N(\mathbf{0},\, \mathbf{H}^{-1})$.
$\mathrm{Sp}(\mathbf{x}_i) = (\Delta_i \mathbf{w})^\intercal \mathbf{H} (\Delta_i \mathbf{w}) \sim$
$\sim \chi^2(|\mathcal{A}|)$.

Модификация определения

$$\mathbf{w} \sim N\left(\mathbf{w}_0,\, \tau \mathbf{I}\right)$$

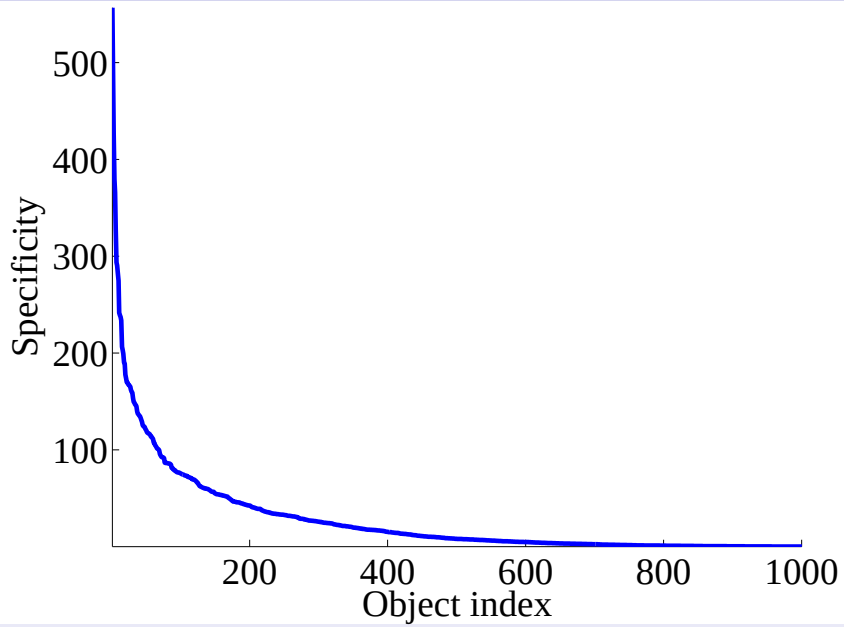
$$\hat{\mathbf{w}} \sim N\left(\mathbf{w}_0,\, \left(\mathbf{H} + \frac{1}{\tau}\mathbf{I}\right)^{-1}\right)$$

$$\mathrm{Sp}(\mathbf{x}_i) = (\Delta_i \mathbf{w})^\intercal \left(\mathbf{H} + \frac{1}{\tau}\mathbf{I}\right) (\Delta_i \mathbf{w}).$$

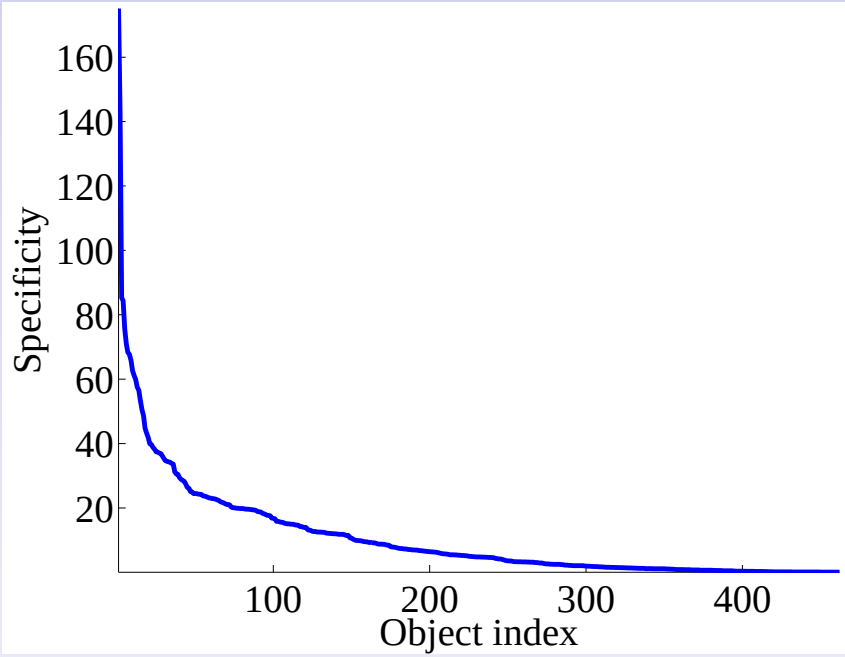
$\mathbf{H} \mapsto \mathrm{diag}(D_j)$, где

$$D_j = \frac{\sum_{i \in \mathcal{S}} (\Delta_i w_j)^2}{|\mathcal{S}| - 1}.$$

Распределение специфичности на выборке



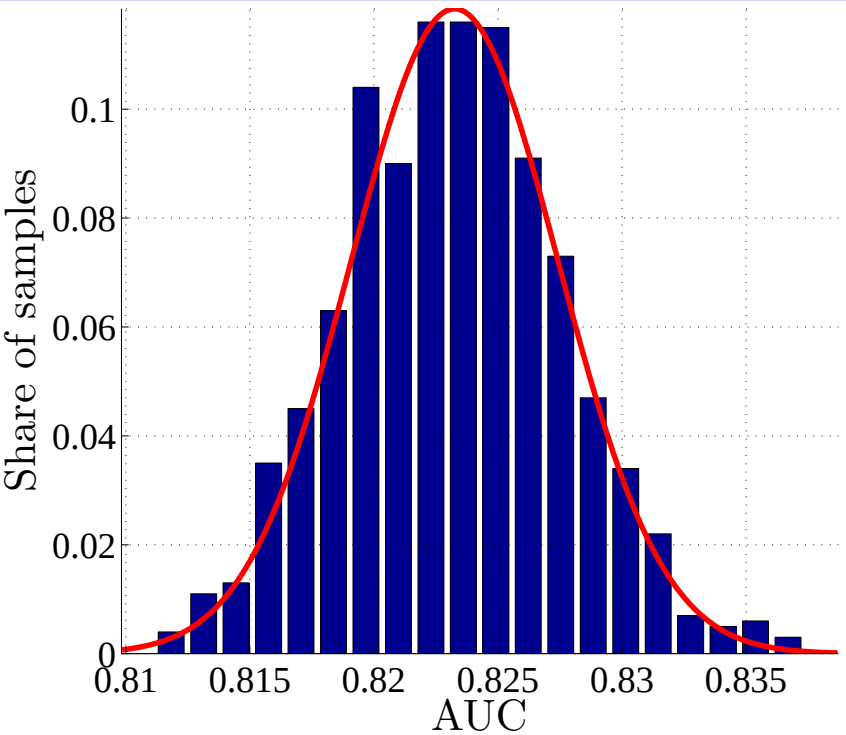
Потребительские кредиты



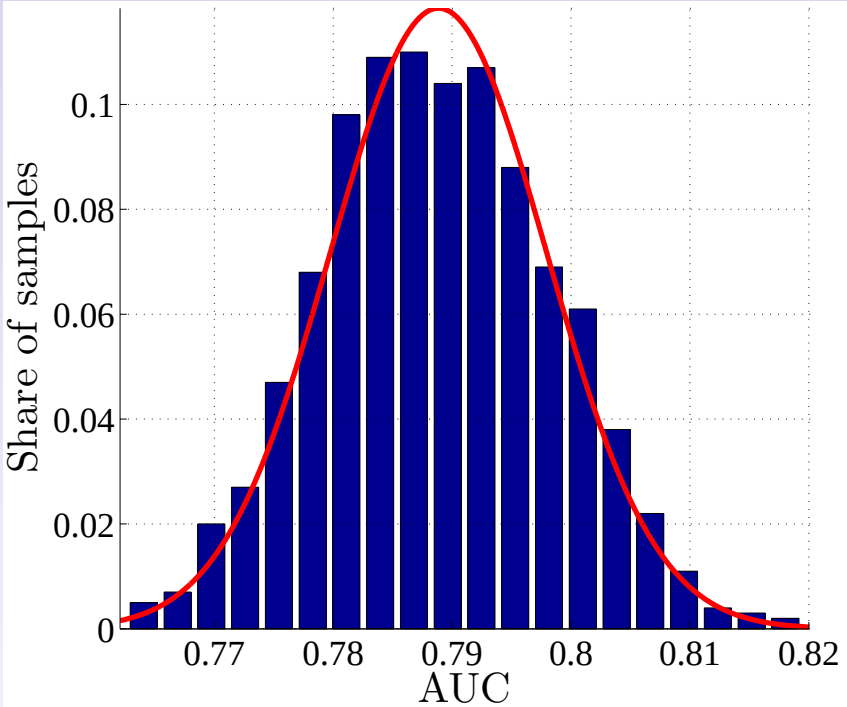
SAHD данные

Данные	AUC до	AUC после	Удалено
SAHD	0.7948	0.8275	15 из 462
Кредиты	0.8179	0.8779	50 из 1000

Проверка значимости улучшения качества



а) Обучение

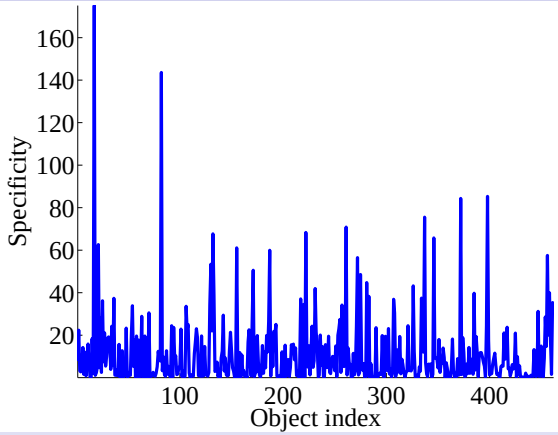


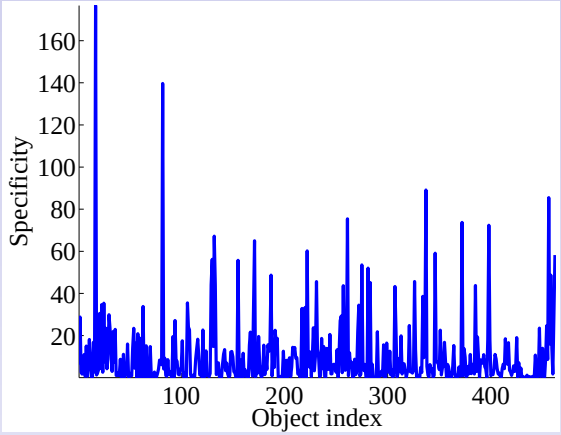
б) Контроль

Проверка значимости улучшения качества

Характеристики	Кредиты	SAHD
$\mathrm{AUC}_{\mathrm{learn}}$	0.8819	0.8308
$\mathrm{AUC}_{\mathrm{test}}$	0.8507	0.8093
p_{SW}	0.2655; 0.2364	0.2786; 0.7879
$\hat{m}$	0.8233; 0.7889	0.7994; 0.7722
$\hat{\sigma}^2$	$1.75 \cdot 10^{-5}$; $8.27 \cdot 10^{-5}$	$3.73 \cdot 10^{-5}$; $1.26 \cdot 10^{-4}$
$\hat{\sigma}$	0.0042; 0.0091	0.0061; 0.011
M	**14.0; 6.8**	**5.15; 3.32**
$\mathbf{p_0}$	$\mathbf{0}$; $\mathbf{5.3 \cdot 10^{-12}}$	$\mathbf{1.33 \cdot 10^{-7}}$; $\mathbf{4.55 \cdot 10^{-4}}$

Исследование различий в определении

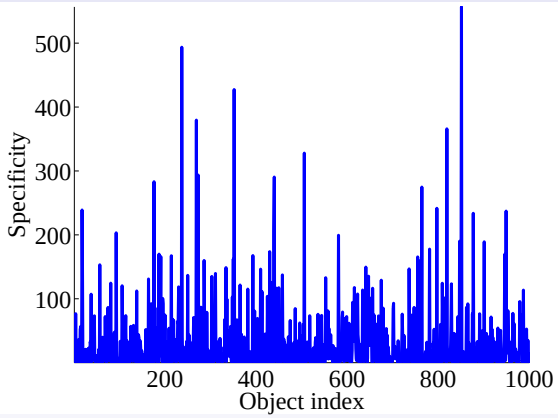


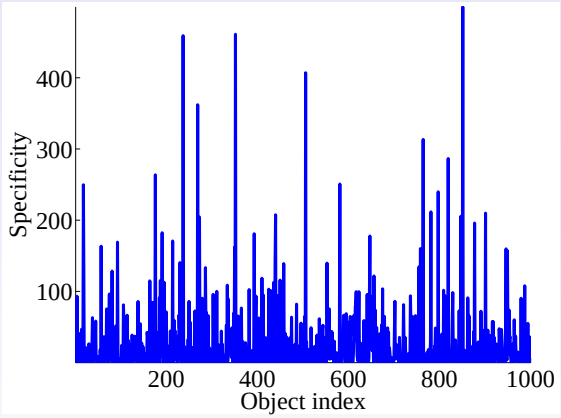


Корреляции

Пирсона	0.9736
Спирмена	0.9901
Кендалла	0.9132

SAHD данные





Корреляции

Пирсона	0.9794
Спирмена	0.9946
Кендалла	0.9377

Потребительские кредиты

EM-алгоритм для смеси логистических моделей

Правдоподобие данных для смеси моделей

$$L(\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}|\mathbf{X}, \mathbf{y}) =$$
$$= \prod_{i=1}^{m} \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i, \mathbf{w}_k)^{y_i} (1 - f(\mathbf{x}_i, \mathbf{w}_k))^{1-y_i} \right),$$
$$L(\mathbf{w}_1, \ldots, \mathbf{w}_K, , \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y}) =$$
$$= \prod_{i=1}^{m} \prod_{k=1}^{K} \{\pi_k f(\mathbf{x}_i, \mathbf{w}_k)^{y_i} (1 - f(\mathbf{x}_i, \mathbf{w}_k))^{1-y_i}\}^{z_{ik}}.$$

E-шаг

$$\gamma_{ik} = \mathbb{E}[z_{ik}] = p(k|\mathbf{x}_i, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}) =$$
$$= \frac{\pi_k f(\mathbf{x}_i, \mathbf{w}_k)^{y_i} (1 - f(\mathbf{x}_i, \mathbf{w}_k))^{1-y_i}}{\sum_{j=1}^{K} \pi_j f(\mathbf{x}_i, \mathbf{w}_j)^{y_i} (1 - f(\mathbf{x}_i, \mathbf{w}_j))^{1-y_i}},$$

M-шаг

$$\tilde{l}(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{y}) = \mathbb{E}_{\mathbf{Z}}[-\log L(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \mathbf{y})] =$$

$$= -\sum_{i=1}^{m}\sum_{k=1}^{K}\gamma_{ik}\{\log\pi_k + y_i\log(f(\mathbf{x}_i,\ \mathbf{w}_k)) + (1-y_i)\log(1-f(\mathbf{x}_i,\ \mathbf{w}_k))\}.$$

$$\pi_k = \frac{1}{m}\sum_{i=1}^{m}\gamma_{ik}.$$

$$\tilde{l}(\mathbf{w}_1,\ldots,\mathbf{w}_K,\boldsymbol{\pi}|\mathbf{X},\mathbf{y}) = -\sum_{k=1}^{K}\{\log\pi_k\sum_{i=1}^{m}\gamma_{ik}\} + \sum_{k=1}^{K}\tilde{l}_k(\mathbf{w}_k|\mathbf{X},\ \mathbf{y}).$$

$$\frac{\partial\tilde{l}_k}{\partial\mathbf{w}_k} = \mathbf{X}^\mathsf{T}\boldsymbol{\Gamma}_k(\mathbf{f}-\mathbf{y}),$$

$$\mathbf{H}_k = \mathbf{X}^\mathsf{T}\mathbf{R}_k\mathbf{X},$$

$$\mathbf{R}_k = \mathrm{diag}(\gamma_{ik}f(\mathbf{x}_i^\mathsf{T}\mathbf{w}_k)f(-\mathbf{x}_i^\mathsf{T}\mathbf{w}_k)).$$

Определение требуемого числа моделей

Начало алгоритма: $K = 1$, $\pi_1 = 1$, $\alpha > 1$ – некоторое заданное число.

$$\tilde{\mathcal{B}} = \left\{ i : \frac{p(\mathbf{w}_1, \ldots, \mathbf{w}_{K-1} | \mathbf{x}_i, y_i)}{\max_j p(\mathbf{w}_1, \ldots, \mathbf{w}_{K-1} | \mathbf{x}_j, y_j)} < \frac{1}{\alpha} \right\}.$$

Если $|\tilde{\mathcal{B}}| > m_0$, строим новую компоненту смеси: $\pi_K = \frac{|\tilde{\mathcal{B}}|}{m}$,

$\pi_j \to (1 - \pi_K)\pi_j$, $j \leq K - 1$,

$\mathbf{w}_K = \mathbf{0}$.

Процедура выбора модели

Решающее правило отнесения к модели на обучении

$$k_i^* = \arg\max_{k \in \{1..l\}} p(y_i \mid f_k, \mathbf{x}_i).$$

Осторожный выбор модели (на контроле)

$$k_i^* = \arg\max_{k \in \{1..l\}} \min_{u \in \{0,1\}} p(u \mid f_k, \boldsymbol{x}_i).$$

Для логистической регрессии

$$k_i^* = \arg\max_{k \in \{1..l\}} \sigma(-|\boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}_k|) =$$

$$= \arg\min_{k \in \{1..l\}} |\boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}_k|.$$

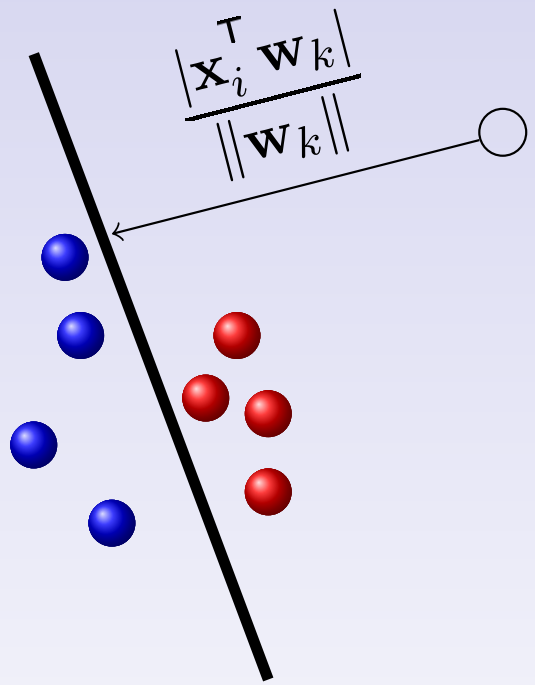
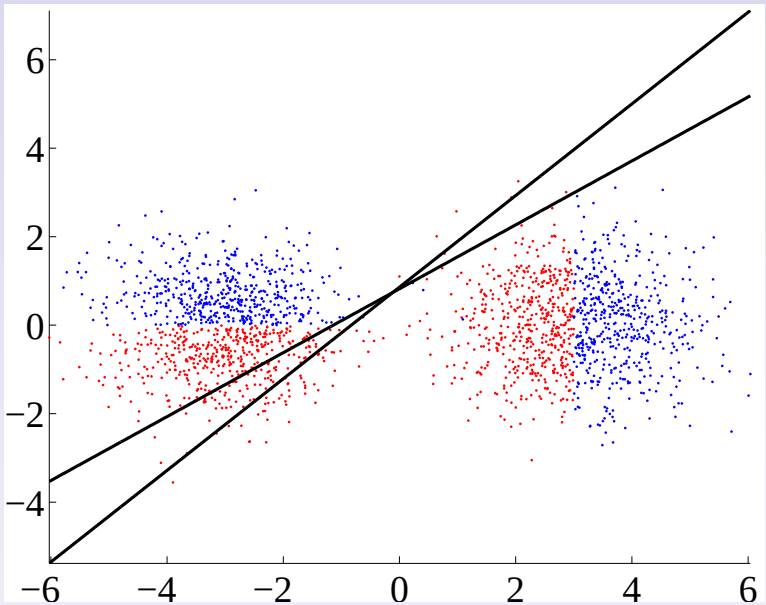
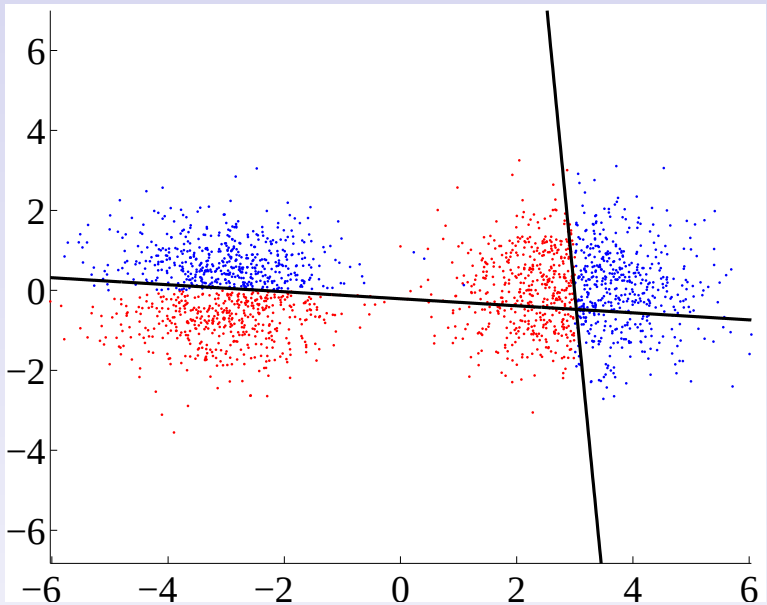


Рис. : Иллюстрация отнесения объекта контроля к модели.

Иллюстрация выбора модели

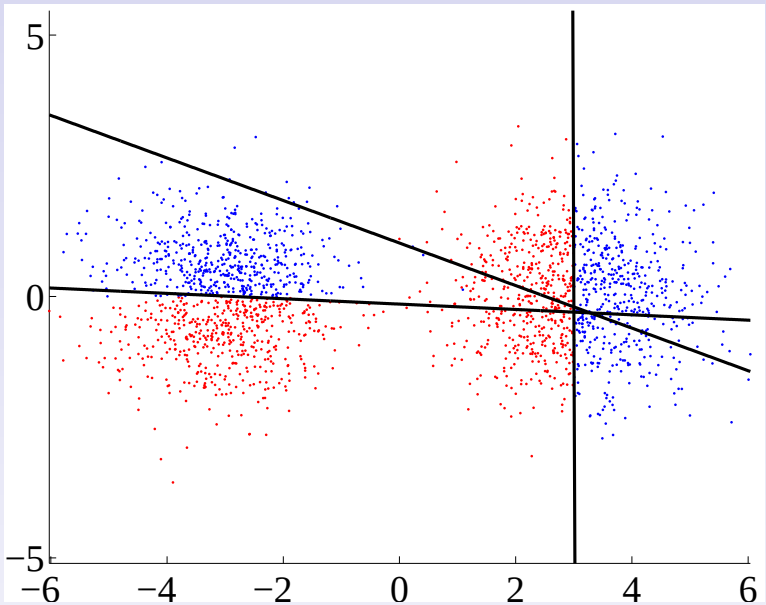


а) Старое решающее правило

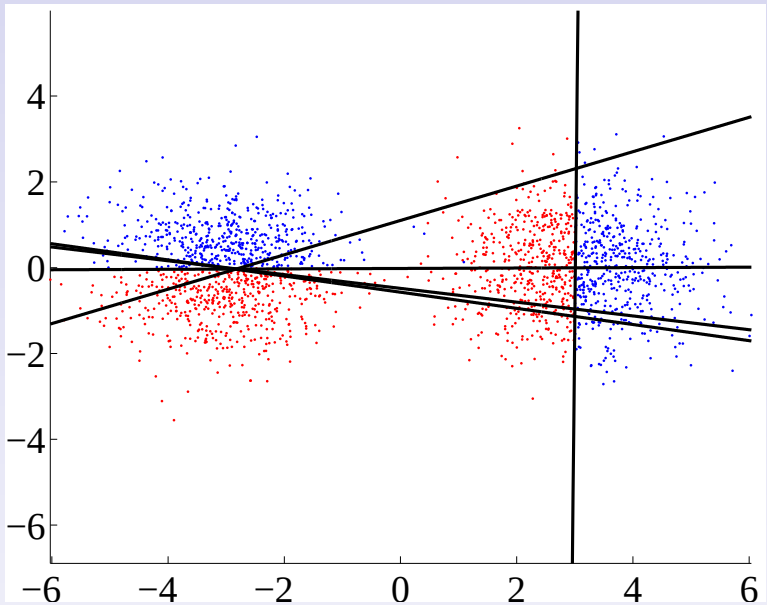


б) Осторожный выбор модели

Сравнение многоуровневых моделей и смеси моделей

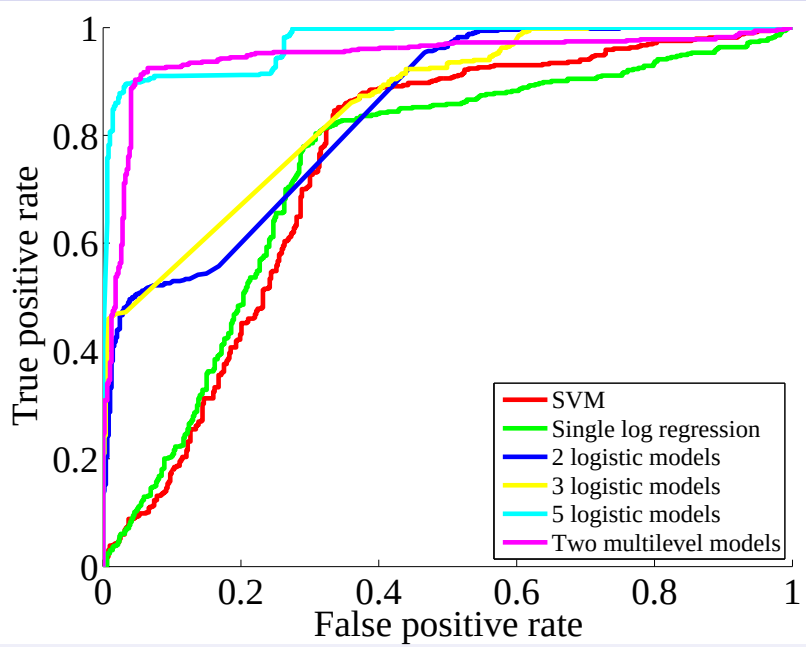


а) 3 многоуровневые модели



б) Смесь 5 моделей

Сравнение многоуровневых моделей и смеси моделей



ROC кривые

Таблица : Площадь под ROC-кривой для разных моделей.

Модель	AUC
1 лог. модель	0.7364
SVM	0.7462
Смесь 2 лог. моделей	0.8393
Смесь 3 лог. моделей	0.8529
2 многоур. модели	0.9460
Смесь 5 лог. моделей	0.9722
3 многоур. модели	0.9757

Данные Яндекса

97290 объектов. 245 признаков. Классы–$\{0, 1, 2, 3, 4\}$.

$$Q_2(\hat{y}) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} DCG_j, \text{ где } DCG_j = \sum_{i=1}^{|\Omega_j|} \frac{y_{\text{ind}(\mathbf{x}_i)}}{\log_2 i + 1}.$$

Предлагаемая модификация многоклассовой логистической регрессии: $\hat{y}_i = \sum_{k=1}^{K} C_k P(C_k|\mathbf{x}_i)$.

Таблица : Сравнение качества Q_2.

Алгоритм отбора	Число признаков	Q_2	$\hat{Q}_2$
Пошаговый	12	3.612	4.028
Генетический	18	3.639	4.058

Публикации по теме

1. Адуенко А. А. Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга // Машинное обучение и анализ данных, 2012. № 3. С. 279–291.
2. Адуенко А. А., Стрижов В. В. Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2013. № 2.
3. Адуенко А. А., Кузьмин А. А., Стрижов В. В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 2012. № 3. С. 119–131.
4. Адуенко А. А., Стрижов В. В. Алгоритм оптимального расположения названий коллекции документов // Программная инженерия, 2013. № 3. С. 21–25.
5. Иванова А. В., Адуенко А. А., Стрижов В. В. Алгоритм построения логических правил при разметке текстов // Программная инженерия, 2013. № 6.

Заключение

- Построен алгоритм отбора объектов и признаков.
- Отбор признаков основан на предложенной модификации метода Белсли и существенно повышает устойчивость построенных моделей.
- Отбор объектов и фильтрация выбросов основаны на введенной функции специфичности объекта и показано, что полученное повышение качества значимо.
- Предложено использовать осторожный выбор модели для объектов обучения в многоуровневых моделях, что значительно снижает переобучение.
- Предложена модификация алгоритма многоклассовой логистической регрессии для ранжирования объектов внутри класса, что позволило значительно улучшить качество прогноза релевантности на данных Яндекса.