Модели и методы интеллектуального анализа данных

Рудаков Константин Владимирович
Воронцов Константин Вячеславович

Международная научная конференция по информатике
и прикладной математике, посвященная 60-летию
Вычислительного центра им. А. А. Дородницына РАН

Москва, ВЦ РАН • 9–10 декабря 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 — ЭМК пациента; y — диагноз, лечение;
  - 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;

Задачи, некорректно поставленные по Адамару

Корректно поставленная задача:

- решение существует,
- решение единственно,
- решение устойчиво (непрерывно зависит от данных в некоторой разумной топологии).

Жак Саломон Адамар (1865–1963), почётный член АН СССР (1929)

Задачи восстановления зависимостей по эмпирическим данным — всегда некорректно поставленные.

Hadamard J. Sur les problèmes aux dérivées partielles et leur signification physique. 1902.

Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. 1974.

Обучение регрессии — это оптимизация

Задача регрессии, $a\colon \mathbb{X} \to \mathbb{R}$

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, $a\colon \mathbb{X} \to \{-1, +1\}$

1. Выбираем модель классификации, например, линейную:
$$a(x, w) = \text{sign}\langle x, w\rangle$$
2. Выбираем функцию потерь, например, число ошибок:
$$\mathscr{L}(a, y) = \big[ay < 0\big]$$
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:
$$Q(a, X^k) = \frac{1}{k}\sum_{i=1}^{k}\big[a(\tilde{x}_i, w)\tilde{y}_i < 0\big]$$

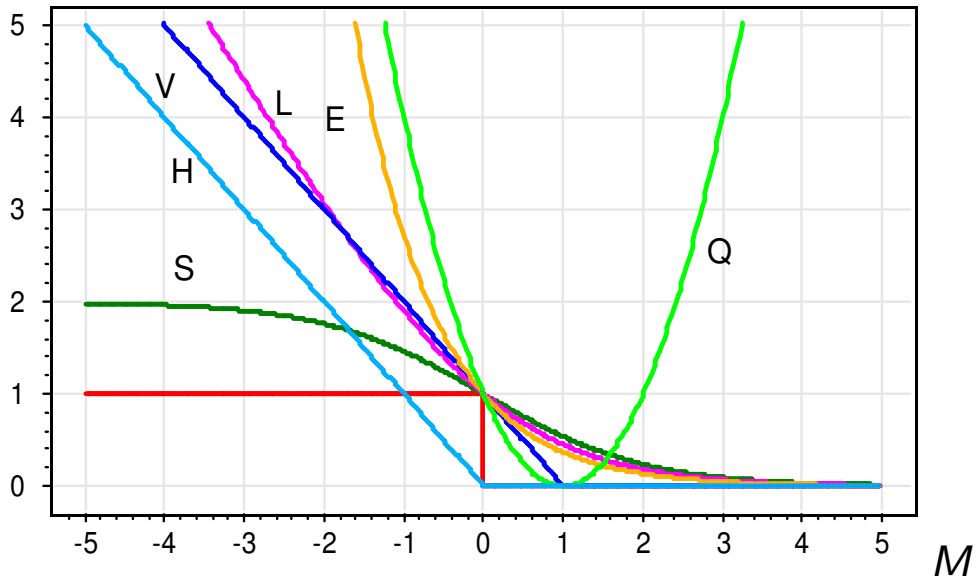
Минимизация аппроксимированного эмпирического риска

Задача классификации с двумя классами, $a\colon \mathbb{X} \to \{-1, +1\}$

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

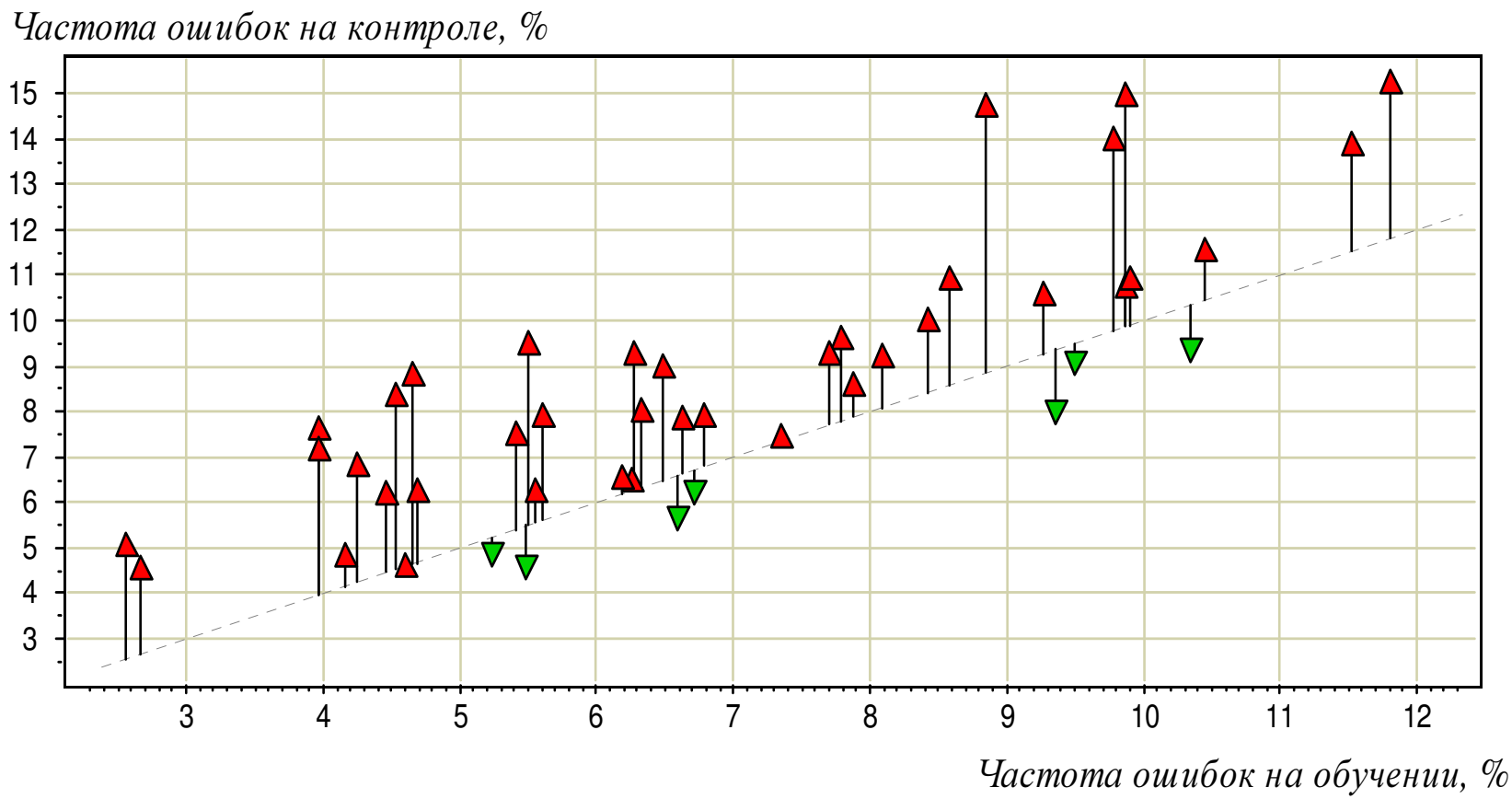
Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1 - M)_+$ — 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)

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

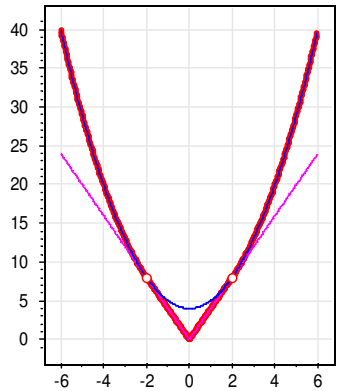
- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊕ Есть эффект группировки
- ⊖ Шумовые признаки также группируются вместе, и группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

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

Support Features Machine (SFM)

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

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



- 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.

Регуляризация нестационарной регрессии

Линейная модель регрессии:

$$a(x_t, w_t) = \langle x_t, w_t \rangle = \sum_{j=1}^{n} x_{tj} w_{tj}, \qquad x_t, w_t \in \mathbb{R}^n$$

где параметр модели $w_t \in \mathbb{R}^n$ зависит от времени t.

Метод наименьших квадратов:

$$Q(w, X^T) = \sum_{t=1}^{T} \big(a(x_t, w_t) - y_t\big)^2 \to \min_{\{w_{jt}\}}$$

Регуляризация:

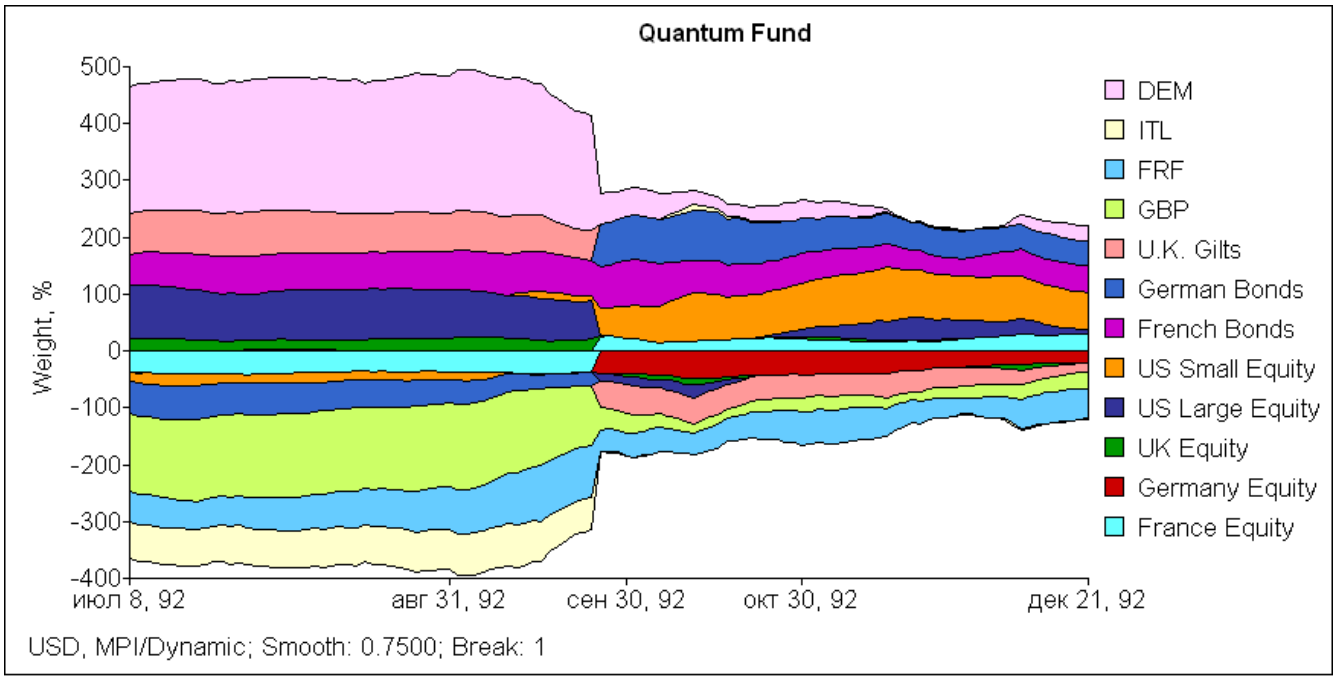
$$Q(w, X^T) + \mu \sum_{t=1}^{T} \sum_{j=1}^{n} w_{tj}^2 + \eta \sum_{t=1}^{T-1} \sum_{j=1}^{n} \big|w_{t+1,j} - w_{tj}\big| \;\to\; \min_{\{w_{jt}\}}$$

μ — параметр селективности;

η — параметр изменчивости коэффициентов во времени;

Пример. Оценивание динамики портфеля хедж-фонда

«Чёрная среда», 16-09-1992, фонд Quantum (Дж.Сорос): обнаружен момент реструктуризации портфеля (продажа GBP)



O. Krasotkina, A. Kopylov, V. Mottl, M. Markov. Bayesian estimation of time-varying regression with changing time-volatility for detection of hidden events in nonstationary signals // SPPRA, 2010.

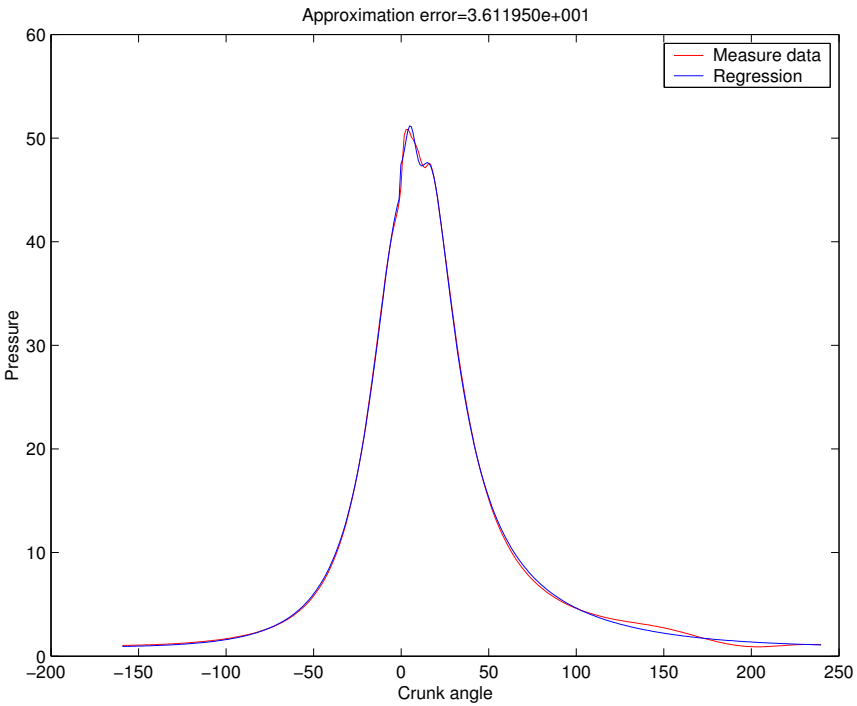
Процедура порождения и выбора моделей

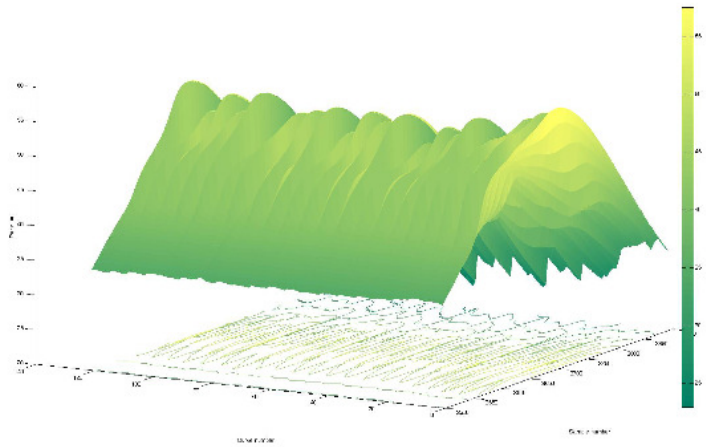
- Задаётся множество достаточно простых параметрических порождающих функций
- Метаэвристиками порождается множество суперпозиций
- Настраиваются параметры модели w;
  байесовский вывод приводит к регуляризатору $w^{\mathsf{T}}Aw$,
  где A^{-1} — ковариационная матрица.
- Выбирается модель с наименьшей длиной описания.

Реализация с открытым кодом:
MVR Composer, http://strijov.com

Стрижов В. В. Порождение и выбор моделей в задачах регрессии и классификации. Диссертация на соискание ученой степени д.ф.-м.н., ВЦ РАН, 2014.

Модель давления в камере сгорания дизельного двигателя





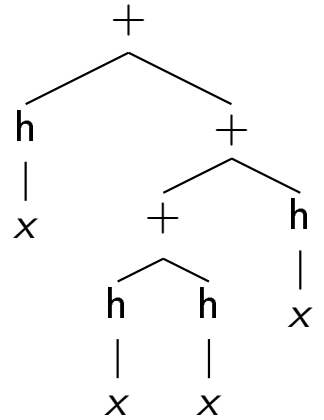
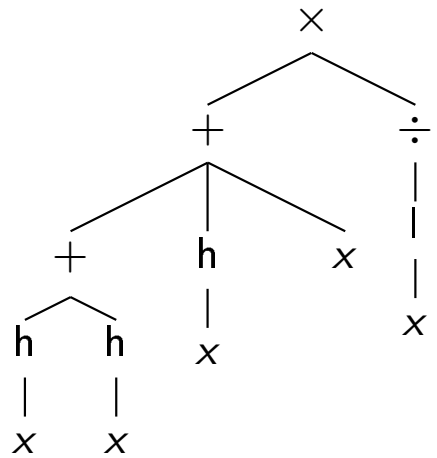
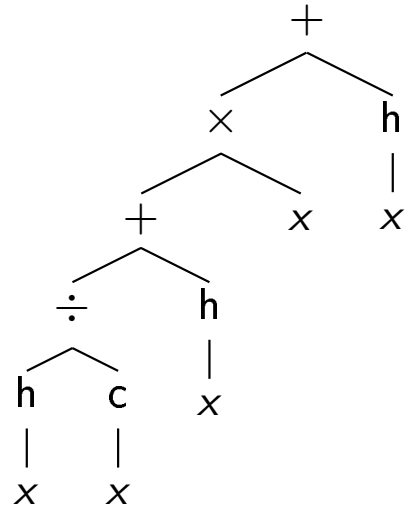
Давление в камере сгорания дизельного двигателя:

x — угол поворота коленчатого вала (нормализованный),

y — давление (нормализованное), выборка включает 4000 элементов.

Стрижов В.В. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010.

Порожденные и выбранные модели давления

Модель 1	Модель 2	Модель 3
		

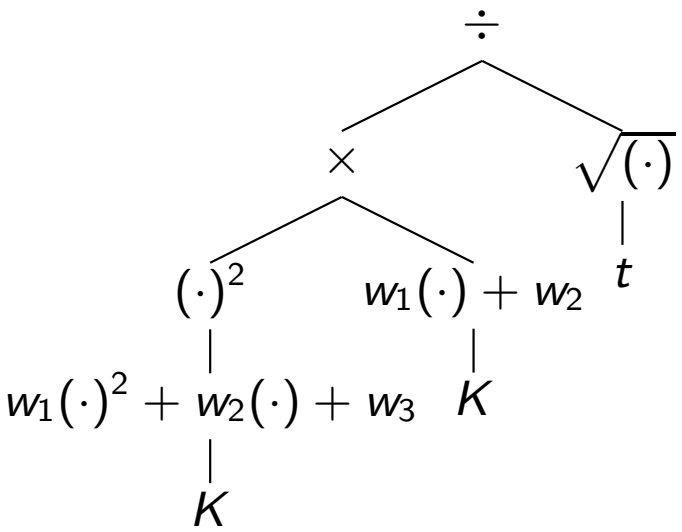
Обозначения: h — гауссова функция $y = \lambda \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\xi)^2 + a\right)$,
c — кубическая $y = ax^3 + bx^2 + cx + d$, | — линейная $y = ax + b$.

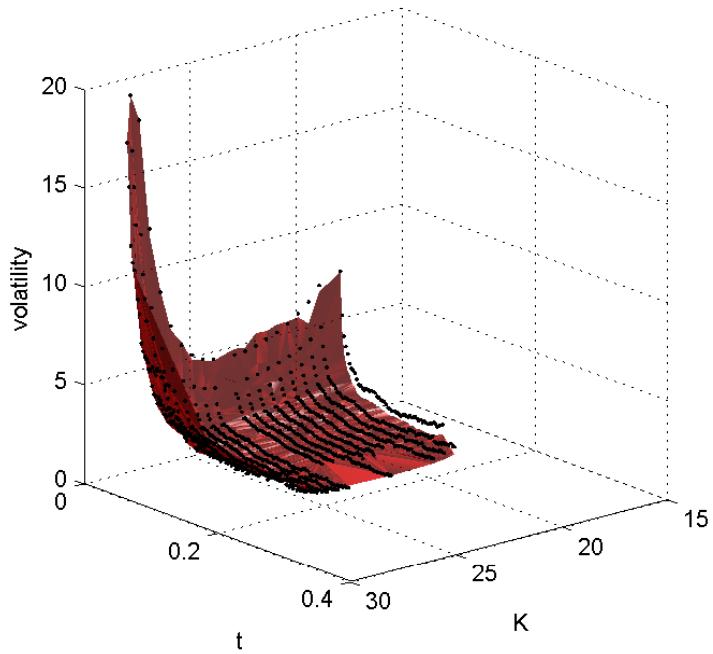
$$y = (ax+b)^{-1}\left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i\right).$$

Рудой Г.И., Стрижов В.В. Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных // Информатика и её применения, 2013, 7(1): 17-26.

Модель волатильности опционов

Модель волатильности европейского опциона на сырую нефть.
Волатильность σ зависит от времени t и цены исполнения K:
$\sigma = \frac{1}{\sqrt{t}}(w_1 K + w_2)(w_1 K^2 + w_2 K + w_3)^2$.





Сологуб Р. А. Алгоритмы индуктивного порождения и трансформации моделей в задачах нелинейной регрессии. Диссертация к.ф.-м.н. ВЦ РАН, 2014.

Задачи матричного разложения

Дано: матрица $Z = \|z_{ij}\|_{n\times m}$.

Найти: матрицы $X = \|x_{it}\|_{n\times k}$ и $Y = \|y_{tj}\|_{k\times m}$ такие, что

$$\|Z - XY\|_D = \sum_{i=1}^{n}\sum_{j=1}^{m} D\Big(z_{ij}, \sum_t x_{it}y_{tj}\Big) \to \min_{X,Y}$$

Типы задач:

- различные нормы D: Фробениуса, KL-дивергенция,...
- неотрицательные матричные разложения: $x_{it} \geqslant 0,\ \ y_{tj} \geqslant 0$
- стохастические матричные разложения: $\sum_i x_{it} = 1,\ \ \sum_t y_{tj} = 1$
- разреженные данные: известны только $z_{ij},\ \ (i,j) \in \Omega$:

$$\sum_{(i,j)\in\Omega} D\Big(z_{ij}, \sum_t x_{it}y_{tj}\Big) \to \min_{X,Y}$$

Примеры прикладных задач матричного разложения

1. **Анализ данных жидкостной хроматографии**

$$z(t,\lambda) = \sum_i X_i(t)\, Y_i(\lambda)$$

дано: $z(t,\lambda)$ — выход сканирующего УФ-детектора;
найти: $X_i(t)$ — хроматограмма i-го вещества, t — время
$Y_i(\lambda)$ — спектр i-го вещества, λ — длина волны.

2. **Анализ данных ДНК-микрочипов**

$$I(p,k) = \sum_g a_{pg}\, C_{gk}$$

дано: $I(p,k)$ — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
C_{gk} — концентрация g-го гена на k-м чипе.

Примеры прикладных задач матричного разложения

3. **Рекомендательные системы**

$$R_{iu} = \sum_t p_i(t) q_u(t)$$

дано: R_{iu} — рейтинги товаров i, поставленные пользователем u;
найти: $p_i(t)$ — профиль интересов товара i;
$q_u(t)$ — профиль интересов пользователя u.

4. **Тематическое моделирование текстовых коллекций**

$$p(w|d) = \sum_t p(w|t) p(t|d)$$

дано: $p(w|d) = \frac{n_{dw}}{n_d}$ — частоты слов w в документах d;
найти: $\phi_{wt} = p(w|t)$ — распределения слов w в темах t,
$\theta_{td} = p(t|d)$ — распределения тем t в документах d.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw}\right) \end{cases}$$

где $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

ARTM — Аддитивная регуляризация тематических моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

Комбинирование регуляризованных тематических моделей

Максимизация log правдоподобия с n регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \theta_{td}}\right) \end{cases}$$

Мультимодальная ARTM [Vorontsov et al, 2015]

W^m — словарь токенов m-й модальности, $m \in M$

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновая параллельная мультимодальная ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

Сообщество:

- Открытый код https://github.com/bigartm (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Эксперимент 1. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

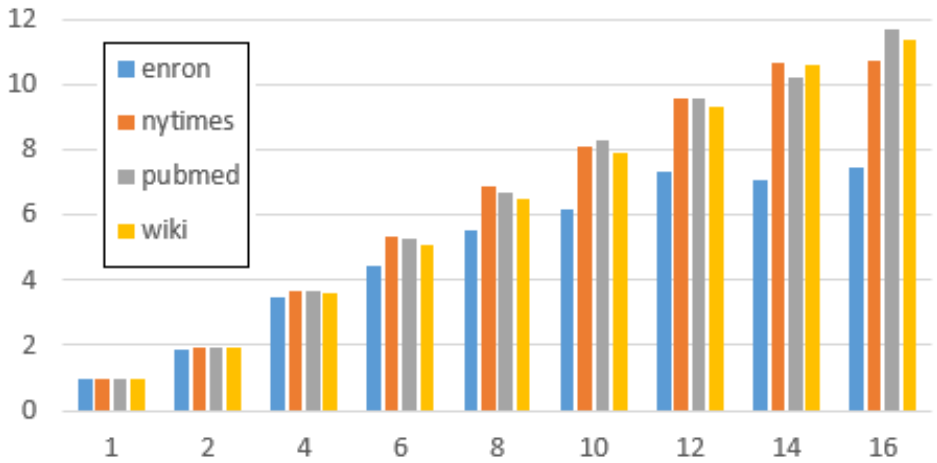
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* вычислена на тестовой выборке документов

Эксперимент 2. Масштабируемость по числу потоков

коллекция	$\|W\|$, 10^3	$\|D\|$, 10^6	n, 10^6	размер, Гб
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

ускорение



число ядер

Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Эксперимент 3. Выявление этничных тем в социальной сети

- Выделение этно-релевантных тем
  - регуляризатор частичного обучения по словарю этнонимов
  - альтернативный вариант: этнонимы как модальность
  - регуляризаторы для контрастирования малых тем
- Повышение качества тематической модели
  - использование модальности авторов
  - выявление тематических сообществ
  - лингвистическая регуляризация и выделение мультиграмм
- Оценивание распределения тем по регионам
  - разреживание региональных тем
  - сглаживание общих тем
- Оценивание распределения тем по времени
  - разреживание событийных тем
  - сглаживание перманентных тем

Примеры этничных тем

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,
(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,
(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,
(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция, операция, селение, сша, нусра, турция,
(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,
(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,
(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,
(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожать, сирия, подразделение, квартал, армейский,
(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этничных тем

(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,
(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,
(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,
(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,
(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,
(украинцы, немцы): украинский, упс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,
(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,
(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

Примеры этничных тем

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Несколько тезисов в заключение

- Регуляризации — инструмент решения некорректно поставленных задач восстановления зависимостей
- Big Data — субквадратичные алгоритмы
- Predictive Analytics — предсказательные модели
- Deep Learning — глубокие суперпозиции моделей
- Data Science — профессия будущего: число прикладных задач растёт во всех областях науки и бизнеса
- «Система Физтеха»: Физтех — база ВЦ РАН — компании

Воронцов Константин Вячеславович
voron@forecsys.ru
www.MachineLearning.ru/wiki?title=User:Vokov