Машина релевантных тегов

Молчанов Д. А.[1]
Кондрашкин Д. А.[2]
Ветров Д. П.[2,3]

[1] Московский Государственный Университет им. М. В. Ломоносова
[2] Национальный Исследовательский Университет Высшая Школа Экономики
[3] Сколковский институт науки и технологий

Математические методы распознавания образов,
19–25 сентября 2015

Машина релевантных векторов

Модель:

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{n=1}^{N} \sigma(t_n \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_n}))$$

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{d} \mathcal{N}(w_j|0, \alpha_j^{-1})$$

Обучение:

$$p(\boldsymbol{t}|\boldsymbol{\alpha}) = \int p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w} \to \max_{\boldsymbol{\alpha}}$$

ARD-эффект: такая процедура обучения осуществляет отбор признаков.
Но будет ли наблюдаться этот эффект на более сложных моделях с другим априорным распределением?

Постановка задачи

Рассмотрим задачу бинарной классификации данных с бинарными признаками.

- $(\boldsymbol{x_i}, t_i)_{i=1}^n$ — обучающая выборка
- $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{id})$ — объект
- $t_i \in \{0, 1\}$ — метка класса объекта $\boldsymbol{x_i}$
- $x_{ij} = 1 \Leftrightarrow \boldsymbol{x_i}$ помечен тегом j
- Все теги влияют на метку класса независимо

Вероятностная модель RTM (для одного объекта $\boldsymbol{x}$):

$$q_j = \mathrm{P}(t = 1 | x_j = 1) \qquad \mathrm{P}(t = 1 | \boldsymbol{x}, \boldsymbol{q}) = \frac{\prod\limits_{j=1}^{d} q_j^{x_j}}{\prod\limits_{j=1}^{d} q_j^{x_j} + \prod\limits_{j=1}^{d} (1 - q_j)^{x_j}}$$

Параметры модели — вектор $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)^T$

Априорное распределение

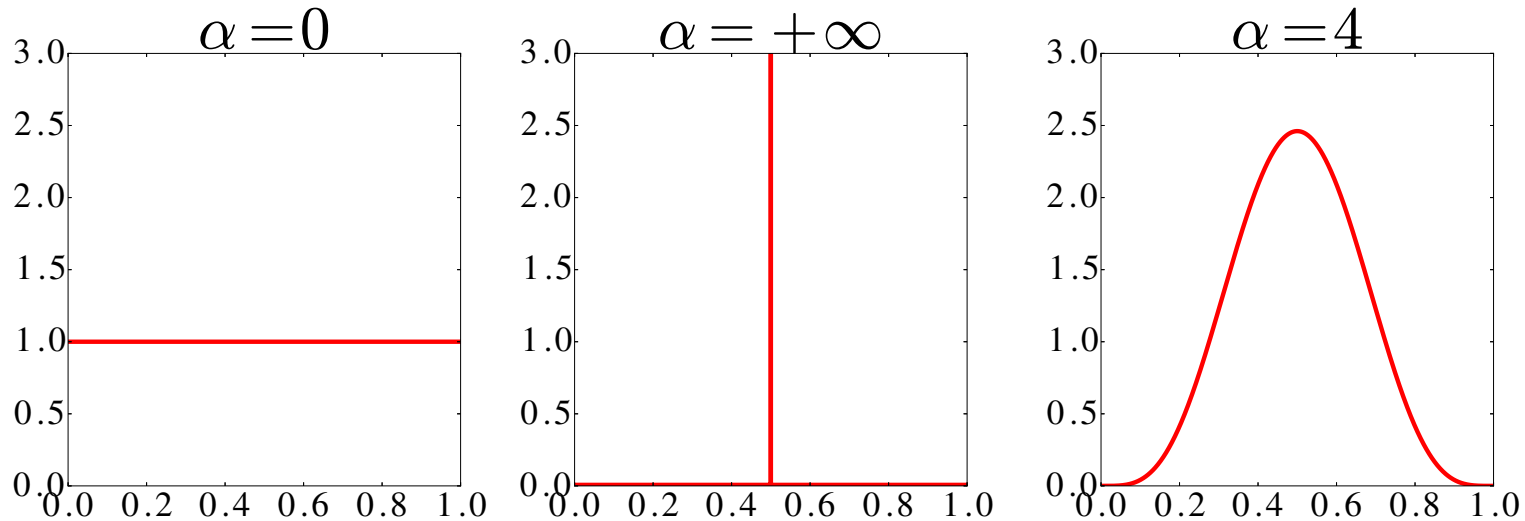
Симметричное бета-распределение:

$$q_j \sim \text{Beta}(\alpha_j + 1, \alpha_j + 1), \alpha_j \in [0, +\infty)$$

Вектор $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)^T$ называется вектором гиперпараметров.

- $\alpha_j = 0 \Rightarrow q_j^{MAP} = q_j^{ML}$
- $\alpha_j = +\infty \Rightarrow q_j = p(t = 1|x_j = 1) = 0.5 \Rightarrow$ тег j убирается из модели
- $\alpha_j \in (0, +\infty) \Rightarrow$ на q_j накладывается регуляризация

Графики $p(q|\alpha)$ при различных значениях α:



Обоснованность

Формула Байеса:

$$p(\boldsymbol{q}|\boldsymbol{X},\boldsymbol{t},\boldsymbol{\alpha}) = \frac{\mathrm{P}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{q})p(\boldsymbol{q}|\boldsymbol{\alpha})}{\int \mathrm{P}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{q})p(\boldsymbol{q}|\boldsymbol{\alpha})d\boldsymbol{q}}$$

- $p(\boldsymbol{q}|\boldsymbol{X},\boldsymbol{t},\boldsymbol{\alpha})$ — апостериорное распределение
- $\mathrm{P}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{q})$ — функция правдоподобия
- $p(\boldsymbol{q}|\boldsymbol{\alpha})$ — априорное распределение
- $E(\boldsymbol{\alpha}) = \int \mathrm{P}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{q})p(\boldsymbol{q}|\boldsymbol{\alpha})d\boldsymbol{q}$ — *обоснованность* (evidence)

Обоснованность может быть использована для выбора модели

Байесовский выбор модели

Максимизация обоснованности дает самую простую модель из тех, что хорошо объясняют данные.

В случае RTM:
Обоснованность $\to \max_{\alpha}$ $\Rightarrow$ $\alpha_j \to +\infty$ для нерелевантных тегов j $\Rightarrow$ нерелевантные теги автоматически убираются из модели.

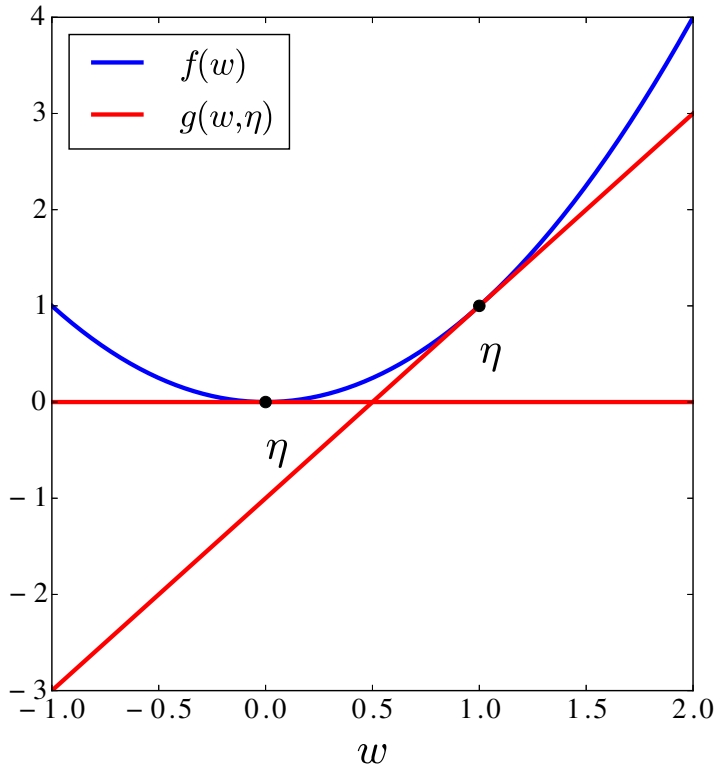
В случае модели RTM ни значение, ни градиент обоснованности не может быть вычислен, поэтому приходится оптимизировать ее приближенно.

Вариационные нижние оценки

Рассмотрим функцию $f(\boldsymbol{w}), \boldsymbol{w} \in M_1 \subseteq \mathbb{R}^m$
Вариационной нижней оценкой на функцию $f(\boldsymbol{w})$ называется функция $g(\boldsymbol{w}, \boldsymbol{\eta})$ такая, что:

$$g(\boldsymbol{w}, \boldsymbol{\eta}) \leq f(\boldsymbol{w}) \; \forall \boldsymbol{w}, \boldsymbol{\eta} \in M_1$$
$$g(\boldsymbol{w}, \boldsymbol{w}) = f(\boldsymbol{w}) \; \forall \boldsymbol{w} \in M_1$$

Пример:
$f(\boldsymbol{w})$ – выпуклая функция,
$g(\boldsymbol{w}, \boldsymbol{\eta})$ – касательная плоскость к функции $f(\boldsymbol{w})$, проведенная через точку касания $\boldsymbol{\eta}$.



Семейство оценок на обоснованность

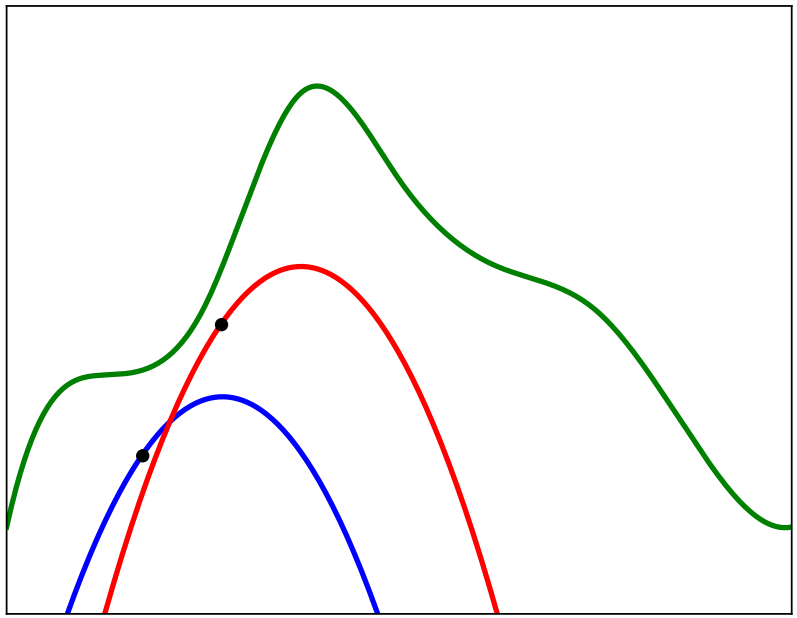
- Вариационная нижняя оценка на функцию правдоподобия одного объекта:
  $L_i(\boldsymbol{q}, \boldsymbol{\eta_i}) \leq p(t_i|\boldsymbol{x_i}, \boldsymbol{q})$
- Вариационные параметры:
  $\mathbf{H} = (\eta_{ij})_{i,j=1}^{n,d}$
- Нижняя оценка на подынтегральную функцию для обоснованности:
  $L(\boldsymbol{q}, \mathbf{H}, \boldsymbol{\alpha}) = p(\boldsymbol{q}|\boldsymbol{\alpha}) \prod_{i=1}^{n} L_i(\boldsymbol{q}, \boldsymbol{\eta_i}) \leq p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q}|\boldsymbol{\alpha})$
- Семейство нижних оценок на обоснованность:
  $\tilde{E}(\mathbf{H}, \boldsymbol{\alpha}) = \int L(\boldsymbol{q}, \mathbf{H}, \boldsymbol{\alpha}) d\boldsymbol{q} \leq E(\alpha)$

Максимизация обоснованности

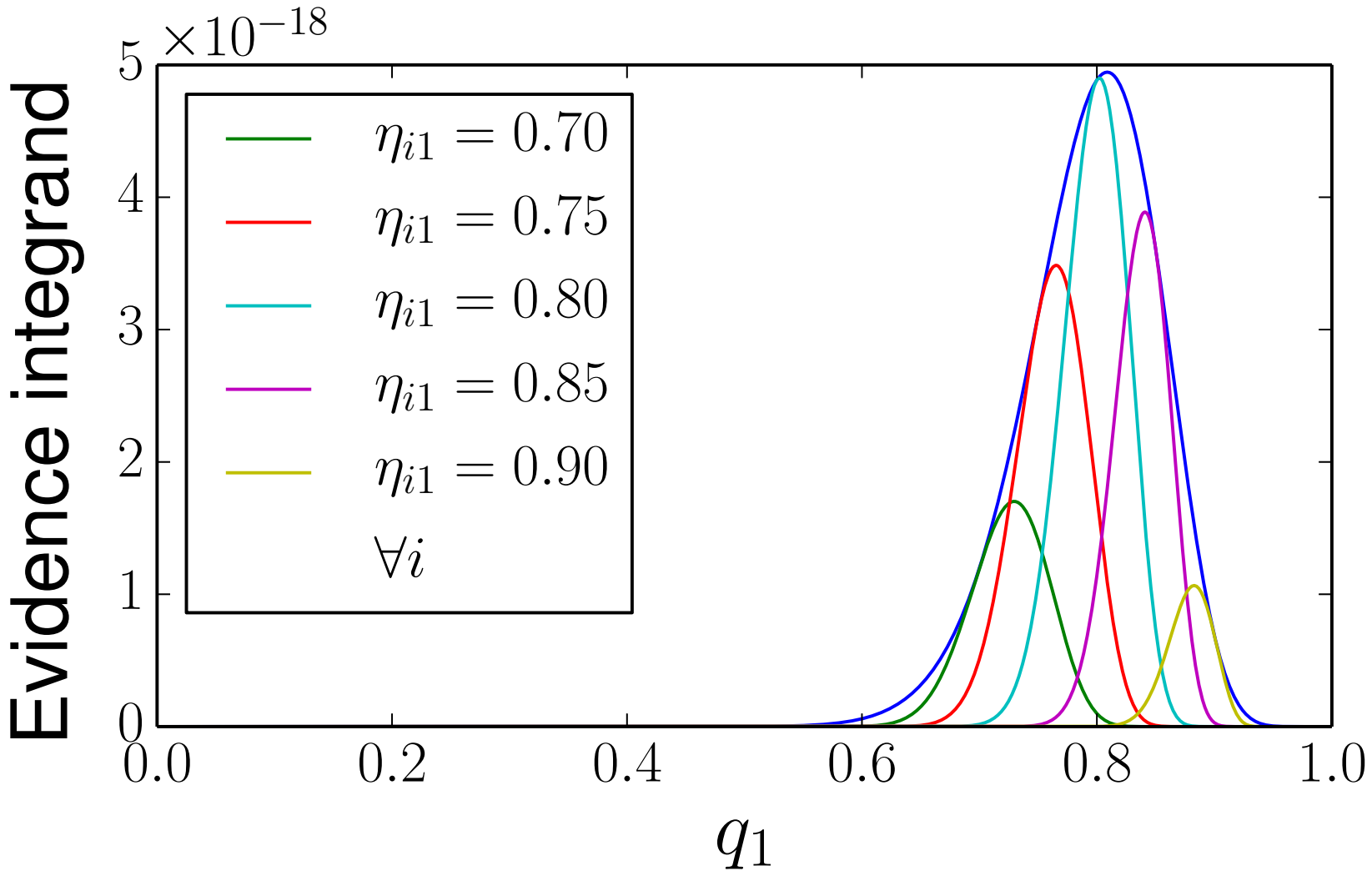
Эти оценки можно использовать для максимизации обоснованности:

$$\begin{aligned}\mathbf{H}^{k+1} &= \arg\max_{\mathbf{H}} \log \tilde{E}(\mathbf{H}, \boldsymbol{\alpha}^k),\\ \boldsymbol{\alpha}^{k+1} &= \arg\max_{\boldsymbol{\alpha}} \log \tilde{E}(\mathbf{H}^{k+1}, \boldsymbol{\alpha})\end{aligned}$$

Здесь не надо вычислять значение оптимизируемой функции!



Нижняя оценка на подынтегральную функцию для обоснованности



Нижняя оценка на обоснованность для модели RTM

$$E(\boldsymbol{\alpha}) \geq \tilde{E}(\boldsymbol{\alpha}, \mathbf{H}) = \left(\prod_{i=1}^{n} c_i(\boldsymbol{\eta}_i)\right) \prod_{j=1}^{d} \int_0^1 \exp\left(\sum_{i:j\in Q_i} \tilde{c}_{ij}(\boldsymbol{\eta}_i) \left(\frac{1-q_j}{q_j}\right)^{|Q_i|(2t_i-1)}\right) p(q_j|\alpha_j) dq_j$$

$$\forall \mathbf{H} \in (0,1)^{n\times d},\ \forall \boldsymbol{\alpha} \in [0,+\infty)^d,$$

где

$$Q_i = \{j | x_{ij} = 1\},$$

$$c_i(\boldsymbol{\eta}_i) = \frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i}(1-\eta_{ij})} \exp\left(\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{1-t_i}(1-\eta_{ij})^{t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i}(1-\eta_{ij})}\right),$$

$$\tilde{c}_{ij}(\boldsymbol{\eta}_i) = -\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i}(1-\eta_{ij})} \left(\frac{\eta_{ij}}{1-\eta_{ij}}\right)^{|Q_i|(2t_i-1)} |Q_i|^{-1}.$$

RTM-full и RTM-MAP

RTM-full:

1. $\mathbf{H}^{new} = \arg\max_{\mathbf{H}} \log \tilde{E}(\boldsymbol{\alpha}^{old}, \mathbf{H})$
2. $\boldsymbol{\alpha}^{new} = \arg\max_{\boldsymbol{\alpha}} \log \tilde{E}(\boldsymbol{\alpha}, \mathbf{H}^{new})$

RTM-MAP:

1. $\boldsymbol{\eta_i}^{new} = \boldsymbol{q}^{MAP} = \arg\max_{\boldsymbol{q}} \mathsf{P}(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q}|\boldsymbol{\alpha}^{old}) \ \forall i.$
2. $\boldsymbol{\alpha}^{new} = \arg\max_{\boldsymbol{\alpha}} \log \tilde{E}(\boldsymbol{\alpha}, \mathbf{H}^{new})$

Сложность RTM-full и RTM-MAP

- n – число объектов
- τ – среднее число тегов у одного объекта
- d – число возможных тегов
- k_i – число подсчетов градиента оценки обоснованности на i-м шаге одной итерации, $i = 1, 2$ (в экспериментах $k_1 \approx k_2 \approx 10$)

Сложность одной итерации метода:

Метод	Число операций интегрирования
RTM-full	$O(n\tau k_1 + dk_2)$
RTM-MAP	$O(dk_2)$

Модельные данные

- 500 объектов
- 50 тегов
- Усреднение по выборкам с различным уровнем зашумления

Точность отбора признаков:

Тип шума	RTM-MAP	RTM-full	RVM	L1-LR
Процент убранных нерелевантных признаков				
Случайный	99.64%	99.46%	99.10%	85.63%
Коррелирующий	84.44%	88.90%	84.65%	100.00%
Процент убранных релевантных признаков				
Случайный	4.50%	4.68%	2.63%	3.54%
Коррелирующий	2.50%	4.34%	1.04%	2.50%

Задача семантического анализа

- Классификация предложений по эмоциональной окраске
- Каждый объект задается набором своих слов (теги = слова)

Точность классификации:

RTM-MAP	RVM	L1-LR	RF	GBDT	SVM
0.9659	0.9586	0.9708	0.9416	0.9683	0.9683

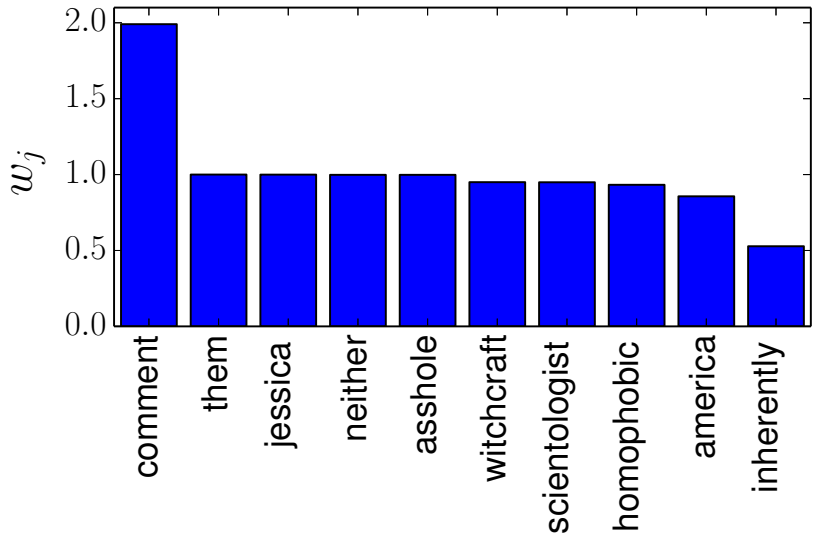
RVM — машина релевантных векторов
L1-LR — логистическая регрессия с L1-регуляризатором
RF — случайный лес
GBDT — градиентный бустинг решающих пней
SVM — машина опорных векторов

Задача семантического анализа: отбор признаков

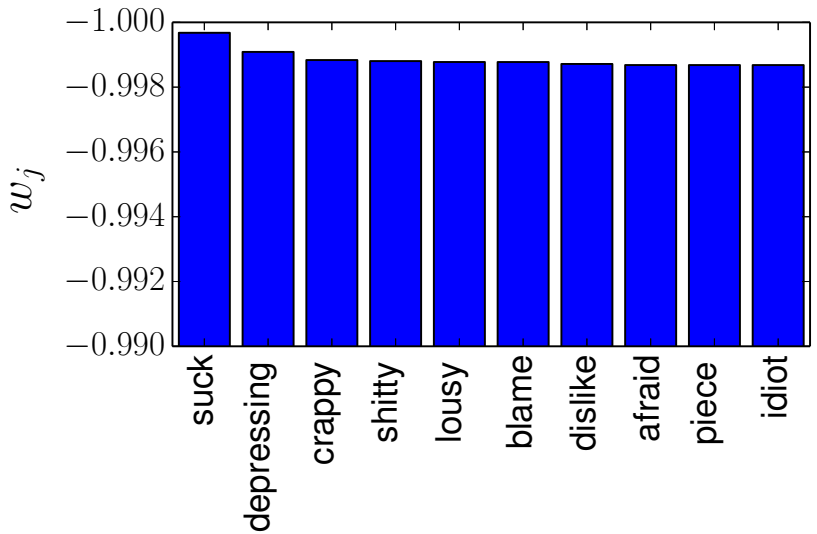
Метод	Число оставленных признаков
RTM-MAP	70
L1-LR	120
RVM	230

Веса линейной модели, настроенной RVM:

Позитивный класс:



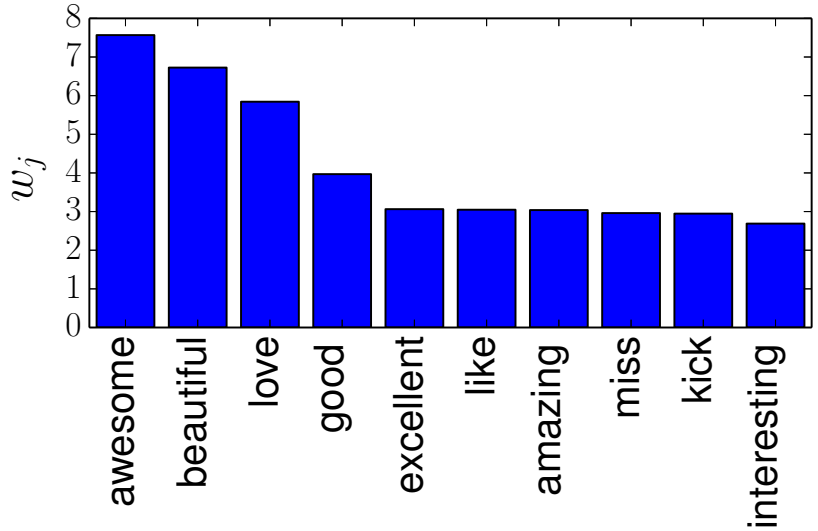
Негативный класс:



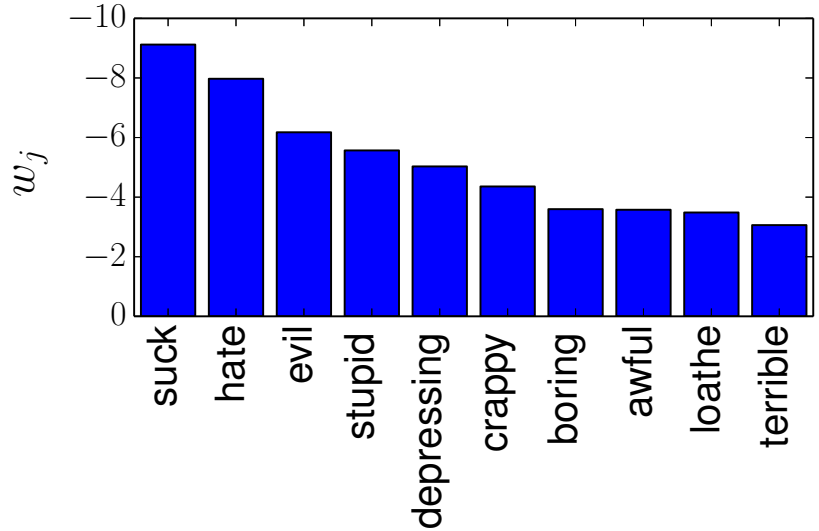
Задача семантического анализа: отбор признаков

Веса линейной модели, настроенной логистической регрессией:

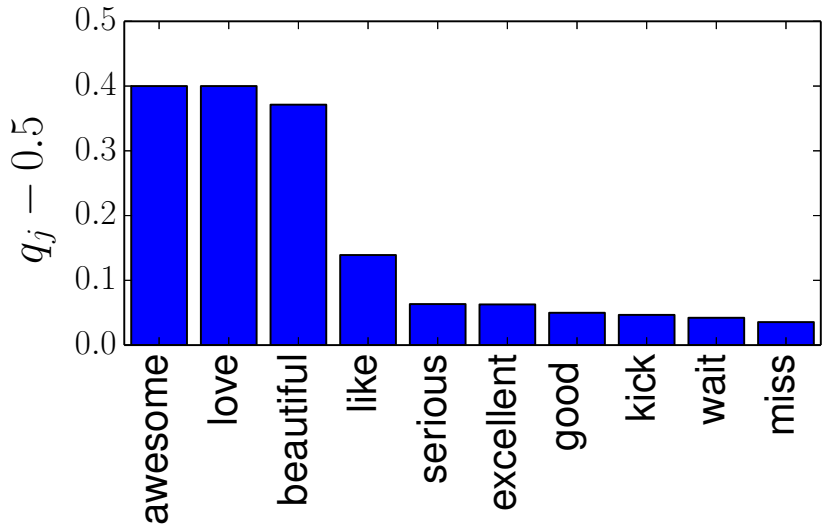
Позитивный класс:

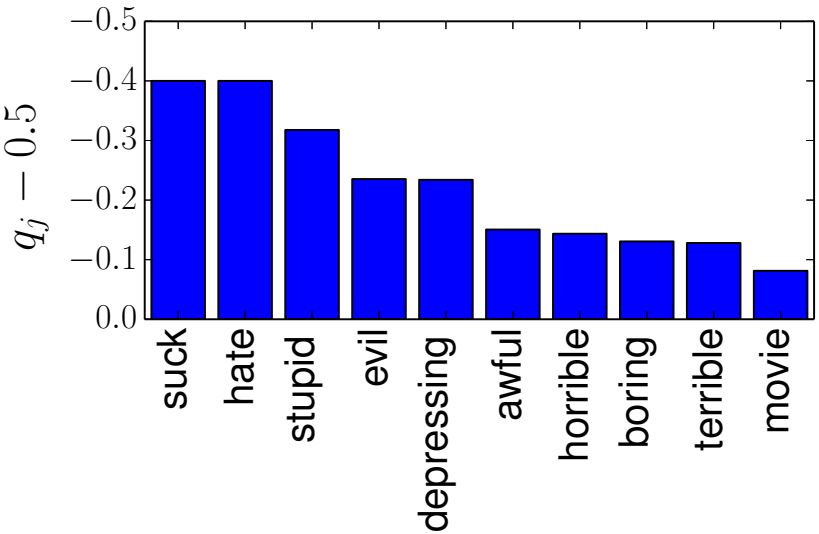


Негативный класс:



$q_j - 0.5$ для самых важных слов у RTM-MAP:





Заключение

Предложено:

- Модель бинарной классификации для данных с бинарными признаками
- Метод обучения, позволяющий проводить автоматический отбор признаков

Показано:

- Качество классификации и отбора признаков предложенного метода сравнимо с качеством классических моделей
- Модель получается более разреженной
- Эффект ARD наблюдается также и на сложных моделях с априорным распределением, отличным от нормального