Машинное обучение на основе минимизации сглаженных оценок средних, нечувствительных к выбросам

З. М. Шибзухов

«МАТЕМАТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ» (ММРО – 2019)

Принцип экстремума средних арифметических

Многие задачи машинного обучения сводятся к задаче минимизации среднего арифметического от конечного набора параметризованных функций:

$$\mathbf{w}^* \in \arg\min_{\mathbf{w}} Q(\mathbf{w}),$$

$$Q(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^{N} \ell_k(\mathbf{w}) + \tau R(\mathbf{w}).$$

Здесь

- $\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})$ – набор неотрицательных параметризованных функций;
- $R(\mathbf{w})$ – регуляризатор, τ – параметр регуляризации.

Проблема выбросов

- Эмпирическое распределение $\ell_{(1)}(\mathbf{w}), \ldots, \ell_{(N)}(\mathbf{w})$ может содержать выбросы:
  - существенные искажения или ошибки в исходных данных;
  - неадекватность модели на части исходных данных.
- Среднее арифметическое не является устойчивой оценкой среднего значения по отношению к выбросам.
- Поэтому выбросы могут приводить к существенному искажению значений целевой функции.
- В свою очередь, это приводит к существенному искажению значений искомых параметров.

Как разрешить проблему выбросов?

- Конечно, можно было бы перейти к взвешенному среднему арифметическому:

$$Q(\mathbf{w}) = \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) + \tau R(\mathbf{w}),$$

  где $v_k \geqslant 0$ и $v_1 + \dots + v_N = 1$.

- Веса элементов выборки $v_1, \dots, v_N$ должны быть способны подавить или существенно ослабить влияние выбросов на значения $\mathbf{w}^*$.
- Вопрос: **как** их выбрать?

Классический метод преодоления влияния выбросов

- Выбирается функция $\varrho(r)$, такая что
  - $\varrho(r)$ монотонно возрастает;
  - $\varrho'(r)$ не возрастает.
- **При этом:**

$$\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\} \longrightarrow \{\varrho(\ell_1(\mathbf{w})), \ldots, \varrho(\ell_N(\mathbf{w}))\},$$

так чтобы выбросы были бы выражены в существенно меньшей степени.

- Целевая функция:

$$\mathcal{Q}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N} \varrho\big(\ell_k(\mathbf{w})\big) + \tau R(\mathbf{w}).$$

- Задача минимизации $\mathcal{Q}(\mathbf{w})$ сводится к решению системы уравнений:

$$\sum_{k=1}^{N} \varrho'\big(\ell_k(\mathbf{w})\big)\nabla\ell_k(\mathbf{w}) + \tau\nabla R(\mathbf{w}) = 0$$

Классический метод преодоления влияния выбросов

Метод итеративного перевзвешивания:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) + \tau R(\mathbf{w}),$$

где $v_k = \varrho'\big(\ell_k(\mathbf{w}_t)\big)$.

Если

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) + \tau R(\mathbf{w}),$$

при $v_k = \varrho'\big(\ell_k(\mathbf{w}^*)\big)$, то $\mathbf{w}^*$ – искомое решение.

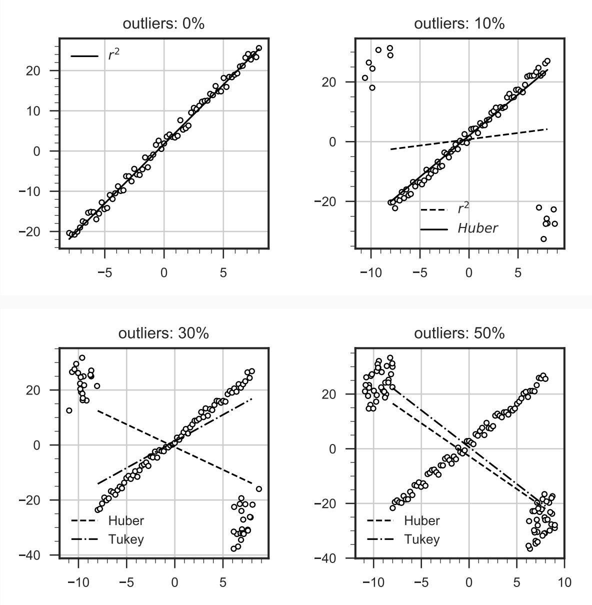
P.S.: Одновременно получаем веса: $v_k^* = \varrho'\big(\ell_k(\mathbf{w}^*)\big)$.

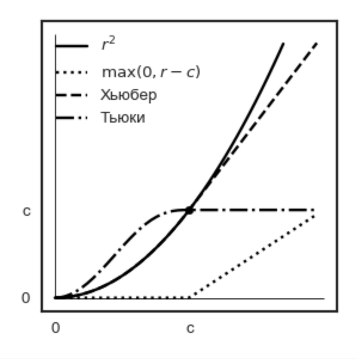
Классический метод подавления влияния выбросов

- Этот метод работает, когда
  - выбросов немного или
  - они невелики
- Однако, **этот метод не работает** если
  - выбросов много (30 – 50%) или
  - они велики (например, превосходят на порядок).

Иллюстративный пример

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N}\sum_{k=1}^{N} \underbrace{\rho(f(\tilde{\mathbf{x}}_k, \mathbf{w}) - \tilde{y}_k)}_{\ell_k(\mathbf{w})}$$





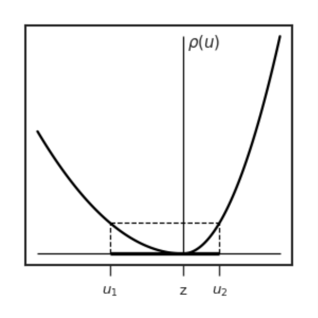
Как преодолеть проблему выбросов?

- Нужно использовать эмпирические оценки среднего значения, нечувствительные или малочувствительные к выбросам, вместо среднего арифметического!
- Медиана, квантили... **Однако**,
  - производные сильно разрывные;
  - градиентные процедуры не применимы.
- Цензурированное среднее арифметическое?

$$\mathrm{M}_u\{z_1, \dots, z_N\} = \frac{1}{N} \sum_{k=1}^{N} \min\{z_k, u\}$$

  Однако,
  - как выбрать u?
  - квантиль $z_1, \dots, z_N$?
- Нужны дифференцируемые оценки среднего значения, нечувствительные к выбросам.

M-средние



Функция несходства

- $\rho(r) \geqslant 0$
- $\rho(0) = 0$
- $\rho(r)$ – (строго) выпуклая функция

$$\bar{z}_\rho = \mathrm{M}_\rho\{z_1, \ldots, z_N\} = \arg\min_u \sum_{k=1}^{N} \rho(z_k - u)$$

$\mathrm{M}\{z_1, \ldots, z_N\}$ – усредняющая агрегирующая функция:

- $z_1' \leqslant z_1'', \ldots, z_N' \leqslant z_N'' \to \mathrm{M}\{z_1', \ldots, z_N'\} \leqslant \mathrm{M}\{z_1'', \ldots, z_N''\}$
- $\min\{z_1, \ldots, z_N\} \leqslant \mathrm{M}\{z_1, \ldots, z_N\} \leqslant \max\{z_1, \ldots, z_N\}$

Вычисление M-средних

- Если существует $\rho'(u) \longrightarrow$ метод градиентного спуска
- Если существует $\rho''(u) \longrightarrow$ метод Ньютона

Итерационная процедура

$$u_{t+1} = \frac{\sum\limits_{k=1}^{N} \varphi(z_k - u_t) z_k}{\sum\limits_{k=1}^{N} \varphi(z_k - u_t)}, \quad \varphi(u) = \frac{\rho'(u)}{u}$$

Достаточное условие сходимости

$$\left| \sum_{k=1}^{N} \varphi'(z_k - \bar{\bar{z}}_\rho)(z_k - \bar{\bar{z}}_\rho) \right| < \left| \sum_{k=1}^{N} \varphi(z_k - \bar{\bar{z}}_\rho) \right|.$$

Дифференцируемые M-средние

Если существует $\rho''(u)$:

$$\frac{\partial \mathrm{M}_\rho\{z_1,\ldots,z_N\}}{\partial z_k} = \frac{\rho''(z_k - \bar{\bar{z}}_\rho)}{\rho''(z_1 - \bar{\bar{z}}_\rho) + \cdots + \rho''(z_N - \bar{\bar{z}}_\rho)}$$

$$\bar{\bar{z}}_\rho = \mathrm{M}_\rho\{z_1,\ldots,z_N\}$$

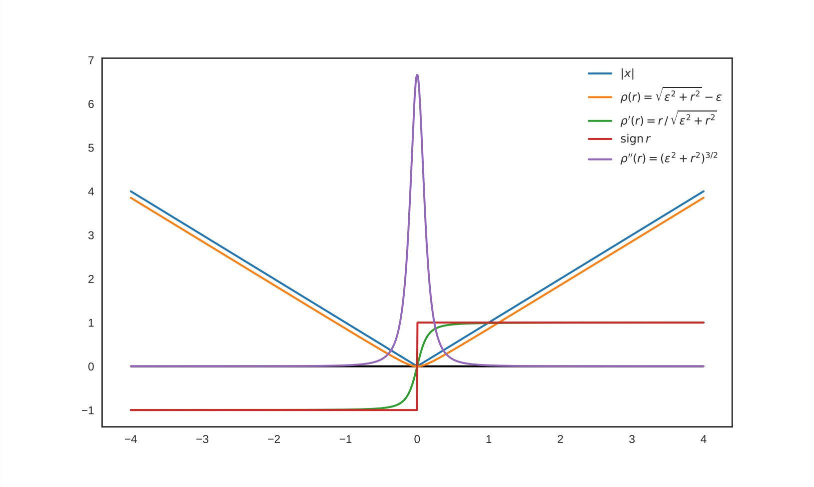
$$\frac{\partial \mathrm{M}_\rho}{\partial z_1} + \cdots + \frac{\partial \mathrm{M}_\rho}{\partial z_N} = 1$$

$$\frac{\partial \mathrm{M}_\rho\{z_1,\ldots,z_N\}}{\partial z_k} \geqslant 0$$

Сглаженная версия медианы

- $\lim\limits_{\varepsilon \to 0} \rho_\varepsilon(r) = |r|$
- $\lim\limits_{\varepsilon \to 0} \rho'_\varepsilon(r) = \operatorname{sign} r$
- $\lim\limits_{\varepsilon \to 0} \rho''_\varepsilon(r) = \begin{cases} +\infty, & \text{если } r = 0 \\ 0, & \text{если } r \neq 0 \end{cases}$

$$\rho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$$



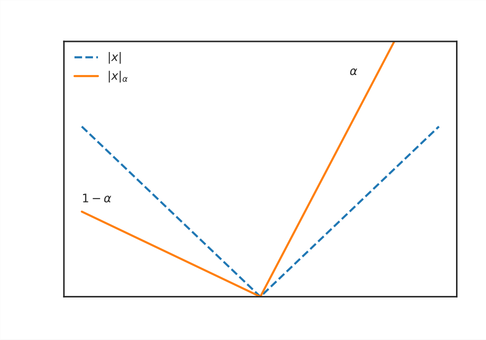
Сглаженная версия квантиля

α-квантиль:

$$\bar{z}_\alpha = \arg\min_u \sum_{k=1}^{N} |z_k - u|_\alpha$$

$$|r|_\alpha = \begin{cases} \alpha r, & \text{если } r > 0 \\ 0, & \text{если } r = 0 \\ (\alpha - 1) r, & \text{если } r < 0 \end{cases}$$

$$\rho_{\alpha,\varepsilon}(r) = \begin{cases} \alpha \rho_\varepsilon(r), & \text{если } r > 0 \\ \frac{\alpha}{2}\rho'_\varepsilon(0^+) + \frac{1-\alpha}{2}\rho'_\varepsilon(0^-)), & \text{если } r = 0 \\ (1-\alpha)\rho_\varepsilon(r), & \text{если } r < 0 \end{cases}$$



По следам функции Хьюбера

Функция Хьюбера (расширенная):

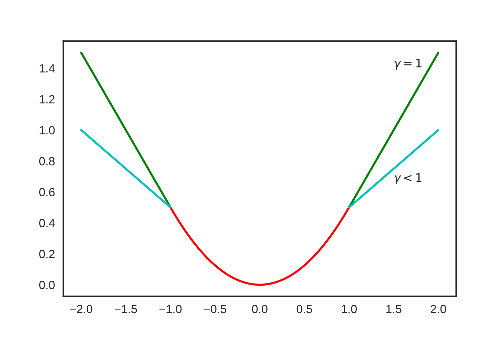
$$H(r) = \begin{cases} \frac{1}{2} r^2, & \text{если } |r| \leqslant c \\ \gamma c r - (\gamma - \frac{1}{2}) c^2, & \text{если } |r| > c \end{cases} \qquad 0 < \gamma \leqslant 1$$

$$c = \bar{\bar{z}}_{\rho_\alpha} = \mathrm{M}_{\rho_\alpha}\{z_1, \ldots, z_N\}.$$

$$\mathrm{M}_H\{z_1, \ldots, z_N\} = \frac{1}{N - m} \sum_{z_k \leqslant \bar{\bar{z}}_{\rho_\alpha}} z_k + \frac{\gamma m}{N - m} \bar{\bar{z}}_{\rho_\alpha},$$

где m – число $z_k > \bar{\bar{z}}_{\rho_\alpha}$.

$$\frac{\partial \mathrm{M}_H}{\partial z_k} = \begin{cases} \frac{1}{N-m} + \frac{\gamma m}{N-m} \frac{\partial \mathrm{M}_{\rho_\alpha}}{\partial z_k}, & \text{if } k \leqslant m \\ \frac{\gamma m}{N-m} \frac{\partial \mathrm{M}_{\rho_\alpha}}{\partial z_k}, & \text{if } k > m \end{cases}$$



Цензурированное среднее арифметическое

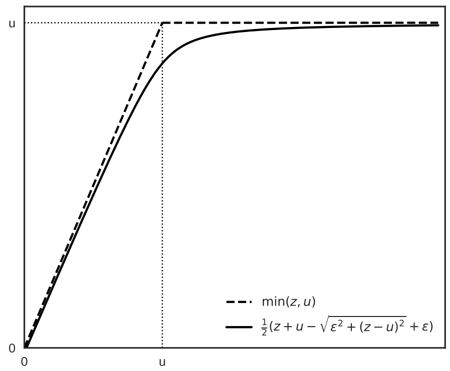
Цензурированное среднее арифметическое:

$$\mathrm{M}_u\{z_1, \ldots, z_N\} = \frac{1}{N}\sum_{k=1}^{N} C(z, u),$$

где $C(z, u)$ – цензурирующая функция.

Например:

- $C(z, u) = \min\{z, u\}$
- $C(z, u) = \frac{1}{2}(z_k + u - \rho(z_k - u))$



Параметр цензуры $u = \bar{\bar{z}}_{\rho_\alpha}$ – значение сглаженного варианта α-квантиля.

Цензурированное среднее арифметическое

- Градиент для $\min(z, \bar{\bar{z}}_{\rho_\alpha})$:

$$\frac{\partial \mathrm{M}_{\bar{\bar{z}}_{\rho_\alpha}}}{\partial z_k} = \begin{cases} \frac{1}{N} + \frac{m}{N}\frac{\partial \mathrm{M}_{\rho_\alpha}}{\partial z_k}, & \text{if } k \leqslant m \\ \frac{m}{N}\frac{\partial \mathrm{M}_{\rho_\alpha}}{\partial z_k}, & \text{if } k > m \end{cases}$$

- Градиент для $\frac{1}{2}\bigl(z_k + \bar{\bar{z}}_{\rho_\alpha} - \rho(z_k - \bar{\bar{z}}_{\rho_\alpha})\bigr)$:

$$\frac{\partial \mathrm{M}_{\bar{\bar{z}}_{\rho_\alpha}}}{\partial z_k} = \frac{1}{2N}\bigl(1 - \rho_\alpha'(z_k - \bar{\bar{z}}_{\rho_\alpha})\bigr) + \frac{1}{2}\frac{\partial \mathrm{M}_{\rho_\alpha}}{\partial z_k}$$

$$\frac{\partial \mathrm{M}_{\bar{\bar{z}}_{\rho_\alpha}}}{\partial z_1} + \cdots + \frac{\partial \mathrm{M}_{\bar{\bar{z}}_{\rho_\alpha}}}{\partial z_N} = 1$$

Принцип минимизации робастной дифференцируемой оценки среднего

- Используем дифференцируемую оценку для среднего значения, нечувствительную или малочувствительную к выбросам.
- Определяем целевую функцию:

$$\mathcal{Q}(\mathbf{w}) = \mathsf{M}\left\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\right\}$$

- Ищем параметры модели, минимизирующие целевую функцию с регуляризатором:

$$\mathbf{w}^\star \in \arg\min_{\mathbf{w}} \left\{\mathcal{Q}(\mathbf{w}) + \tau R(\mathbf{w})\right\}$$

Градиент робастной целевой функции

Целевая функция:

$$\mathscr{Q}(\mathbf{w}) = \mathrm{M}\bigl\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\bigr\}$$

Градиент целевой функции:

$$\nabla\mathscr{Q}(\mathbf{w}) = \sum_{k=1}^{N} v_k(\mathbf{w})\nabla\ell_k(\mathbf{w}),$$

$$\sum_{k=1}^{N} v_k(\mathbf{w}) = 1, \qquad v_k(\mathbf{w}) \geqslant 0$$

$$v_k(\mathbf{w}) = \frac{\partial \mathrm{M}\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}}{\partial z_k}$$

Метод поиска решения

Оптимальный набор параметров $\mathbf{w}^*$ является решением системы:

$$\begin{cases} z_k = \ell_k(\mathbf{w}), & k = 1, \ldots, N \\ v_k = \dfrac{\partial \mathrm{M}\{z_1, \ldots, z_N\}}{\partial z_k}, & k = 1, \ldots, N \\ \sum\limits_{k=1}^{N} v_k \nabla \ell_k(\mathbf{w}) + \tau \nabla R(\mathbf{w}) = 0 & (*) \end{cases}$$

Для его решения применяется метод итеративного перевзвешивания.

P.S.: если v_k – числовые константы, то $(*)$ соответствует решению задачи

$$\arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) + \tau R(\mathbf{w})$$

Метод итеративно перевзвешенного эмпирического среднего

procedure IR-EM($\mathbf{w}_0$)

$t \leftarrow 0$

repeat

$z_1, \ldots, z_N \leftarrow \ell_1(\mathbf{w}_t), \ldots, \ell_N(\mathbf{w}_t)$

$(v_1, \ldots, v_N) = \nabla \mathrm{M}\{z_1, \ldots, z_N\}$

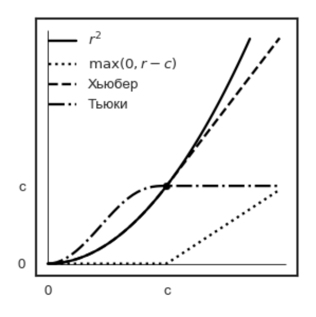
$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}) + \tau R(\mathbf{w})$$

$t \leftarrow t + 1$

until $\{u_t\}$ не сконцентрируется около минимума

end procedure

Классическая робастная M–регрессия



данные: $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N\}$, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_N\}$

невязка: $r_k(\mathbf{w}) = f(\tilde{\mathbf{x}}_k, \mathbf{w}) - \tilde{y}_k$

потери: $\ell_k(\mathbf{w}) = \varrho(r_k(\mathbf{w}))$

функция влияния: $\varrho(r)$

Приницип минимизации эмпирического риска:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N}\sum_{k=1}^{N} \varrho\big(r_k(\mathbf{w})\big) + \tau R(\mathbf{w})$$

Алгоритм **IRLS** (Iteratively Reweighted Least Squares):

$$\mathbf{w}_{t+1} \in \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k r_k(\mathbf{w})^2 + \tau R(\mathbf{w}),$$

где $v_k = \psi(r_k(\mathbf{w}_t))$, $\psi(r) = \varrho'(r)/r$.

Робастная регрессия

Принцип минимизации робастного эмпирического риска

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathrm{M}\{\varrho\big(r_1(\mathbf{w})\big), \dots, \varrho\big(r_N(\mathbf{w})\big)\} + \tau R(\mathbf{w})$$

Алгоритм **IR-ERM** (Iteratively Reweighted Empirical Risk Minimization):

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k r_k(\mathbf{w})^2 + \tau R(\mathbf{w}),$$

$$v_k = \frac{\partial \mathrm{M}\{z_1, \dots, z_k\}}{\partial z_k} \psi\big(r_k(\mathbf{w}_t)\big) \qquad \psi(r) = \frac{\varrho'(r)}{r}$$

$$z_k = \varrho\big(r_k(\mathbf{w}_t)\big)$$

Сглаженная медиана (квантиль):
$$\frac{\partial \mathrm{M}_\rho}{\partial z_N} = \frac{\rho''(z_k - \bar{\bar{z}}_\rho)}{\rho''(z_1 - \bar{\bar{z}}_\rho) + \dots + \rho''(z_N - \bar{\bar{z}}_\rho)}$$

Цензурированное среднее арифметическое:
$$\frac{\partial \mathrm{M}_{\bar{\bar{z}}_\rho}}{\partial z_k} = \begin{cases} \dfrac{1}{N} + \dfrac{m}{N}\dfrac{\partial \mathrm{M}_\rho}{\partial z_k}, & \text{if } k \leqslant m \\ \dfrac{m}{N}\dfrac{\partial \mathrm{M}_\rho}{\partial z_k}, & \text{if } k > m \end{cases}$$

Линейная регрессия

Принцип минимизации робастного эмпирического риска

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathrm{M}_{\rho_\alpha}\{r_1(\mathbf{w})^2, \dots, r_N(\mathbf{w})^2\}$$

где M_{ρ_α} – сглаженная оценка α-квантиля,

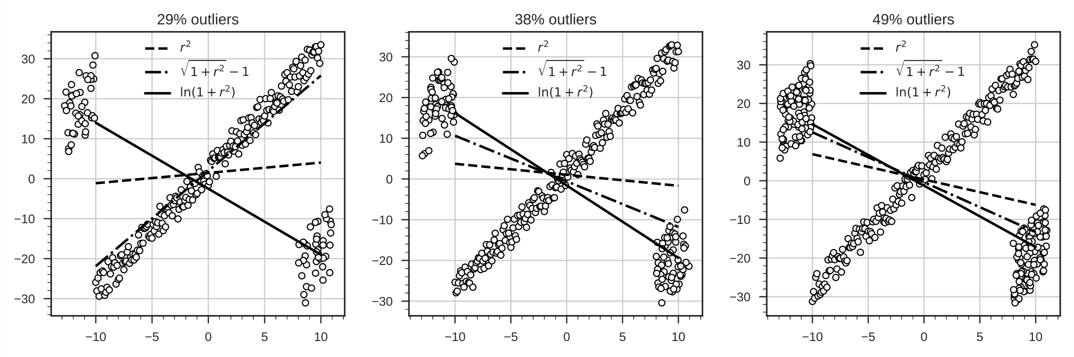
$$\rho_\alpha(r) = \begin{cases} \alpha\rho(r), & \text{if } r > 0 \\ \frac{\alpha}{2}\rho'(0^+) + \frac{1-\alpha}{2}\rho'(0^-), & \text{if } r = 0 \\ (1-\alpha)\rho(r), & \text{if } r < 0 \end{cases}$$

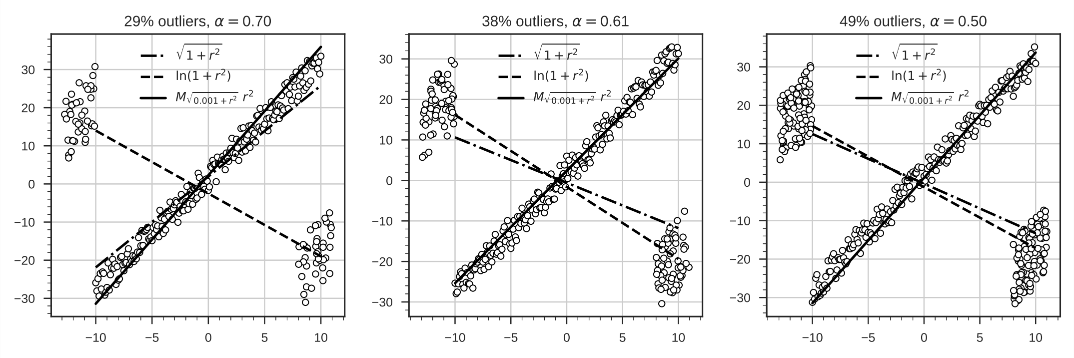
$$\rho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon, \ \varepsilon = 0.001$$

Для сравнения: M-регрессия с $\varrho(r)$:

- $\varrho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$
- $\varrho(r) = \ln(1 + r^2)$

Линейная регрессия





Классическая робастная SVM-классификация (2 класса)

- данные: $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N\}$, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$
- отступ: $r_k(\mathbf{w}) = \big(-\tilde{y}_k f(\tilde{\mathbf{x}}_k, \mathbf{w})\big)_+$ (функция Хинжа от отступа)
- потери: $\ell_k(\mathbf{w}) = \varrho\big(r_k(\mathbf{w})\big)$ ϱ – убывающая функция влияния

Принцип минимизации эмпирического риска (потери от отрицательных отступов)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{k=1}^{N} \underbrace{\varrho\big(r_k(\mathbf{w})\big)}_{\ell_k(\mathbf{w})}$$

Алгоритм **IRHM** (Iteratively Reweighted Hinge Margins):

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k r_k(\mathbf{w}),$$

где $v_k = \varrho'\big(r_k(\mathbf{w}_t)\big)$.

Робастная классификация (SVC)

Принцип минимизации робастного эмпирического риска

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathrm{M}_{\rho_\alpha}\left\{\varrho\big(r_1(\mathbf{w})\big), \ldots, \varrho\big(r_N(\mathbf{w})\big)\right\} + \tau R(\mathbf{w})$$

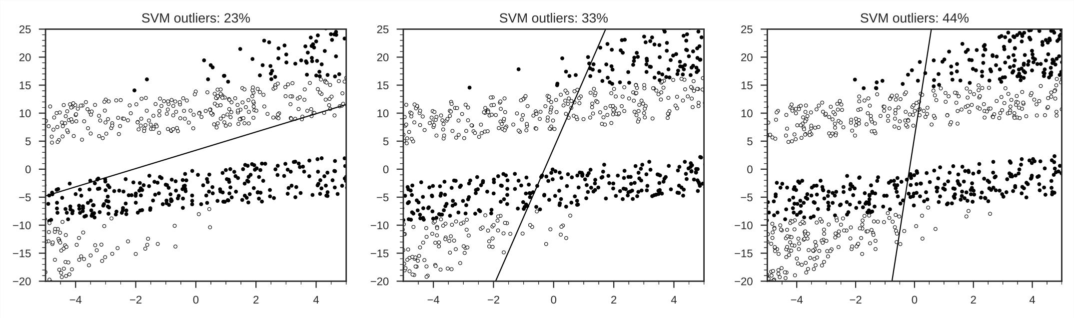
где $\varrho(r)$ – функция Хинжа, M_{ρ_α} – сглаженная оценка α-квантиля,

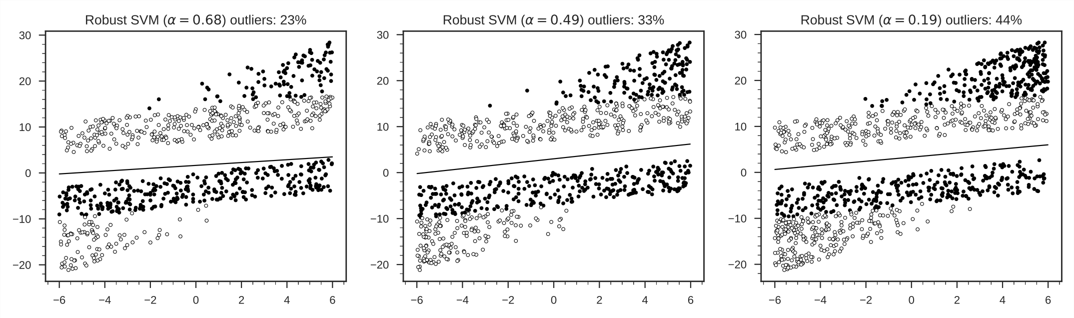
$$\rho_\alpha(r) = \begin{cases} \alpha\rho(r), & \text{if } r > 0 \\ \frac{\alpha}{2}\rho'(0^+) + \frac{1-\alpha}{2}\rho'(0^-), & \text{if } r = 0 \\ (1-\alpha)\rho(r), & \text{if } r < 0 \end{cases}$$

$$\rho(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon, \qquad \varepsilon = 0.001$$

Для сравнения: SVC

Робастная классификация (SVC)





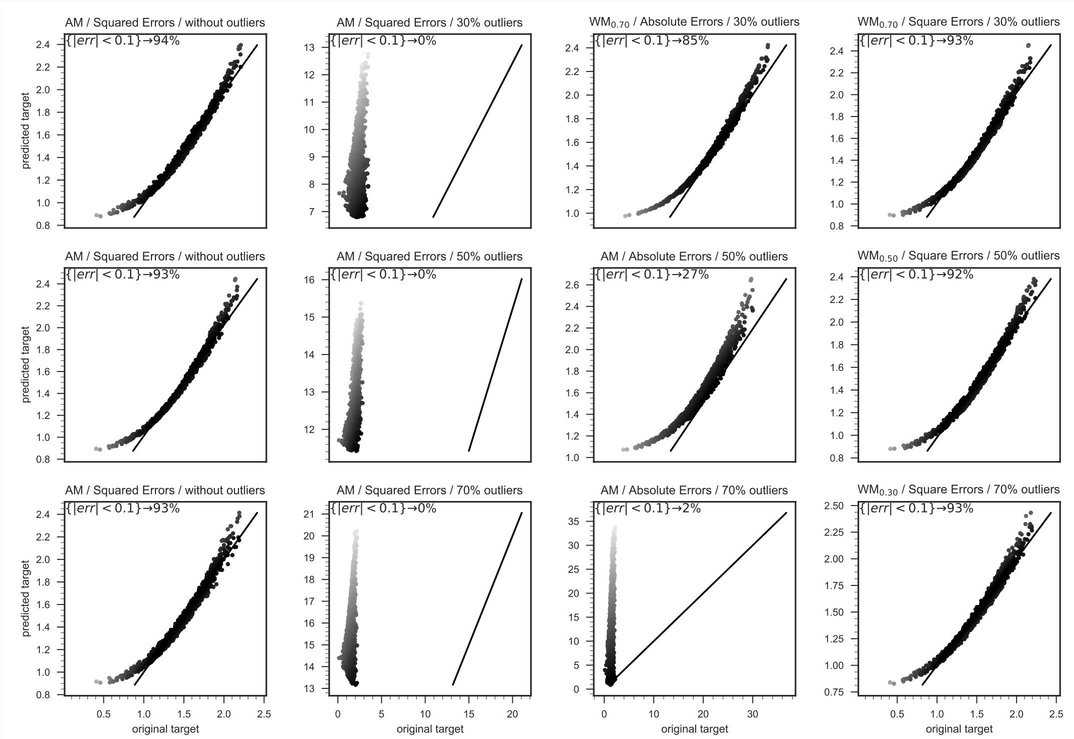
Два примера обучения НС

- Аппроксимируем функцию $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^3$.
- Обучающая/контрольная выборка равномерно распределена на $[-2, 2]^3$.
- Количество точек в обучающей/контрольной выборках $1000/1000$.
- Добавлен равномерно распределенный шум $\varepsilon \in [-0.2, 0.2]$.
- Для выбросов в обучающей выборке исходные значения заменены на значение 30.

Пример 1. $f(\mathbf{x}) = \|\mathbf{x}\|^{2/3}$.
($\max f(\mathbf{x}) < 3$ на $[-2, 2]^3$).

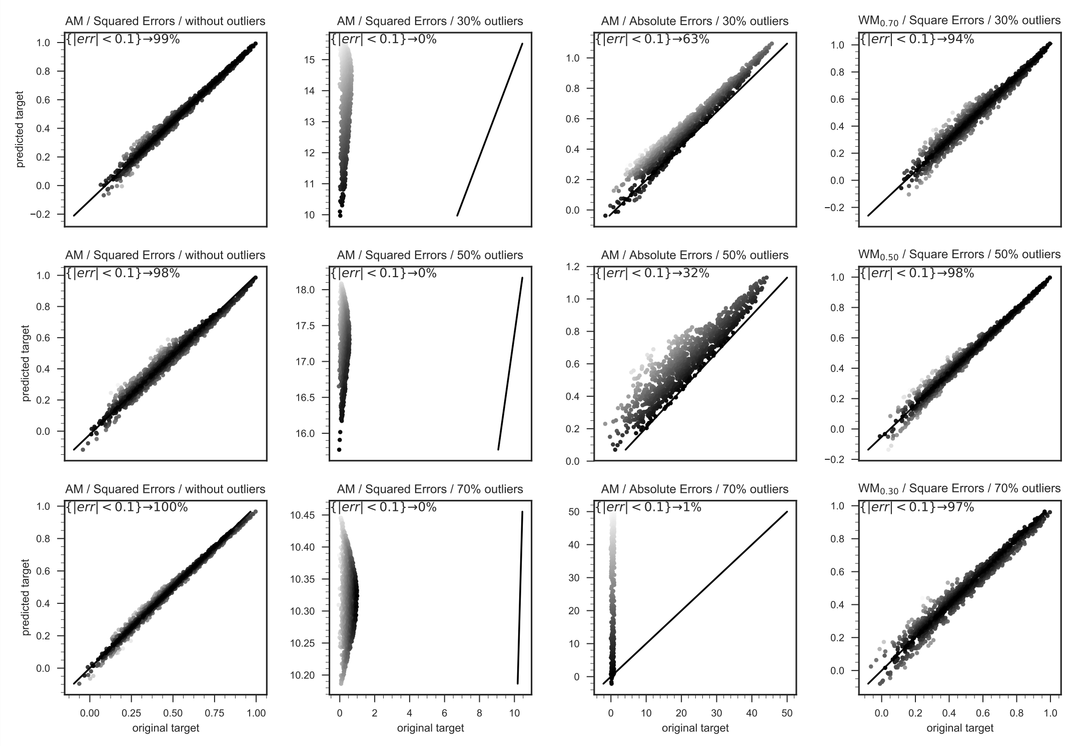
Пример 2. $f(\mathbf{x}) = \sin \|\mathbf{x}\| / (\|\mathbf{x}\| + 10^{-8})$.
($\max f(\mathbf{x}) < 1$ на $[-2, 2]^3$).

Пример 1



AM / Squared Errors / without outliers
{|err| < 0.1}→94%
AM / Squared Errors / 30% outliers
{|err| < 0.1}→0%
WM0.70 / Absolute Errors / 30% outliers
{|err| < 0.1}→85%
WM0.70 / Square Errors / 30% outliers
{|err| < 0.1}→93%
AM / Squared Errors / without outliers
{|err| < 0.1}→93%
AM / Squared Errors / 50% outliers
{|err| < 0.1}→0%
AM / Absolute Errors / 50% outliers
{|err| < 0.1}→27%
WM0.50 / Square Errors / 50% outliers
{|err| < 0.1}→92%
AM / Squared Errors / without outliers
{|err| < 0.1}→93%
AM / Squared Errors / 70% outliers
{|err| < 0.1}→0%
AM / Absolute Errors / 70% outliers
{|err| < 0.1}→2%
WM0.30 / Square Errors / 70% outliers
{|err| < 0.1}→93%
predicted target
original target

Пример 2



AM / Squared Errors / without outliers
{|err| < 0.1}→99%
AM / Squared Errors / 30% outliers
{|err| < 0.1}→0%
AM / Absolute Errors / 30% outliers
{|err| < 0.1}→63%
WM0.70 / Square Errors / 30% outliers
{|err| < 0.1}→94%
AM / Squared Errors / without outliers
{|err| < 0.1}→98%
AM / Squared Errors / 50% outliers
{|err| < 0.1}→0%
AM / Absolute Errors / 50% outliers
{|err| < 0.1}→32%
WM0.50 / Square Errors / 50% outliers
{|err| < 0.1}→98%
AM / Squared Errors / without outliers
{|err| < 0.1}→100%
AM / Squared Errors / 70% outliers
{|err| < 0.1}→0%
AM / Absolute Errors / 70% outliers
{|err| < 0.1}→1%
WM0.30 / Square Errors / 70% outliers
{|err| < 0.1}→97%
predicted target
original target

Классический центр и корреляционная матрица

- $\mathbf{X} = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_N\right\} \subset \mathbb{R}^n$ – конечное множество точек
- $p(\mathbf{x}) \propto \frac{1}{\sqrt{|\mathbf{S}|}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})}$ – функция нормального распределения точек с центром $\mathbf{c}$ и корреляционной матрицей $\mathbf{S}$
- $D(\mathbf{x}; \mathbf{c}, \mathbf{S}) = \ln|\mathbf{S}| + (\mathbf{x}-\mathbf{c})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})$ – расстояние Махаланобиса с учетом корреляционной матрицы $\mathbf{S}$

$$\mathcal{Q}(\mathbf{c}, \mathbf{S}) = \ln|\mathbf{S}| + \frac{1}{N}\sum_{k=1}^{N} D(\mathbf{x}_k; \mathbf{c}, \mathbf{S}),$$

$$\begin{cases} \mathbf{c}^* = \frac{1}{N}\sum\limits_{k=1}^{N} \mathbf{x}_k \\ \mathbf{S}^* = \frac{1}{N}\sum\limits_{k=1}^{N} (\mathbf{x}_k - \mathbf{c})'(\mathbf{x}_k - \mathbf{c}). \end{cases}$$

М-эстиматор: центр и корреляционная матрица

Используется функция ρ, растущая медленнее линейной.

- $p(\mathbf{x}) \propto \frac{1}{\sqrt{|\mathbf{S}|}} e^{-\frac{1}{2}\rho\left((\mathbf{x}-\mathbf{c})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})\right)}$ – функция распределения точек с центром $\mathbf{c}$ и корреляционной матрицей $\mathbf{S}$
- $D(\mathbf{x}; \mathbf{c}, \mathbf{S}) = \ln|\mathbf{S}| + \rho\left((\mathbf{x}-\mathbf{c})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{c})\right)$ – псевдо-расстояние

$$Q(\mathbf{c}, \mathbf{S}) = \ln|\mathbf{S}| + \frac{1}{N}\sum_{k=1}^{N} \rho\left(D(\mathbf{x}_k; \mathbf{c}, \mathbf{S})\right),$$

$z_k = z_k(\mathbf{c}, \mathbf{S}) = D(\mathbf{x}_k; \mathbf{c}, \mathbf{S})$.

$$\begin{cases} v_k = \rho'(z_k) \\ \mathbf{c}^* = \frac{1}{N}\sum_{k=1}^{N} v_k \mathbf{x}_k \\ \mathbf{S}^* = \frac{1}{N}\sum_{k=1}^{N} v_k (\mathbf{x}_k - \mathbf{c})'(\mathbf{x}_k - \mathbf{c}). \end{cases}$$

Принцип минимизации робастной оценки

- Используем робастное среднее для оценки эмпирического среднего:

$$\mathcal{Q}(\mathbf{c}, \mathbf{S}) = \mathrm{M}\big\{z_1(\mathbf{c}, \mathbf{S}), \ldots, z_N(\mathbf{c}, \mathbf{S})\big\},$$

  где $z_k = z_k(\mathbf{c}, \mathbf{S}) = \ln|\mathbf{S}| + D(\mathbf{x}_k; \mathbf{c}, \mathbf{S})$.

- Искомые значения центра и корреляционной матрицы – решения системы уравнений:

$$\begin{cases} \mathbf{v} = \nabla \mathrm{M}\{z_1, \ldots, z_N\} \\ 0 = \sum\limits_{k=1}^{N} v_k(\mathbf{x}_k - \mathbf{c}) \\ \mathbf{S} = \sum\limits_{k=1}^{N} v_k(\mathbf{x}_k - \mathbf{c})'(\mathbf{x}_k - \mathbf{c}). \end{cases}$$

Алгоритм

1. На начальном шаге $t = 0$ выбираются начальные $\mathbf{c}_0$ и $\mathbf{S}_0$.
2. На t-ом шаге последовательно решаются два уравнения:
   1. Сначала решается следующее уравнение для нахождения $\mathbf{c}_{t+1}$:
   $$\mathbf{c} = \sum_{k=1}^{N} \frac{\partial \mathrm{M}\{z_1, \dots, z_N\}}{\partial z_k} \mathbf{x}_k,$$
   где $z_k = D(\mathbf{x}_k; \mathbf{c}, \mathbf{S}_t)$.
   2. Затем решается следующее уравнение для нахождения $\mathbf{S}_{t+1}$:
   $$\mathbf{S} = \sum_{k=1}^{N} \frac{\partial \mathrm{M}\{z_1, \dots, z_N\}}{\partial z_k} (\mathbf{x}_k - \mathbf{c}_{t+1})'(\mathbf{x}_k - \mathbf{c}_{t+1}),$$
   где $z_k = D(\mathbf{x}_k; \mathbf{c}_{t+1}, \mathbf{S})$.
3. Шаг $\boxed{2}$ повторяется до тех пор, пока $t < T$ (максимальное число итераций) или последовательность $\{Q(\mathbf{c}_t, \mathbf{S}_t)\}$ не сконцентрируется вокруг своей точки сгущения.

Алгоритм

- Для решения уравнения

$$\mathbf{c} = f(\mathbf{c})$$

используем итерационную схему

$$\mathbf{c}_{t+1} = (1 - h)\mathbf{c}_t + hf(\mathbf{c}_t)$$

- Для решения уравнения

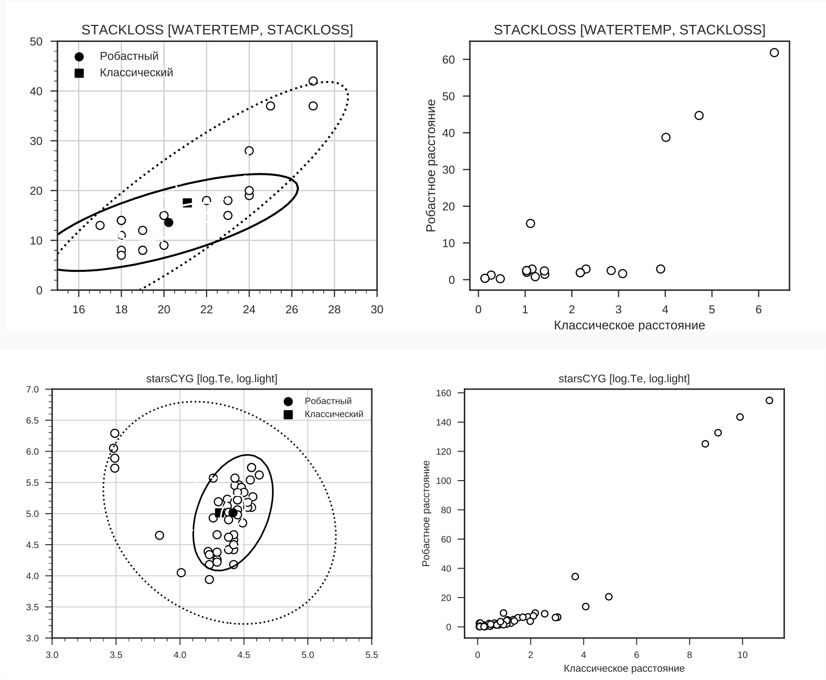
$$\mathbf{S} = F(\mathbf{S})$$

используем итерационную схему

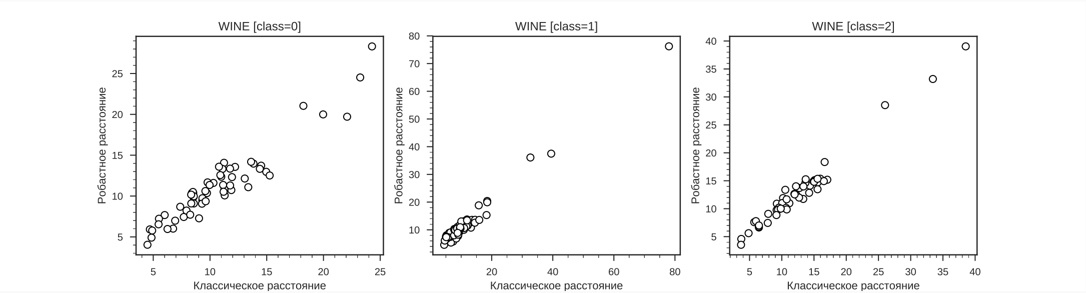
$$\mathbf{S}_{t+1} = (1 - h)\mathbf{S}_t + hF(\mathbf{S}_t),$$

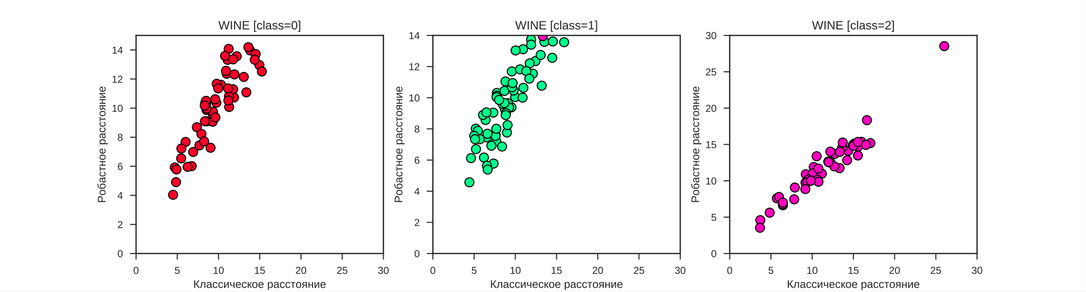
$0 < h < 1$

Пример: stackloss, starCYG



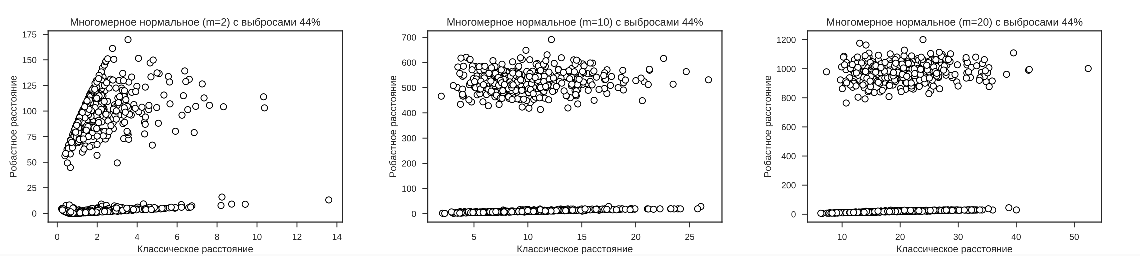
Пример: wine


WINE [class=0]
WINE [class=1]
WINE [class=2]
Робастное расстояние
Классическое расстояние


WINE [class=0]
WINE [class=1]
WINE [class=2]
Робастное расстояние
Классическое расстояние

Пример: нормальное распределение

$N = 500$ точек, $n = 2, 10, 20$, $M = 400$ 44% выбросов



Задача кластеризации

Классическая задача кластеризации (K классов):

$$\mathbf{c}_1^*, \ldots, \mathbf{c}_K^* = \arg\min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} \frac{1}{N} \sum_{k=1}^{N} D(\mathbf{x}_k),$$

где

$$D(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_K) = \min_{j=1,\ldots,K} \|\mathbf{x} - \mathbf{c}_j\|^2$$

Итерационная процедура:

$$\mathbf{c}_{j,t+1} = \frac{1}{|\mathbf{I}_j|} \sum_{k \in \mathbf{I}_j} \mathbf{x}_k,$$

где $\mathbf{I}_j = \{k : j = j(\mathbf{x}_k)\}$ – индексы точек, попадающих в j-ый кластер,

$$j(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_K) = \arg\min_{j=1,\ldots,K} \|\mathbf{x} - \mathbf{c}_j\|^2.$$

Задача робастной кластеризации

Классическая задача кластеризации (K классов):

$$\mathbf{c}_1^*, \ldots, \mathbf{c}_K^* = \arg\min_{\mathbf{c}_1, \ldots, \mathbf{c}_K} \mathrm{M}\left\{D(\mathbf{x}_1; \mathbf{c}_1, \ldots, \mathbf{c}_K), \ldots, D(\mathbf{x}_N; \mathbf{c}_1, \ldots, \mathbf{c}_K)\right\},$$

Система уравнений:

$$\begin{cases} \mathbf{v} = \nabla \mathrm{M}\left\{D(\mathbf{x}_1; \mathbf{c}_1, \ldots, \mathbf{c}_K), \ldots, D(\mathbf{x}_N; \mathbf{c}_1, \ldots, \mathbf{c}_K)\right\} \\ \sum\limits_{k \in \mathbf{I}_j} v_k (\mathbf{x}_k - \mathbf{c}_j) = 0 \end{cases}$$

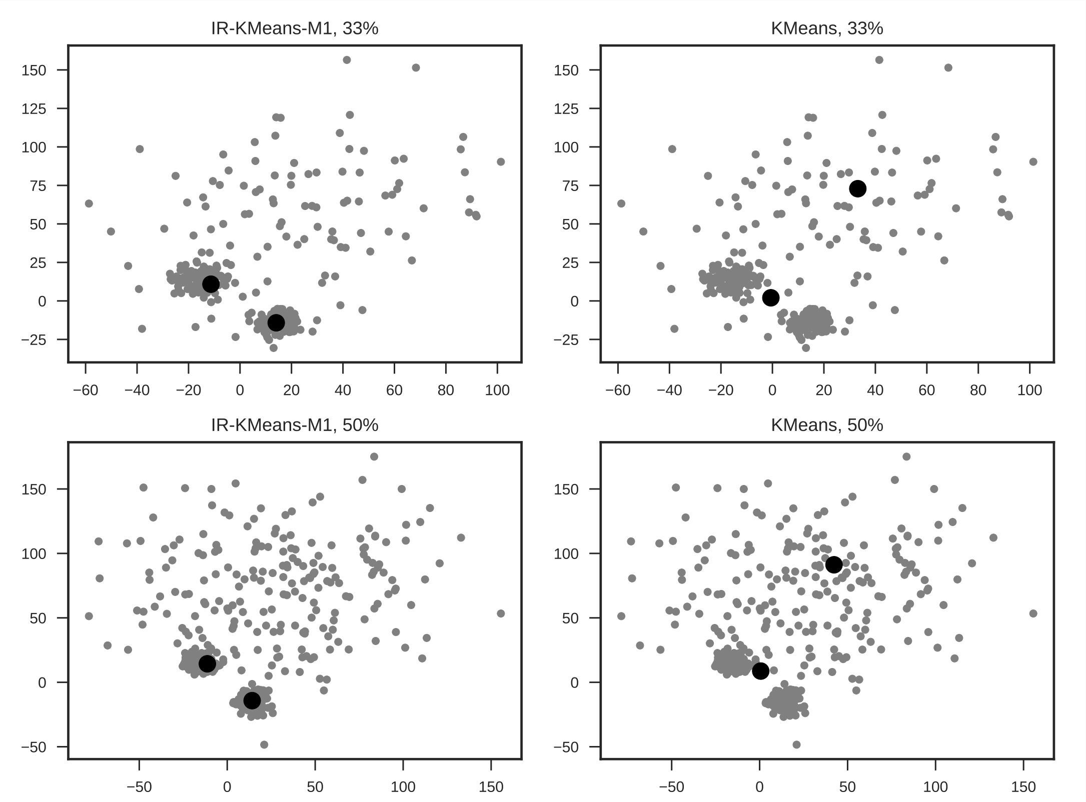
Итерационная процедура:

$$\mathbf{c}_{j,t+1} = \frac{1}{V_j} \sum_{k \in \mathbf{I}_j} v_k \mathbf{x}_k,$$

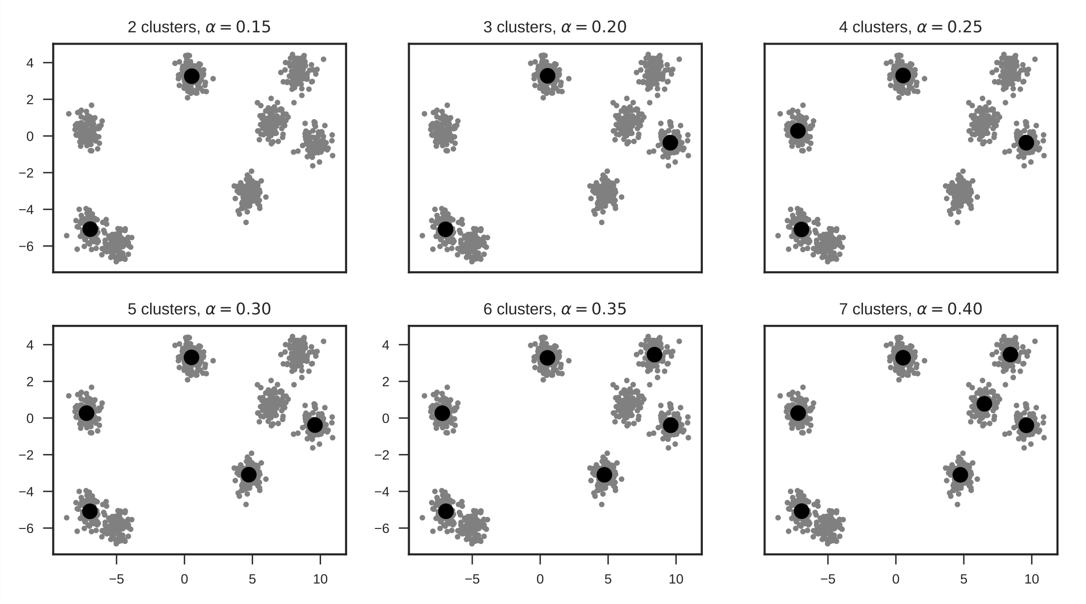
где $V_j = \sum\limits_{k \in \mathbf{I}_j} v_k$,

$$\mathbf{v} = \nabla \mathrm{M}\left\{D(\mathbf{x}_1; \mathbf{c}_{1,t}, \ldots, \mathbf{c}_{K,t}), \ldots, D(\mathbf{x}_N; \mathbf{c}_{1,t}, \ldots, \mathbf{c}_{K,t})\right\}$$

Пример 1



Пример 2



2 clusters, $\alpha = 0.15$
3 clusters, $\alpha = 0.20$
4 clusters, $\alpha = 0.25$
5 clusters, $\alpha = 0.30$
6 clusters, $\alpha = 0.35$
7 clusters, $\alpha = 0.40$
4
2
0
−2
−4
−6
−5
0
5
10

Благодарю за внимание!