Машинное обучение

Лекция 8. Линейные модели, линейная регрессия

МФТИ 2018

Алексей Романенко, alexromsput@gmail.com

Материалы: В. Кантор, К. Воронцов

План

1. Сингулярное разложение

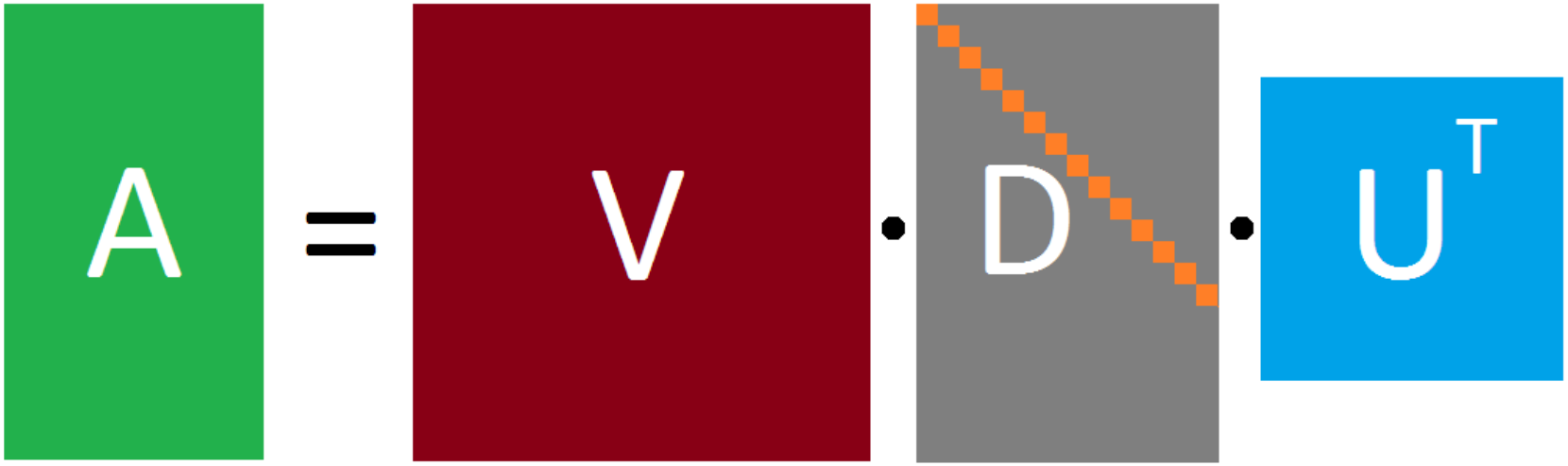
2. Линейная регрессия

3. Дополнительные темы
   - SVM в задаче SemiSupervised
   - Робастные модели

1. Сингулярное разложение

Сингулярное разложение

A – произвольная (вещественная) матрица $n \times m$



$$A = VDU^T$$

- V - ортогональная матрица $n \times n$, $(V^T V = I_n)$
- D - диагональная матрица размером $n \times m$
- U - ортогональная матрица $m \times m$, $(UU^T = I_m)$

Understanding SVD

- $G = A^T A$ – имеет ортонормированный базис из собственных векторов

$$Gx = \lambda x$$

- Базис обозначим как $P = (x_1, \dots, x_m)$

- В этом базисе G имеет диагональный вид:

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_m \end{bmatrix}$$

$$\lambda_i \geq 0, i = \overline{1, m}$$

- Для матрицы $G' = AA^T$ - те же собственные числа

Understanding SVD

- Обозначения: $\sigma_i = \sqrt{\lambda_i}$

- Построим матрицы, D:

$$D = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix}$$

- *V:*

$$V = [v_1 \quad \cdots \quad v_n]$$

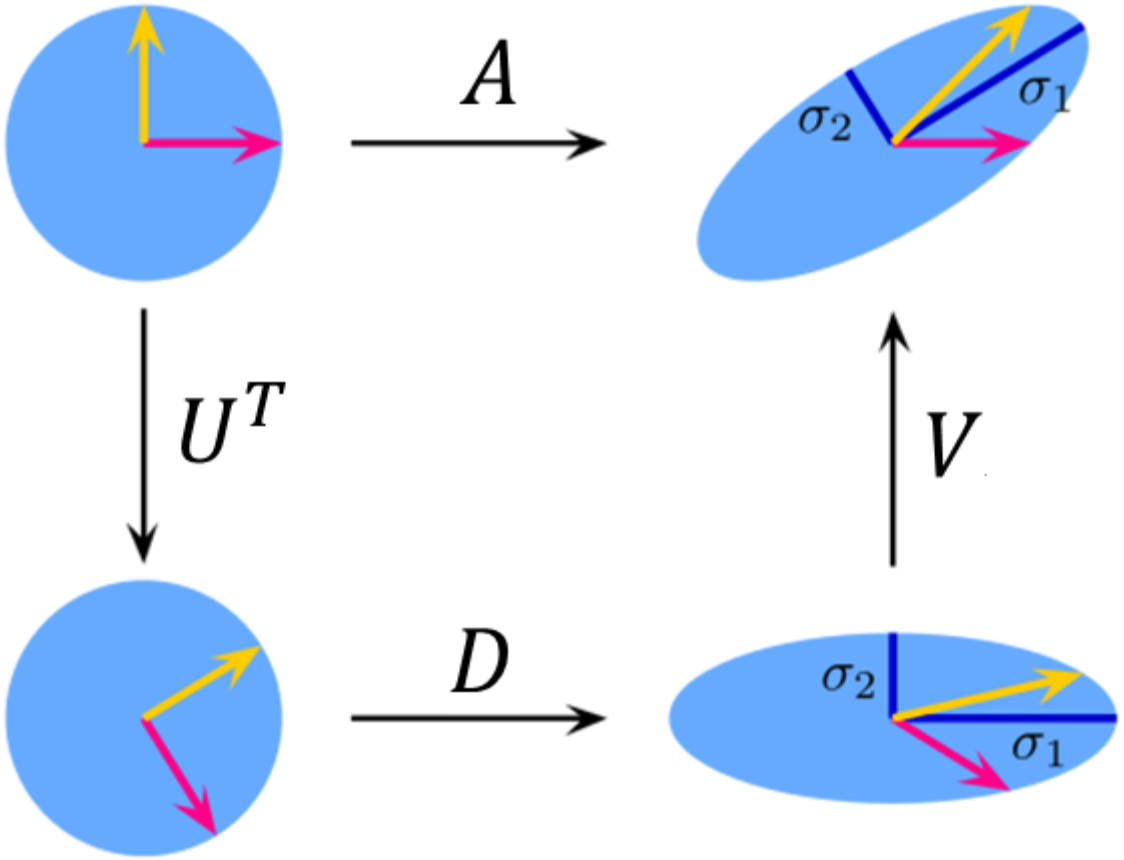
v_i - собственный вектор матрицы AA^T

- *U:*

$$U = [x_1 \quad \cdots \quad x_m]$$

x_i - собственный вектор матрицы $A^T A$

Understanding SVD



$$A = VDU^T$$

2. Линейная регрессия

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

$$L\big(y_i, a(x_i)\big) = (y_i - a(x_i))^2$$

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^{N} L\big(y_i, a(x_i)\big)$$

$$L\big(y_i, a(x_i)\big) = (y_i - a(x_i))^2$$

$$L\big(y_i, a(x_i)\big) = |y_i - a(x_i)|$$

Линейная регрессия: формальная постановка

- X — объекты (часто $\mathbb{R}^n$); Y — ответы (часто $\mathbb{R}$, реже $\mathbb{R}^m$);
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
  $y_i = y(x_i)$, $y : X \to Y$ — неизвестная зависимость;

- $a(x) = f(x, w)$ — модель зависимости,
  $w \in \mathbb{R}^p$ — вектор параметров модели.

- Метод наименьших квадратов (МНК):

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} k_i \big(f(x_i, w) - y_i\big)^2 \to \min_w,$$

где k_i — вес, степень важности i-го объекта.

$Q(w^*, X^\ell)$ — *остаточная сумма квадратов*
(residual sum of squares, RSS).

Линейная регрессия: формальная постановка

- X — объекты (часто $\mathbb{R}^n$); Y — ответы (часто $\mathbb{R}$, реже $\mathbb{R}^m$);
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
  $y_i = y(x_i)$, $y : X \to Y$ — неизвестная зависимость;

- $a(x) = f(x, w)$ — модель зависимости,
  $w \in \mathbb{R}^p$ — вектор параметров модели.

- Метод наименьших квадратов (МНК):

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} k_i \big(f(x_i, w) - y_i\big)^2 \to \min_w,$$

где k_i — вес, степень важности i-го объекта.

$Q(w^*, X^\ell)$ — *остаточная сумма квадратов*
(residual sum of squares, RSS).

Линейная регрессия: формальная постановка

$f_1(x), \ldots, f_n(x)$ — числовые признаки;
Модель многомерной линейной регрессии:

$$f(x, w) = \sum_{j=1}^{n} w_j f_j(x), \qquad w \in \mathbb{R}^n.$$

Матричные обозначения:

$$\underset{\ell \times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}, \quad \underset{n \times 1}{w} = \begin{pmatrix} w_1 \\ \ldots \\ w_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} \big(f(x_i, w) - y_i\big)^2 = \|Fw - y\|^2 \to \min_w.$$

Линейная регрессия: система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial w}(w) = 2F^{\intercal}(Fw - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

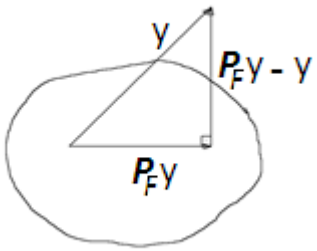
$$F^{\intercal}Fw = F^{\intercal}y,$$

где $\underset{n\times n}{F^{\intercal}F}$ — ковариационная матрица набора признаков $f_1, \ldots, f_n$.

Решение системы: $w^* = (F^{\intercal}F)^{-1}F^{\intercal}y = F^{+}y.$

Значение функционала: $Q(w^*) = \|P_F y - y\|^2,$

где $P_F = FF^{+} = F(F^{\intercal}F)^{-1}F^{\intercal}$ — *проекционная матрица.*



SVD для F:

$$F = VDU^T$$

Линейная регрессия: МНК через SVD

Псевдообратная F^+, вектор МНК-решения w^*, МНК-аппроксимация целевого вектора Fw^*:

$$F^+ = (UDV^\intercal VDU^\intercal)^{-1} UDV^\intercal = UD^{-1}V^\intercal = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^\intercal;$$

$$w^* = F^+ y = UD^{-1}V^\intercal y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^\intercal y);$$

$$Fw^* = P_F y = (VDU^\intercal) UD^{-1} V^\intercal y = VV^\intercal y = \sum_{j=1}^{n} v_j (v_j^\intercal y);$$

$$\|w^*\|^2 = \|D^{-1}V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^\intercal y)^2.$$

Линейная регрессия: проблема переобучения

Если имеются $\lambda_j \to 0$, то

- МНК-решение w^* неустойчиво и неинтерпретируемо: $\|w\| \to \infty$;
- ответы на новых объектах $y' = F'w^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(w^*) = \|Fw^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|w\| \to \min$;
- Преобразование признаков: $f_1, \ldots, f_n \to g_1, \ldots, g_m, \;\; m \ll n$;
- Отбор признаков: $f_1, \ldots, f_n \to f_{j_1}, \ldots, f_{j_m}, \;\; m \ll n$.

Регуляризация: гребневая регрессия

Штраф за увеличение нормы вектора весов $\|w\|$:

$$Q_\tau(w) = \|Fw - y\|^2 + \frac{1}{2\sigma}\|w\|^2,$$

где $\tau = \frac{1}{\sigma}$ — неотрицательный *параметр регуляризации*.

Вероятностная интерпретация: априорное распределение вектора w — гауссовское с ковариационной матрицей σI_n.

Модифицированное МНК-решение (τI_n — «гребень»):

$$w_\tau^* = (F^\intercal F + \tau I_n)^{-1} F^\intercal y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризация : Гребневая регрессия

Вектор регуляризованного МНК-решения w_τ^* и МНК-аппроксимация целевого вектора Fw_τ^*:

$$w_\tau^* = U(D^2 + \tau I_n)^{-1} D V^\intercal y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^\intercal y);$$

$$Fw_\tau^* = VDU^\intercal w_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^\intercal y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^\intercal y);$$

$$\|w_\tau^*\|^2 = \|D^2(D^2 + \tau I_n)^{-1} D^{-1} V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j + \tau} (v_j^\intercal y)^2.$$

$Fw_\tau^* \neq Fw^*$, но зато решение становится гораздо устойчивее.

Регуляризация: гребневая регрессия

Сжатие (shrinkage) или *сокращение весов* (weight decay):

$$\|w_\tau^*\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j + \tau}(v_j^\intercal y)^2 \quad < \quad \|w^*\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j}(v_j^\intercal y)^2.$$

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

$$\operatorname{tr} F(F^\intercal F)^{-1}F^\intercal = \operatorname{tr}(F^\intercal F)^{-1}F^\intercal F = \operatorname{tr} I_n = n.$$

При использовании регуляризации:

$$\operatorname{tr} F(F^\intercal F + \tau I_n)^{-1}F^\intercal = \operatorname{tr}\operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} < n.$$

Регуляризация: Лассо Тибширани

$$\begin{cases} Q(w) = \|Fw - y\|^2 \to \min\limits_{w}; \\ \sum\limits_{j=1}^{n} |w_j| \leqslant \varkappa; \end{cases}$$

Лассо приводит к отбору признаков! Почему?

После замены переменных

$$\begin{cases} w_j = w_j^+ - w_j^-; \\ |w_j| = w_j^+ + w_j^-; \end{cases} \qquad w_j^+ \geqslant 0; \quad w_j^- \geqslant 0.$$

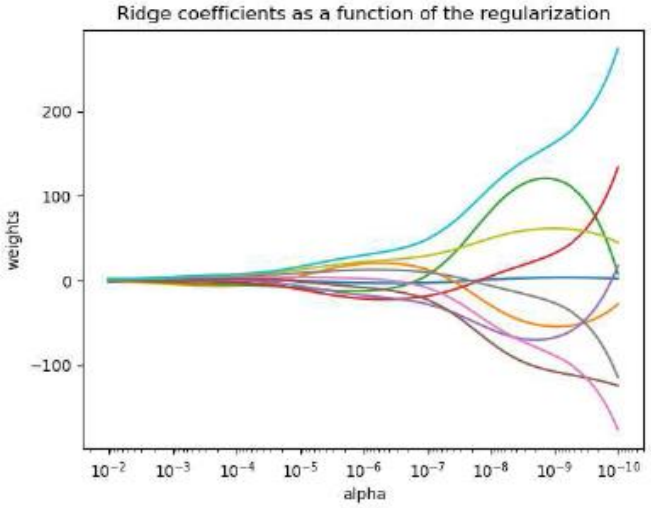
ограничения принимают канонический вид:

$$\sum_{j=1}^{n} w_j^+ + w_j^- \leqslant \varkappa; \quad w_j^+ \geqslant 0; \quad w_j^- \geqslant 0.$$

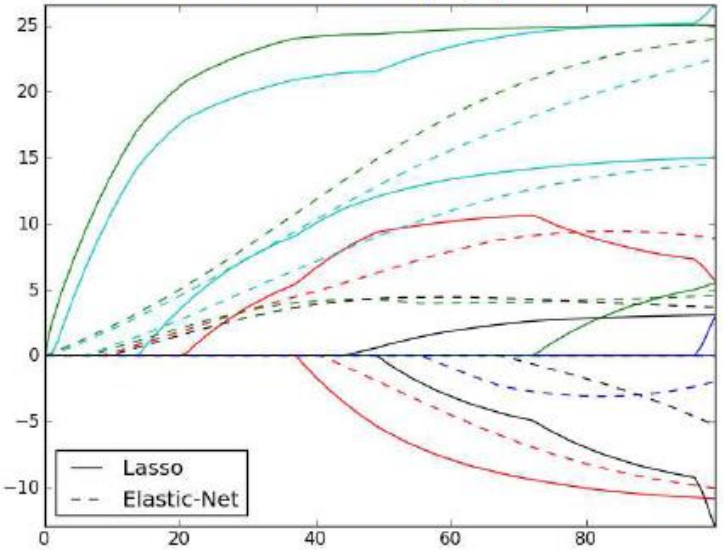
Чем меньше $\varkappa$, тем больше j таких, что $w_j^+ = w_j^- = 0$.

Регуляризация: Lasso vs Ridge

Зависимость $\{w_j\}$ от σ



Зависимость $\{w_j\}$ от $\varkappa$

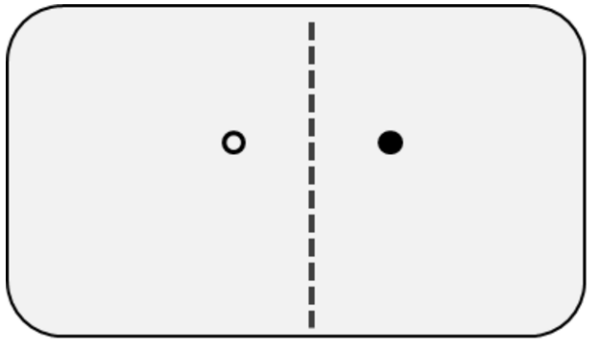


Задача диагностики рака (prostate cancer, UCI)

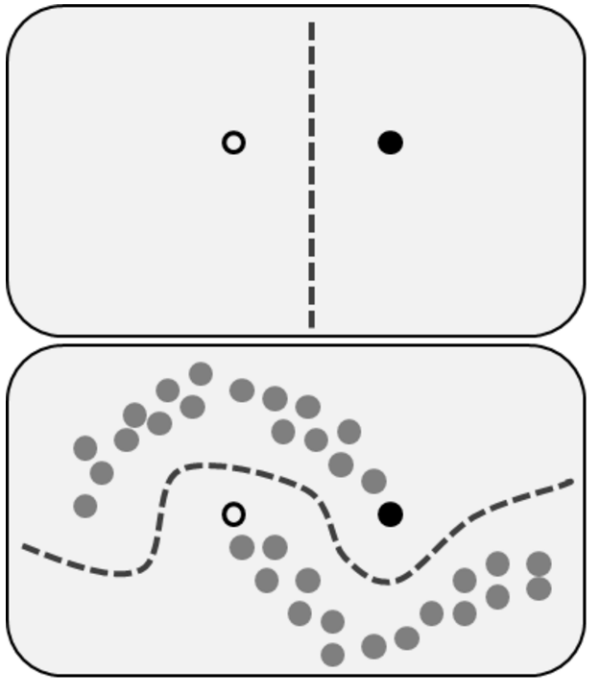
T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Springer, 2001.

3. Дополнительные темы
   - SVM в задаче SemiSupervised
   - Робастные модели

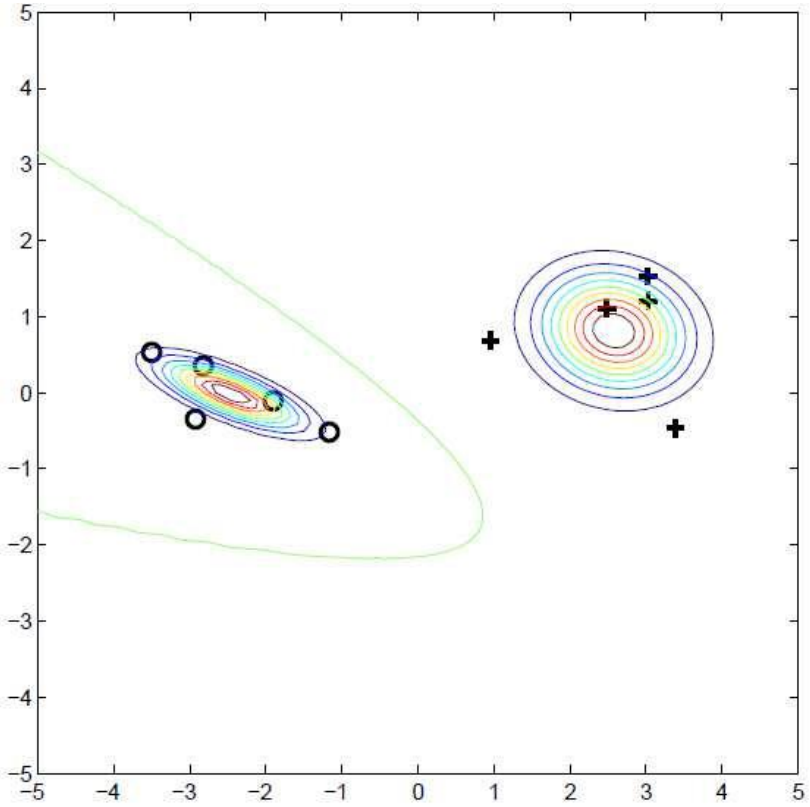
Semi-supervised обучение: мотивация

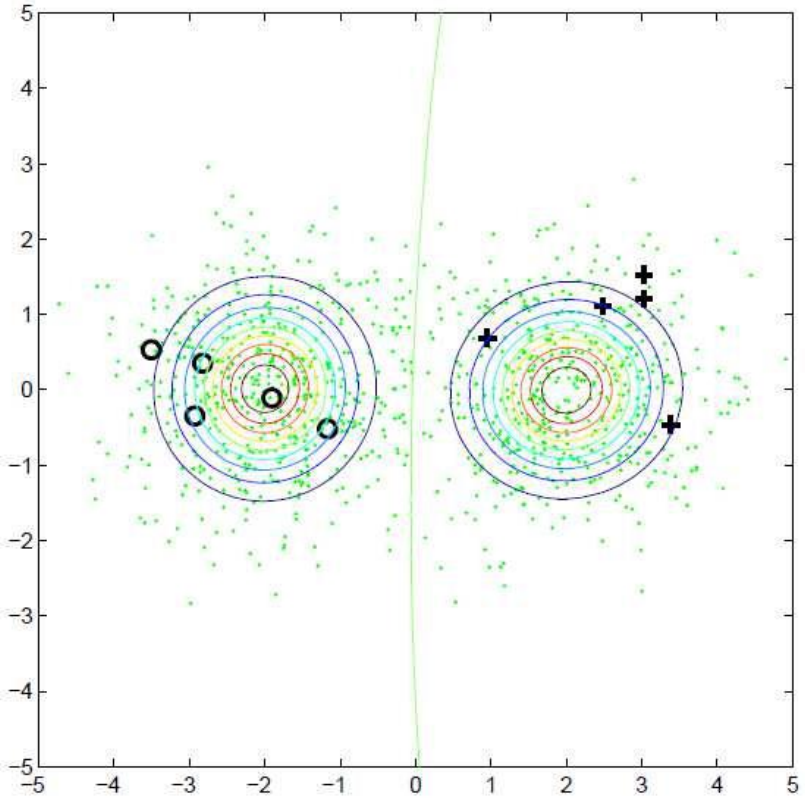


Semi-supervised обучение: мотивация

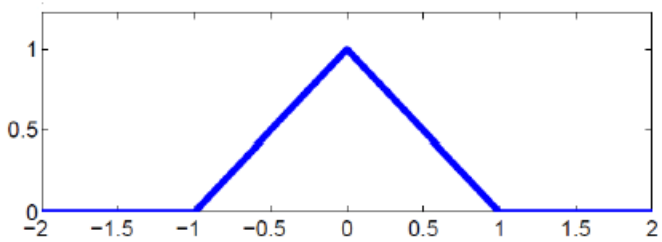


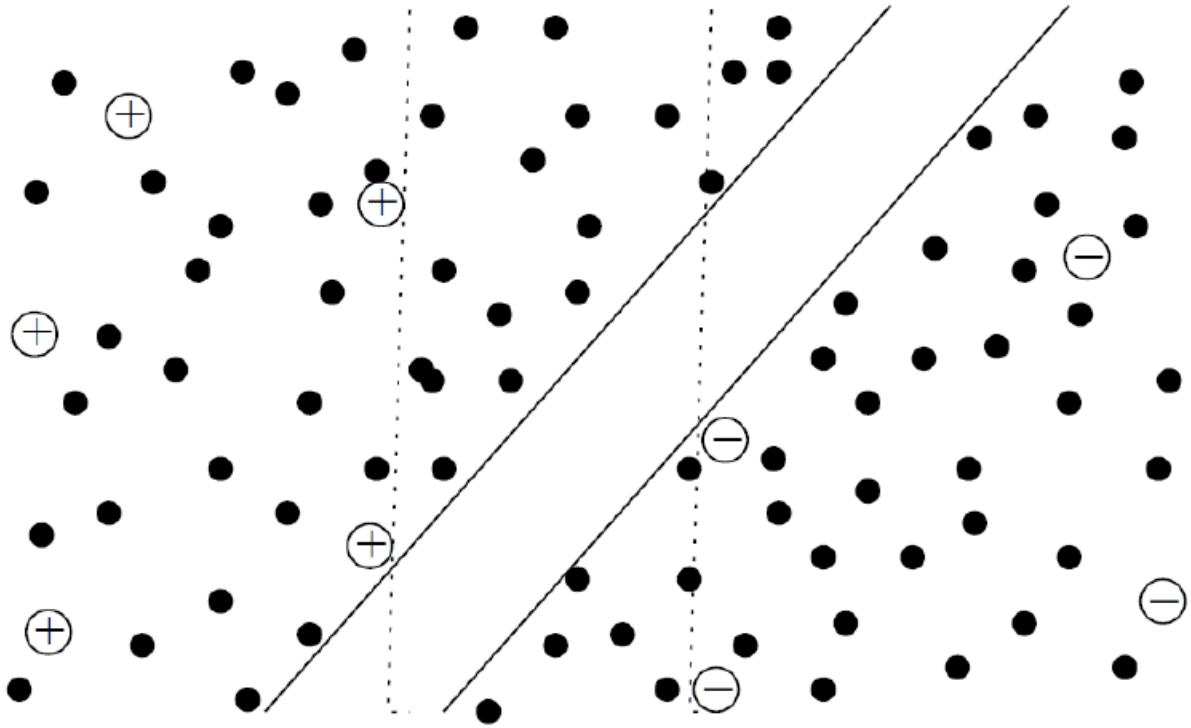
Semi-supervised обучение: мотивация





Функция потерь $\mathscr{L}(M) = (1 - |M|)_+$ штрафует за попадание объекта внутрь разделяющей полосы.





Обучение весов w, w_0 по частично размеченной выборке:

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

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

Эффективная реализация:

Sindhwani, Keerthi. Large scale semisupervised linear SVMs. SIGIR 2006.

Гауссовская функция штрафа:

Chapelle, Zien. Semi-supervised classification by low density separation. AISTAT 2005.

Недостатки TSVM:

- решение неустойчиво, если нет области разреженности;
- требуется настройка двух параметров C, γ;

Semi-supervised SVM (S3VM)

SVM:

$$\sum_{i=1}^{l} \max\{0; 1 - y_i < w, x_i >\} \; + \; \alpha ||w||_{l_2}^2 \to \min_{w}$$

Semi-supervised SVM (S3VM)

SVM:

$$\sum_{i=1}^{l} \max\{0; 1 - y_i < w, x_i >\} \; + \; \alpha ||w||_{l_2}^2 \to \min_w$$

Идея:

$$y_i < w, x_i > \; \to \; a(x_i) < w, x_i >=$$
$$= sign\{< w, x_i >\} < w, x_i >= | < w, x_i > |$$

Semi-supervised SVM (S3VM)

SVM:

$$\sum_{i=1}^{l} \max\{0; 1 - y_i < w, x_i >\} \; + \; \alpha ||w||_{l_2}^2 \to \min_w$$

Идея:

$$y_i < w, x_i > \; \to \; a(x_i) < w, x_i >=$$
$$= sign\{< w, x_i >\} < w, x_i >= | < w, x_i > |$$

$$\sum_{i=1}^{l} \max\{0; 1 - y_i < w, x_i >\} \; + \; \beta \sum_{i=l+1}^{l+u} \max\{0; 1 - | < w, x_i > |\} \; + \alpha ||w||_{l_2}^2$$

Робастные модели

Линейные модели

- Linear Regression $\min_w ||Xw - y||_2{}^2$
- Ridge $\min_w ||Xw - y||_2{}^2 + \alpha||w||_2{}^2$
- LASSO $\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha||w||_1$
- Multi-task LASSO $\min_w \frac{1}{2n_{samples}}||XW - Y||_{Fro}^2 + \alpha||W||_{21}$

$$||A||_{Fro} = \sqrt{\sum_{ij} a_{ij}^2} \qquad ||A||_{21} = \sum_i \sqrt{\sum_j a_{ij}^2}$$

Другие модели

- Elastic Net

$$\min_{w} \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

- Multi-task Elastic Net

$$\min_{W} \frac{1}{2n_{samples}}||XW - Y||_{Fro}^2 + \alpha\rho||W||_{21} + \frac{\alpha(1-\rho)}{2}||W||_{Fro}^2$$

- OMP

$$\arg\min ||y - X\gamma||_2^2 \text{ subject to } ||\gamma||_0 \leq n_{nonzero\_coefs}$$

- Logistic Regression

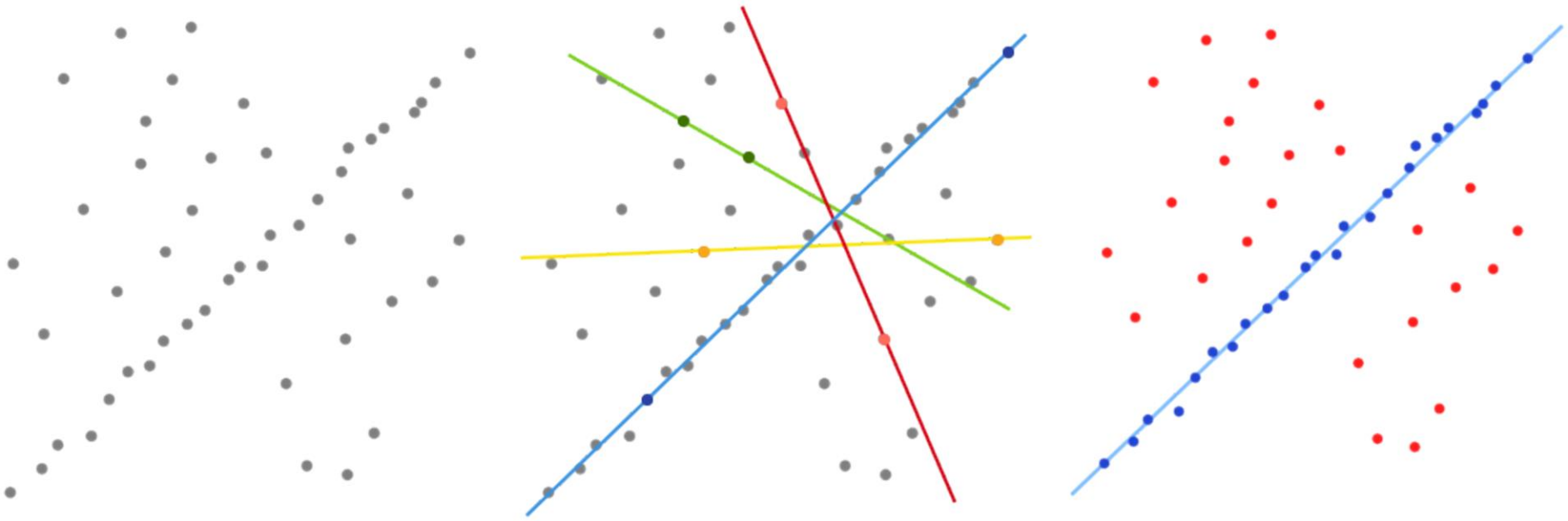
$$\min_{w,c} ||w||_1 + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

$$\min_{w,c} \frac{1}{2}w^T w + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

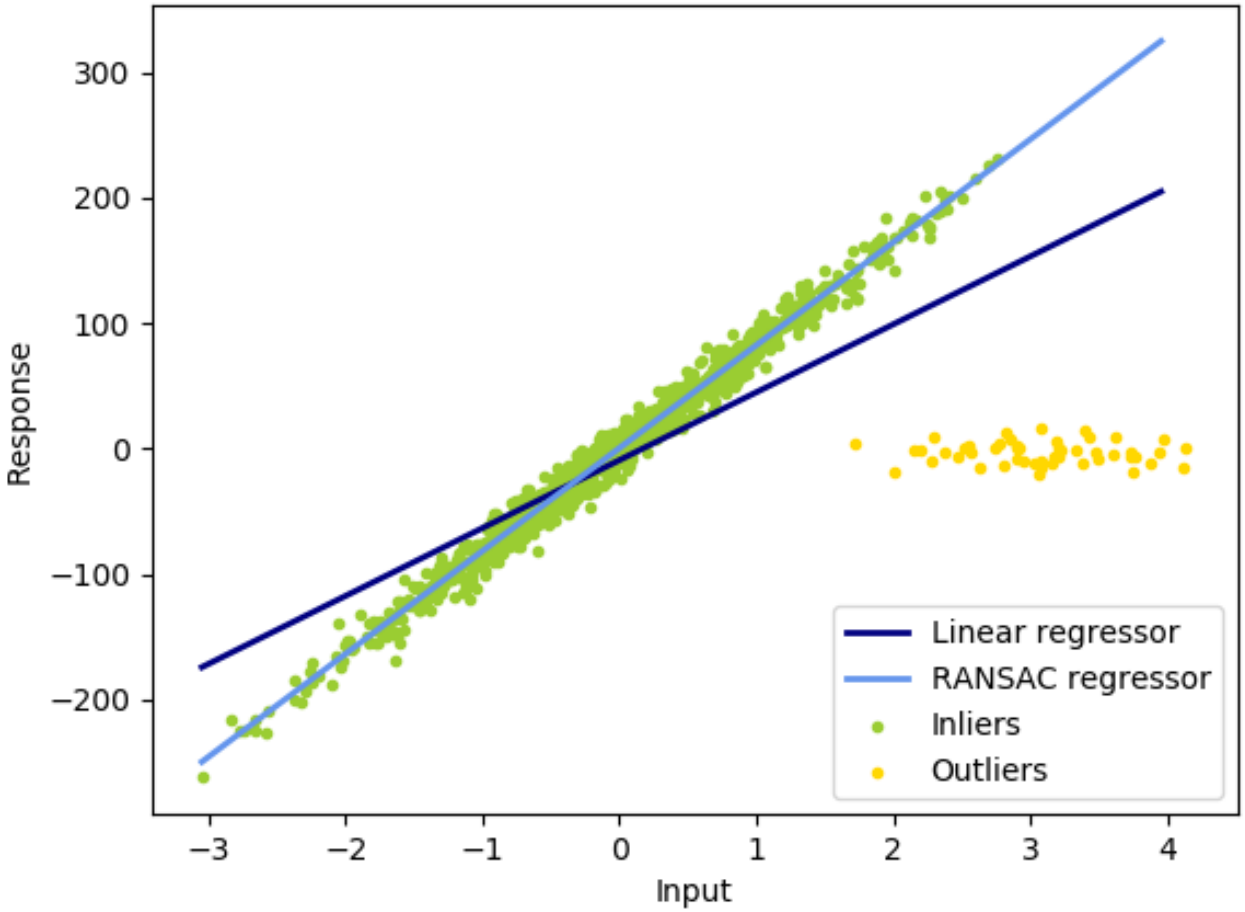
Робастные модели в linear_model

- RANSACRegressor
- HuberRegressor
- Theil-Sen Regressor

RANSACRegressor



RASCANRegressor (RANdom SAmple Consensus)



http://scikit-learn.org/stable/modules/linear_model.html#ransac-regression

HuberRegressor

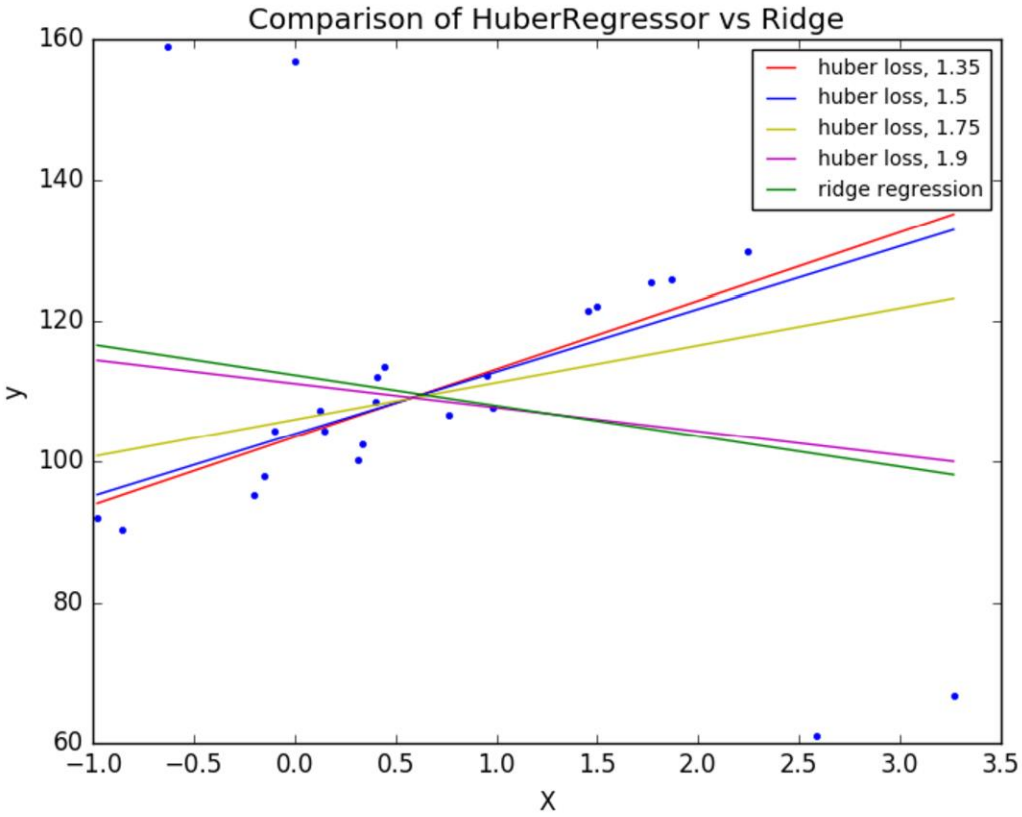
$$\min_{w,\sigma} \sum_{i=1}^{n} \left(\sigma + H_m \left(\frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha ||w||_2{}^2$$

$$H_m(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$

HuberRegressor

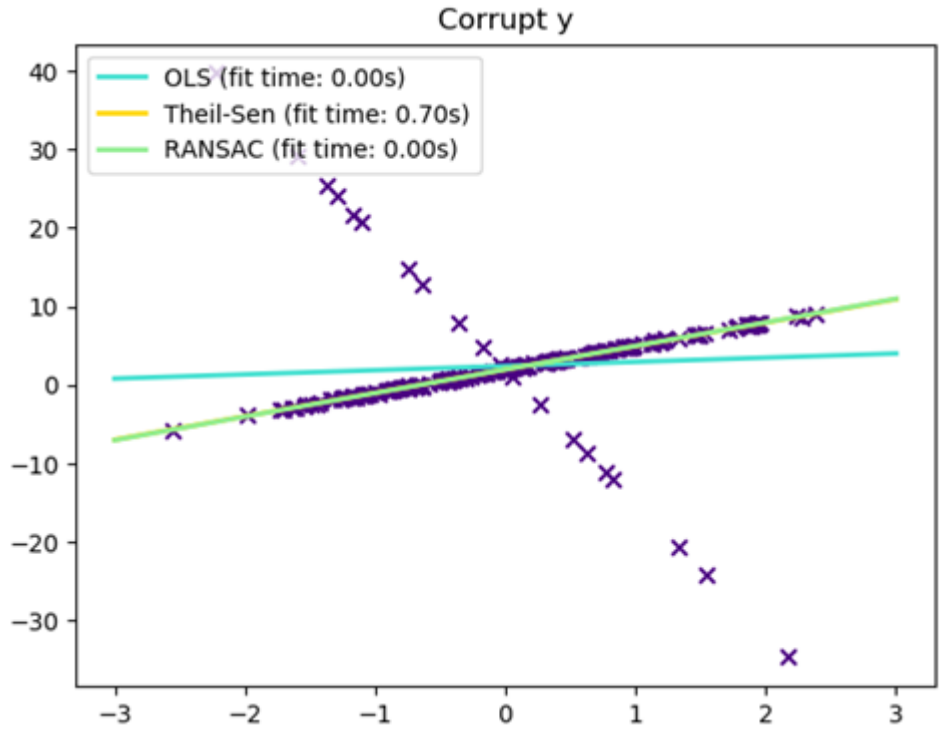
$$\min_{w,\sigma} \sum_{i=1}^{n} \left(\sigma + H_m \left(\frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha ||w||_2{}^2$$

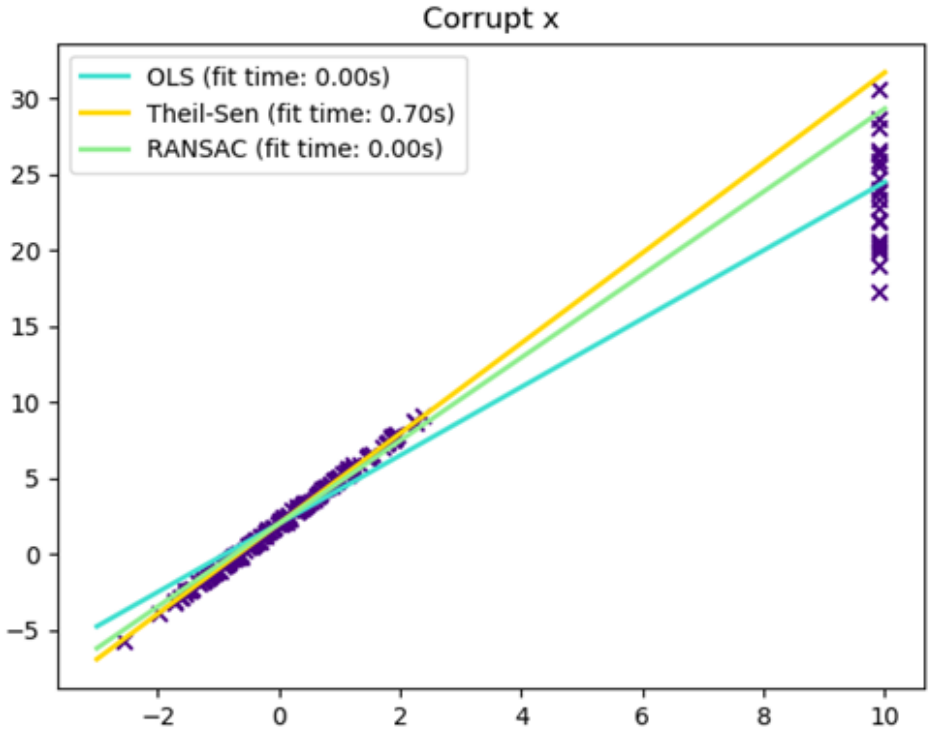
$$H_m(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$



http://scikit-learn.org/stable/modules/linear_model.html#ransac-regression

Theil-Sen estimator (median of slope)





http://scikit-learn.org/stable/auto_examples/linear_model/plot_theilsen.html

Резюме

1. SVD – очень сильный алгебраический приём, позволяющий решать линейные уравнения

2. Линейная регрессия – простой метод,

3. МНК часто приводит к переобучению из-за мультиколлинеарности

4. Гребневая регрессия и Лассо Тибширани – простые модификации

5. Задачи, где очень много неразмеченных данных, часто решаются методами semi-supervised learning

Отзывы о лекции: https://goo.gl/forms/zeZiu1fSgrpPGp6T2