Мат.модели машинного обучения: линейные модели регрессии и метод главных компонент

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 11 октября 2024

Содержание

Метод наименьших квадратов (МНК)

- 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;
- $a(x, w)$ — модель зависимости,
  $w \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

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

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

Многомерная линейная регрессия

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

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

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:

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

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

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

откуда следует *нормальная система* задачи МНК:

$$F^{\mathsf{T}}Fw = F^{\mathsf{T}}y,$$

где $F^{\mathsf{T}}F$ — матрица размера $n \times n$ полного ранга n.

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

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

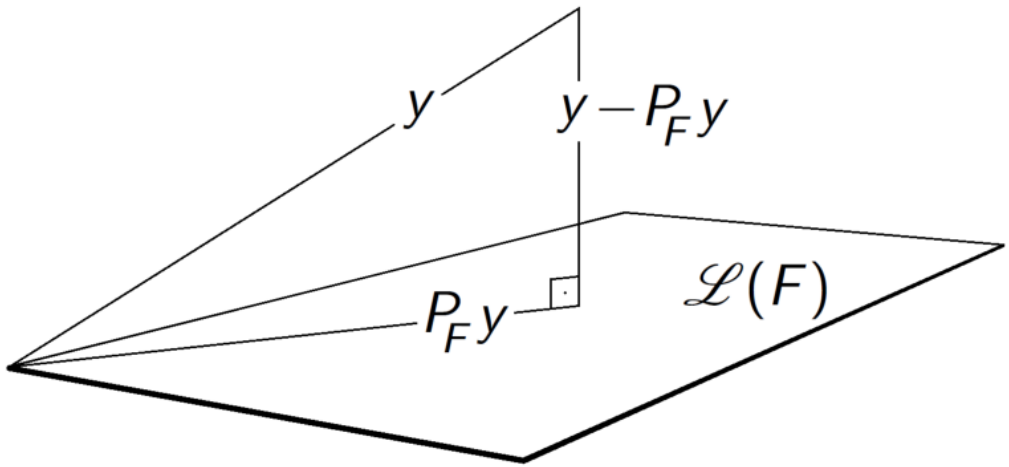
где $P_F = FF^{+} = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Геометрическая интерпретация МНК

Линейная оболочка столбцов матрицы $F = (f_1, \ldots, f_n)$, $f_j \in \mathbb{R}^\ell$:

$$\mathscr{L}(F) = \left\{ \sum_{j=1}^{n} w_j f_j \;\middle|\; w \in \mathbb{R}^n \right\}$$

$P_F = F(F^\mathsf{T}F)^{-1}F^\mathsf{T}$ — проекционная матрица
$P_F y$ — проекция вектора $y \in \mathbb{R}^\ell$ на подпространство $\mathscr{L}(F)$
$(I_\ell - P_F)y$ — проекция y на его ортогональное дополнение



МНК — это опускание перпендикуляра в $\mathbb{R}^\ell$ из y на $\mathscr{L}(F)$

Сингулярное разложение

Произвольная $\ell \times n$-матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^{\mathsf{T}}.$$

Основные свойства сингулярного разложения:

1. $\ell \times n$-матрица $V = (v_1, \ldots, v_n)$ ортогональна, $V^{\mathsf{T}}V = I_n$, столбцы v_j — собственные векторы $\ell \times \ell$-матрицы FF^{T};
2. $n \times n$-матрица $U = (u_1, \ldots, u_n)$ ортогональна, $U^{\mathsf{T}}U = I_n$, столбцы u_j — собственные векторы $n \times n$-матрицы $F^{\mathsf{T}}F$;
3. $n \times n$-матрица D диагональна, $D = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)$, $\lambda_j \geqslant 0$ — общие собственные значения матриц $F^{\mathsf{T}}F$ и FF^{T}.

Решение МНК через сингулярное разложение

Псевдообратная $F^+ = (F^\intercal F)^{-1} F^\intercal$, вектор МНК-решения w^*, МНК-аппроксимация целевого вектора Fw^*:

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

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

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

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

Тождества: $(AB)^{-1} = B^{-1}A^{-1}$, $(AB)^\intercal = B^\intercal A^\intercal$, $\|w\|^2 = w^\intercal w$

Мультиколлинеарность и число обусловленности матрицы

Если $\exists u \in \mathbb{R}^n$: $\underset{n\times n}{S}\, u \approx 0$, то некоторые с.з. S близки к нулю

Число обусловленности $n\times n$-матрицы S:

$$\mu(S) = \|S\|\|S^{-1}\| = \frac{\max\limits_{u:\,\|u\|=1} \|Su\|}{\min\limits_{u:\,\|u\|=1} \|Su\|} = \frac{\lambda_{\max}}{\lambda_{\min}}$$

При умножении обратной матрицы на вектор, $z = S^{-1}u$, относительная погрешность усиливается в $\mu(S)$ раз:

$$\frac{\|\delta z\|}{\|z\|} \leqslant \mu(S)\frac{\|\delta u\|}{\|u\|}$$

В нашем случае: $S = F^\intercal F$, $u = F^\intercal y$, $w^* = S^{-1}u = (F^\intercal F)^{-1}F^\intercal y$, погрешности измерения признаков $f_j(x_i)$ и ответов y_i усиливаются в $\mu(F^\intercal F)$ раз!

Стратегии устранения мультиколлинеарности

Если матрица $S = F^\mathsf{T}F$ плохо обусловлена, то:

- решение w^* неустойчиво и плохо интерпретируемо, содержит большие по модулю w_j^* разных знаков;
- $\|w^*\|$ велико;
- возникает переобучение:
  на обучении $Q(w^*, X^\ell) = \|Fw^* - y\|^2$ мало́;
  на контроле $Q(w^*, X^k) = \|F'w^* - y'\|^2$ велико;

Стратегии устранения мультиколлинеарности и переобучения:

1. регуляризация: $\|w\| \to \min$;
2. отбор признаков: $f_1, \ldots, f_n \to f_{j_1}, \ldots, f_{j_m}$, $m \ll n$.
3. преобразование признаков: $f_1, \ldots, f_n \to g_1, \ldots, g_m$, $m \ll n$;

Гребневая регрессия (ridge regression)

Штраф за увеличение L_2-нормы вектора весов $\|w\|$:

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

где τ — неотрицательный *параметр регуляризации*.

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

$$\frac{\partial Q_\tau(w)}{\partial w} = 2F^\intercal(Fw - y) + 2\tau w = 0$$
$$w_\tau^* = (F^\intercal F + \tau I_n)^{-1} F^\intercal y.$$

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

Регуляризованный МНК через сингулярное разложение

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

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

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

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

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

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x'_i, y'_i)_{i=1}^k$;

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1(x'_1) & \ldots & f_n(x'_1) \\ \ldots & \ldots & \ldots \\ f_1(x'_k) & \ldots & f_n(x'_k) \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y'_1 \\ \ldots \\ y'_k \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O(kn^2 + knT)$ операций:

$$Q(w^*_\tau, X^k) = \|F'w^*_\tau - y'\|^2 = \Big\| \underbrace{F'U}_{k\times n} \operatorname{diag}\Big(\frac{\sqrt{\lambda_j}}{\lambda_j+\tau}\Big) \underbrace{V^{\mathsf{T}}y}_{n\times 1} - y' \Big\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Регуляризация сокращает «эффективную размерность»

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

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

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

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

При использовании регуляризации:

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

Тождество: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

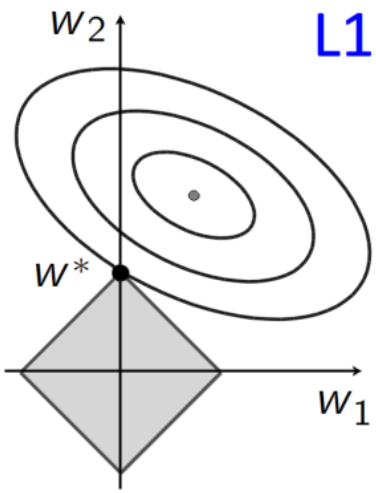
Регуляризация по L_1-норме для отбора признаков

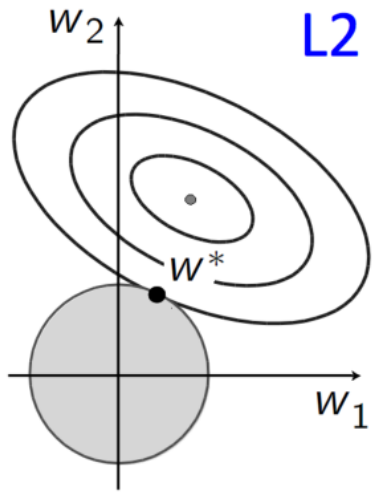
LASSO — Least Absolute Shrinkage and Selection Operator

$$\|Fw - y\|^2 + \mu \sum_{j=1}^{n} |w_j| \to \min_w \quad \Longleftrightarrow \quad \begin{cases} \|Fw - y\|^2 \to \min\limits_w; \\ \sum\limits_{j=1}^{n} |w_j| \leqslant \varkappa; \end{cases}$$

LASSO (L_1):

$$\sum_{j=1}^{n} |w_j| \leqslant \varkappa$$





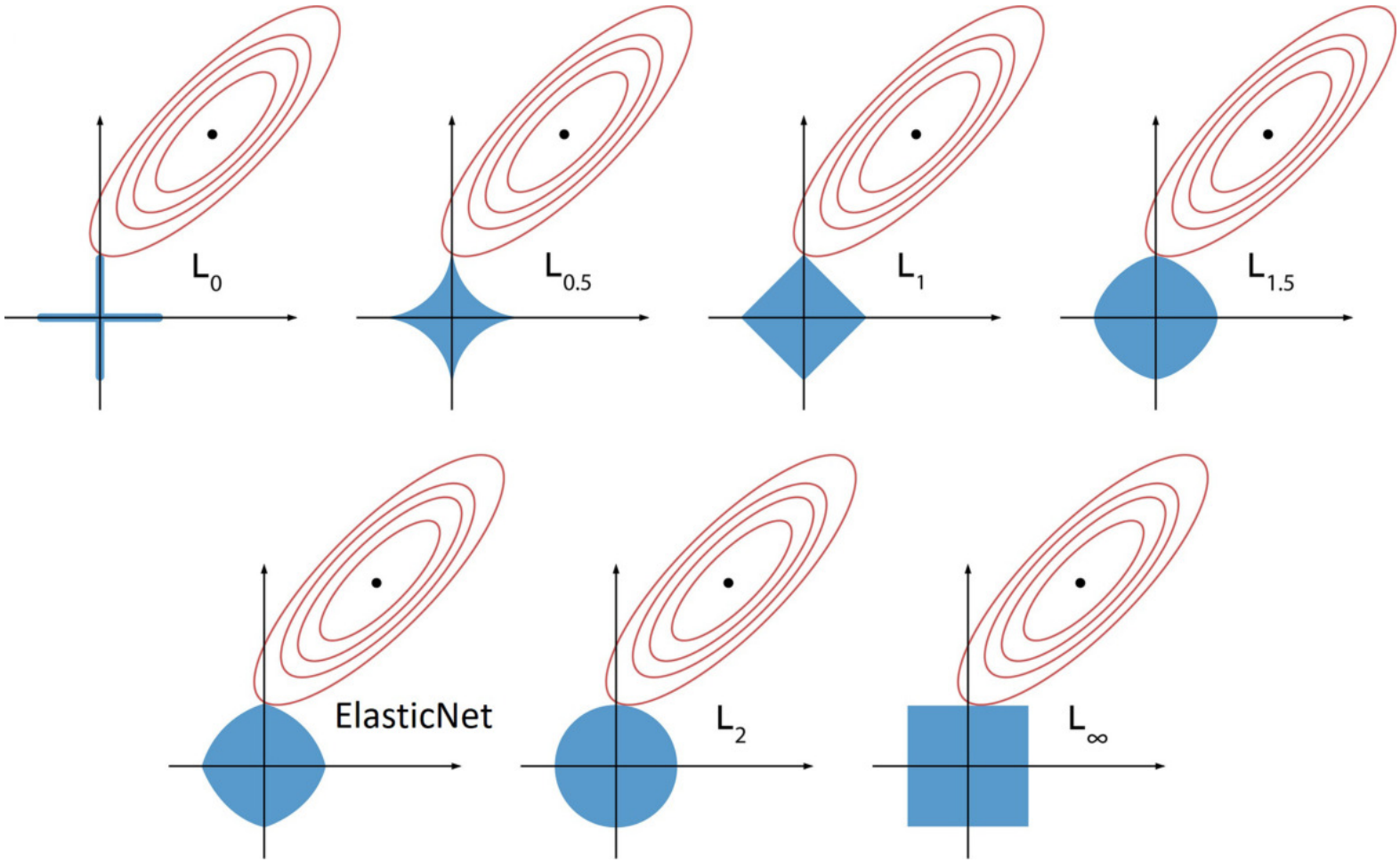
Ridge (L_2):

$$\sum_{j=1}^{n} w_j^2 \leqslant \varkappa$$

T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. 2017.

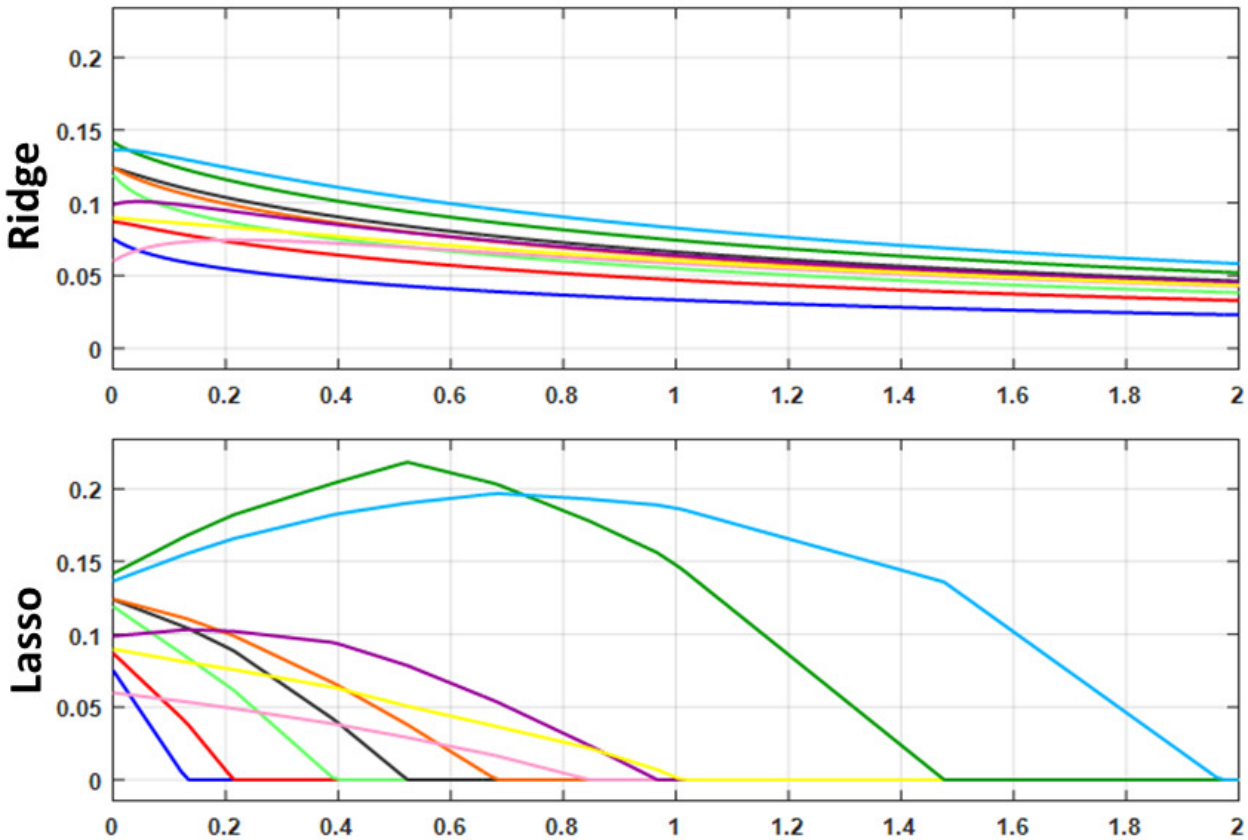
Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным L_p-нормам:



Сравнение L_2 (Ridge) и L_1 (LASSO) регуляризации

Типичный вид зависимости весов w_j от селективности μ



В LASSO с увеличением μ усиливается отбор признаков

Негладкие регуляризаторы для отбора и группировки признаков

Общий вид регуляризаторов (μ — параметр селективности):

$$\sum_{i=1}^{\ell} \mathscr{L}(w, x_i) + \sum_{j=1}^{n} R_\mu(w_j) \;\to\; \min_w .$$

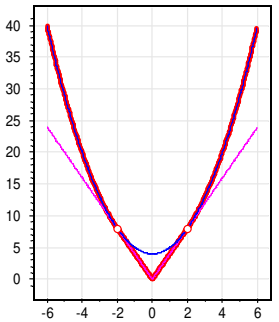
Регуляризаторы с эффектами отбора и группировки признаков:

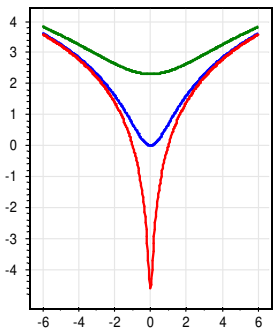
LASSO (L_1): $R_\mu(w) = \mu|w|$

Elastic Net: $R_\mu(w) = \mu|w| + \tau w^2$

Support Feature Machine (SFM):

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





Relevance Feature Machine (RFM):

$$R_\mu(w) = \ln\left(\mu w^2 + 1\right)$$

Метод главных компонент (Principal Component Analysis, PCA)

$f_1(x), \ldots, f_n(x)$ — исходные числовые признаки;
$g_1(x), \ldots, g_m(x)$ — новые числовые признаки, $m \leqslant n$;

Требование: старые признаки $f_j(x)$ должны линейно восстанавливаться по новым признакам $g_s(x)$:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1, \ldots, n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке $x_1, \ldots, x_\ell$:

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n} \big(\hat{f}_j(x_i) - f_j(x_i)\big)^2 \to \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

Задача преобразования признаков (feature transformation) — это задача обучения без учителя, тут нет ответов y_i

Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}; \quad \underset{\ell\times m}{G} = \begin{pmatrix} g_1(x_1) & \ldots & g_m(x_1) \\ \ldots & \ldots & \ldots \\ g_1(x_\ell) & \ldots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \ldots & u_{1m} \\ \ldots & \ldots & \ldots \\ u_{n1} & \ldots & u_{nm} \end{pmatrix}; \qquad \hat{F} = GU^{\mathsf{T}} \overset{\text{хотим}}{\approx} F.$$

Найти: сразу и новые признаки G, и преобразование U:

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n}\big(\hat{f}_j(x_i) - f_j(x_i)\big)^2 = \big\|GU^{\mathsf{T}} - F\big\|^2 \to \min_{G,U},$$

Основная теорема метода главных компонент

Теорема

Если $m \leqslant \operatorname{rk} F$, *то минимум* $\left\|GU^{\mathsf{T}} - F\right\|^2$ *достигается, когда столбцы* U *— это с.в. матрицы* $F^{\mathsf{T}}F$, *соответствующие* m *максимальным с.з.* $\lambda_1, \ldots, \lambda_m$, *а матрица* $G = FU$.

При этом:

1. *матрица* U *ортонормирована:* $U^{\mathsf{T}}U = I_m$;
2. *матрица* G *ортогональна:* $G^{\mathsf{T}}G = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$;
3. $U\Lambda = F^{\mathsf{T}}FU$; $G\Lambda = FF^{\mathsf{T}}G$;
4. $\left\|GU^{\mathsf{T}} - F\right\|^2 = \|F\|^2 - \operatorname{tr}\Lambda = \sum\limits_{j=m+1}^{n} \lambda_j$.

Тождество: $\|F\|^2 = \operatorname{tr}(F^{\mathsf{T}}F) = \sum_{j=1}^{n} \lambda_j$

Связь с сингулярным разложением

Произвольная $\ell \times n$-матрица F представима в виде SVD:

$$F = VDU^{\mathsf{T}}; \quad U^{\mathsf{T}}U = I_m; \quad V^{\mathsf{T}}V = I_m; \quad D = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)$$

Если взять $m = n$, то:

1. $\left\|GU^{\mathsf{T}} - F\right\|^2 = 0$
2. представление $\hat{F} = GU^{\mathsf{T}} = F$ точное и совпадает с SVD, если положить $G = V\sqrt{\Lambda}$, $D = \sqrt{\Lambda}$:
$$F = GU^{\mathsf{T}} = V\sqrt{\Lambda}U^{\mathsf{T}};$$
3. линейное преобразование U работает в обе стороны:
$$F = GU^{\mathsf{T}}; \quad G = FU.$$

$G^{\mathsf{T}}G = \Lambda$ — новые признаки некоррелированы,
U — *декоррелирующе преобразование* Карунена–Лоэва

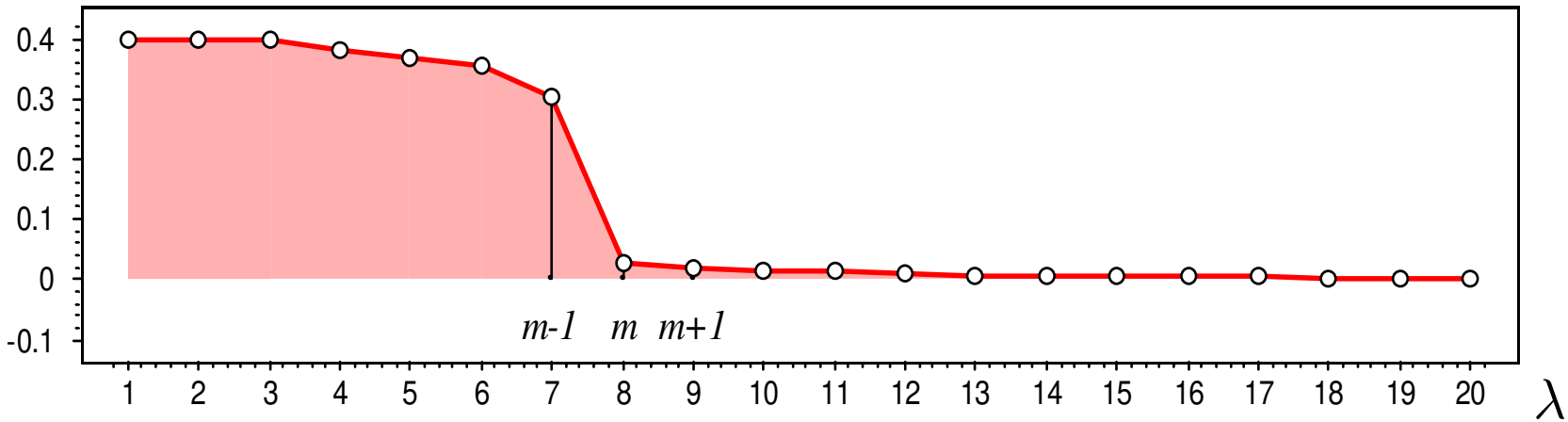
Эффективная размерность выборки

Упорядочим с.з. $F^{\mathsf{T}}F$ по убыванию: $\lambda_1 \geqslant \ldots \geqslant \lambda_n \geqslant 0$.

Эффективная размерность выборки — это наименьшее целое m, при котором

$$E_m = \frac{\|GU^{\mathsf{T}} - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leqslant \varepsilon.$$

Критерий «крутого склона»: находим m: $E_{m-1} \gg E_m$:



Решение задачи НК в новых признаках

Заменим $\underset{\ell\cdot n}{F}$ на её приближение $\underset{\ell\cdot m}{G}\cdot\underset{m\cdot n}{U^{\mathsf{T}}}$, предполагая $m \leqslant n$:

$$\|G\underbrace{U^{\mathsf{T}}w}_{z} - y\|^2 = \|Gz - y\|^2 \to \min_z.$$

Связь нового и старого вектора коэффициентов:

$$z = U^{\mathsf{T}}w; \qquad w = Uz.$$

Решение задачи наименьших квадратов относительно z:

$$z^* = D^{-1}V^{\mathsf{T}}y; \qquad w^* = UD^{-1}V^{\mathsf{T}}y = \sum_{j=1}^{m}\frac{1}{\sqrt{\lambda_j}}u_j(v_j^{\mathsf{T}}y);$$

$$Gz^* = VV^{\mathsf{T}}y = \sum_{j=1}^{m} v_j(v_j^{\mathsf{T}}y);$$

Единственное отличие от прежнего w^* — m слагаемых вместо n

Спектральный метод наименьших квадратов

1. Построить SVD-разложение, упорядочить $\lambda_1 \geqslant \cdots \geqslant \lambda_n$
2. Отделить $n - m$ наименьших с. з. от нуля: $\lambda'_j := \lambda_j + \delta_j$

Частные случаи:

- $\lambda'_j := \lambda_j + \tau$ — гребневая регрессия
- $\lambda'_j := \lambda_j + \infty[j > m]$ — метод главных компонент
- $\lambda'_j := \lambda_j + \tau[j > m]$ — нечто промежуточное

3. Применить формулы SVD для модификации МНК-решения:

$$w^* = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j(v_j^{\mathsf{T}} y) \quad \longrightarrow \quad w^* = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda'_j} u_j(v_j^{\mathsf{T}} y)$$

$$Fw^* = \sum_{j=1}^{n} v_j(v_j^{\mathsf{T}} y) \quad \longrightarrow \quad Fw^* = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda'_j} v_j(v_j^{\mathsf{T}} y)$$

Задачи низкорангового матричного разложения

- Генерация новых признаков $f_1, \ldots, f_n \to g_1, \ldots, g_m, \quad m \ll n$;
- Снижение размерности вектора признаков
- Восстановление пропущенных значений (missing values)

Дано: матрица $F = (f_{ij})_{\ell \times n}$, $(i,j) \in \Omega \subseteq \{1..\ell\} \times \{1..n\}$

Найти: матрицы $G = (g_{is})_{\ell \times m}$ и $U^{\mathsf{T}} = (u_{sj})_{m \times n}$

Критерий: $\left\| F - GU^{\mathsf{T}} \right\|^2 = \sum_{(i,j) \in \Omega} \left(f_{ij} - \sum_s g_{is} u_{sj} \right)^2 \to \min_{X,Y}$

Когда вместо SVD применяется SG или другие методы:

- разреженные данные: $|\Omega| \ll \ell n$
- неквадратичная функция потерь
- неотрицательное матричное разложение: $g_{is} \geqslant 0, \quad u_{sj} \geqslant 0$

Примеры задач неотрицательного матричного разложения

1. **Выявление интересов в рекомендательных системах (recommender systems, collaborative filtering)**

$$f_{iu} = \sum_t p_{it} q_{tu}$$

дано: f_{iu} — рейтинги товаров i, поставленные пользователем u
найти: p_{it} — профиль интересов товара i
q_{tu} — профиль интересов пользователя u

2. **Разделение смеси химических веществ по данным жидкостной хроматографии**

$$f_{t\lambda} = \sum_i c_{ti} s_{i\lambda}$$

дано: $f_{t\lambda}$ — концентрация на выходе УФ-детектора
найти: c_{ti} — хроматограмма i-го вещества, t — время
$s_{i\lambda}$ — спектр i-го вещества, λ — длина волны

Примеры задач неотрицательного матричного разложения

3. **Латентный семантический анализ коллекций текстов (тематическое моделирование)**

$$f_{wd} = \sum_t \varphi_{wt}\theta_{td}$$

дано: $f_{wd} = p(w|d)$ — частоты слов w в документах d
найти: $\varphi_{wt} = p(w|t)$ — распределения слов w в темах t
$\theta_{td} = p(t|d)$ — распределения тем t в документах d

4. **Оценивание экспрессии генов по данным ДНК-микрочипов с учётом кросс-гибридизации**

$$f_{pk} = \sum_g a_{pg} c_{gk}$$

дано: f_{pk} — 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

Резюме в конце лекции

- Многомерная линейная регрессия
  — через *сингулярное разложение*
- Три приёма против мультиколлинеарности и переобучения
  — регуляризация, отбор и преобразование признаков
- L_2-регуляризация, она же гребневая регрессия
  — тоже через *сингулярное разложение*
- L_1-регуляризация (LASSO) и др. негладкие регуляризаторы
  — регулируемый отбор признаков
- Метод главных компонент — задача матричного разложения
  — снова через *сингулярное разложение*
- Другие методы матричных разложений и их приложения
  — в следующем семестре