Методы машинного обучения
Многомерная линейная регрессия и метод главных компонент

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

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-21-22
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 26 сентября 2023

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;
  $y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i \bigl(f(x_i, \alpha) - y_i\bigr)^2 \to \min_{\alpha},$$

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

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

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — числовые признаки;

Модель многомерной линейной регрессии:

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

Матричные обозначения:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell\times 1}{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \underset{n\times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \bigl(f(x_i,\alpha) - y_i\bigr)^2 = \|F\alpha - y\|^2 \to \min_\alpha .$$

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

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

откуда следует *нормальная система* задачи МНК:

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

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

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

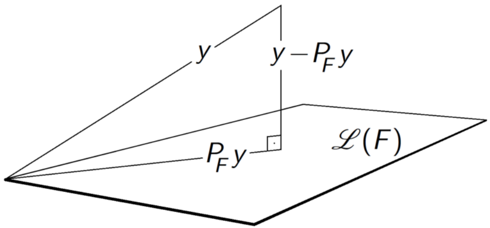
Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,
где $P_F = FF^+ = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Геометрическая интерпретация МНК

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

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

$P_F = F(F^\top F)^{-1} F^\top$ — 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, $D = \mathrm{diag}\bigl(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\bigr)$, $\lambda_j \geqslant 0$ — 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 α^*, МНК-аппроксимация целевого вектора $F\alpha^*$:

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

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

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

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

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

Проблема мультиколлинеарности

Если $\exists\gamma \in \mathbb{R}^n$: $F\gamma \approx 0$, то некоторые λ_j близки к нулю

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

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

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: $\alpha^* = (F^{\mathsf T}F)^{-1}F^{\mathsf T}y$, $S = F^{\mathsf T}F$, $u = F^{\mathsf T}y$, погрешности измерения признаков $f_j(x_i)$ и ответов y_i усиливаются в $\mu(F^{\mathsf T}F)$ раз!

Проблема мультиколлинеарности и переобучения

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

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

Стратегии устранения мультиколлинеарности и переобучения:

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

Гребневая регрессия (ridge regression)

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

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \tfrac{1}{\sigma}\|\alpha\|^2,$$

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

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

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

$$\frac{\partial Q_\tau(\alpha)}{\partial \alpha} = 2F^\intercal(F\alpha - y) + 2\tau\alpha = 0$$

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

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

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α_τ^*
и МНК-аппроксимация целевого вектора $F\alpha_\tau^*$:

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

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

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

$F\alpha_\tau^* \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x_i', y_i')_{i=1}^k$;

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1(x_1') & \dots & f_n(x_1') \\ \dots & \dots & \dots \\ f_1(x_k') & \dots & f_n(x_k') \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y_1' \\ \dots \\ y_k' \end{pmatrix}.$$

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

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

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

Регуляризация сокращает «эффективную размерность»

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

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

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:

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

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

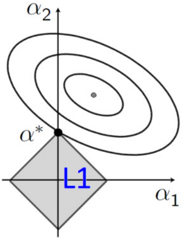
Регуляризация по L_1-норме для отбора признаков

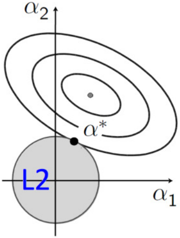
LASSO — Least Absolute Shrinkage and Selection Operator

$$\|F\alpha - y\|^2 + \mu \sum_{j=1}^{n} |\alpha_j| \to \min_{\alpha} \quad \Longleftrightarrow \quad \begin{cases} \|F\alpha - y\|^2 \to \min\limits_{\alpha}; \\ \sum\limits_{j=1}^{n} |\alpha_j| \leqslant \varkappa; \end{cases}$$

LASSO (L_1):

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





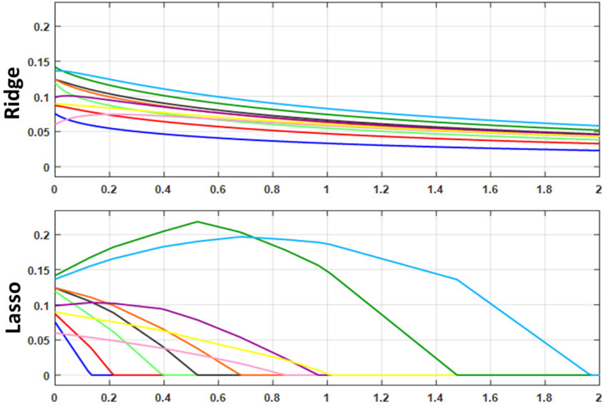
Ridge (L_2):

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

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

Сравнение L_2 (Ridge) и L_1 (LASSO) регуляризации

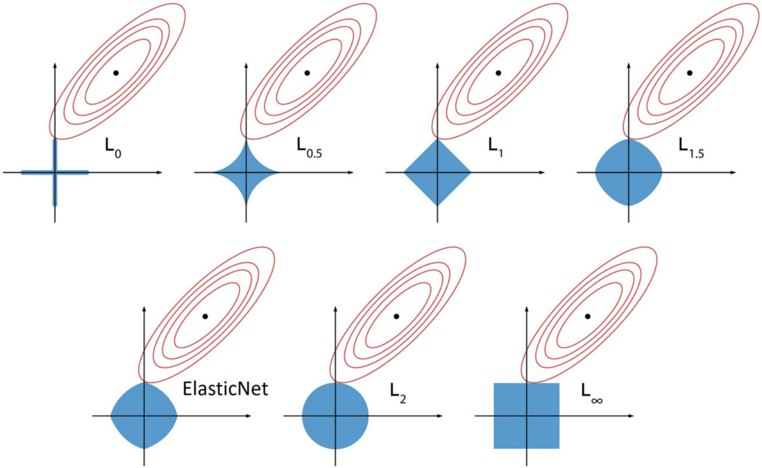
Типичный вид зависимости весов α_j от селективности μ



В LASSO с увеличением μ усиливается отбор признаков

Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным L_p-нормам:



Метод главных компонент (Principal Component Analysis, PCA)

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

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

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

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

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

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

Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

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

Матрица линейного преобразования новых признаков в старые:

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

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

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

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

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 сингулярным разложением при $G = V\sqrt{\Lambda}$:
$$F = GU^{\mathsf{T}} = V\sqrt{\Lambda}U^{\mathsf{T}}; \quad U^{\mathsf{T}}U = I_m; \quad V^{\mathsf{T}}V = I_m.$$
3. линейное преобразование U работает в обе стороны:
$$F = GU^{\mathsf{T}}; \quad G = FU.$$
Поскольку новые признаки некоррелированы $(G^{\mathsf{T}}G = \Lambda)$, преобразование U называется *декоррелирующим* (или преобразованием Карунена–Лоэва).

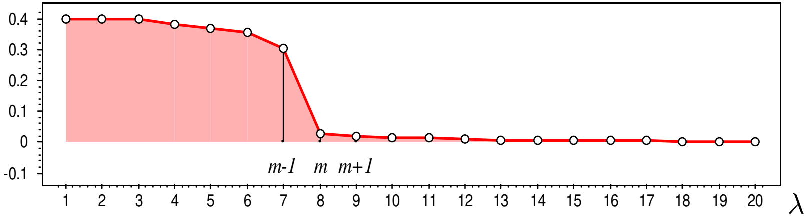
Эффективная размерность выборки

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 $m \leqslant n$:

$$\|G\underbrace{U^{\mathsf{T}}\alpha}_{\beta} - y\|^2 = \|G\beta - y\|^2 \to \min_{\beta}.$$

Связь нового и старого вектора коэффициентов:

$$\beta = U^{\mathsf{T}}\alpha; \qquad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно β (единственное отличие — m слагаемых вместо n):

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

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

Спектральный метод наименьших квадратов

1. Построить SVD-разложение, упорядочить $\lambda_1 \geqslant \dots \geqslant \lambda_n$
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:

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

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

Задачи низкорангового матричного разложения

- Понижение размерности в задачах регрессии/классификации
- Генерация новых признаков
- Формирование сжатого представления данных

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 $U^{\mathsf{T}} = (u_{sj})_{m\times n}$ такие, что

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

Дополнительные ограничения, вынуждающие отказаться от SVD:

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

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