Методы машинного обучения
Многомерная линейная регрессия

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: k.vorontsov@iai.msu.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-24-25
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 17 сентября 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;
  $y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная зависимость;
- $a(x, w)$ — модель зависимости,
  $w \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

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

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

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — 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:

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

Функционал квадрата ошибки:

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

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 $n \times n$.

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

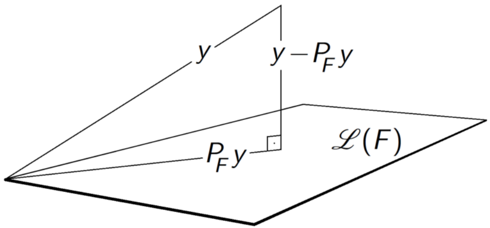
Значение функционала: $Q(w^*) = \|P_F y - y\|^2$,
где $P_F = FF^+ = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Геометрическая интерпретация МНК

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

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

$P_F = F(F^\top F)^{-1} F^\top$ — 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 $V = (v_1, \dots, v_n)$ ортогональна, $V^{\mathsf{T}} V = I_n$, столбцы v_j — собственные векторы $\ell \times \ell$-матрицы FF^{T};
2. $n \times n$-матрица $U = (u_1, \dots, u_n)$ 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, $D = \mathrm{diag}\bigl(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}\bigr)$, $\lambda_j \geqslant 0$ — 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 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 переобучение в линейных моделях

Возможные причины переобучения:

- слишком мало объектов, слишком много признаков
- линейная зависимость (мультиколлинеарность) признаков:

  линейная модель: $a(x, w) = \langle w, x \rangle$
  мультиколлинеарность: $\exists u \in \mathbb{R}^n\colon\ \forall x \in X\ \ \langle u, x \rangle = 0$
  неединственность решения: $\forall \gamma \in \mathbb{R}\ \ a(x, w) = \langle w + \gamma u, x \rangle$

Проявления переобучения:

- слишком большие веса $|w_j|$ разных знаков
- неустойчивость линейной модели $\langle w, x \rangle$
- $Q(X^\ell) \ll Q(X^k)$

Простой способ уменьшить переобучение:

- регуляризация $\|w\| \to \min$ (сокращение весов, weight decay)

Неустойчивость модели и число обусловленности матрицы

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

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

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

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

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: $f_1, \dots, f_n \to f_{j_1}, \dots, f_{j_m}$, $m \ll n$.
3. преобразование признаков: $f_1, \dots, f_n \to g_1, \dots, g_m$, $m \ll n$;

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^*:

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

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

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

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

Выбор параметра регуляризации τ

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

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

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

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

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

Регуляризация сокращает «эффективную размерность»

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

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

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:

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

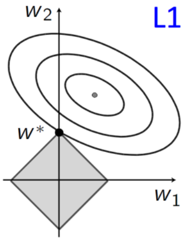
Тождество: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

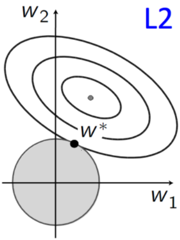
Регуляризация по L_1-норме для отбора признаков

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



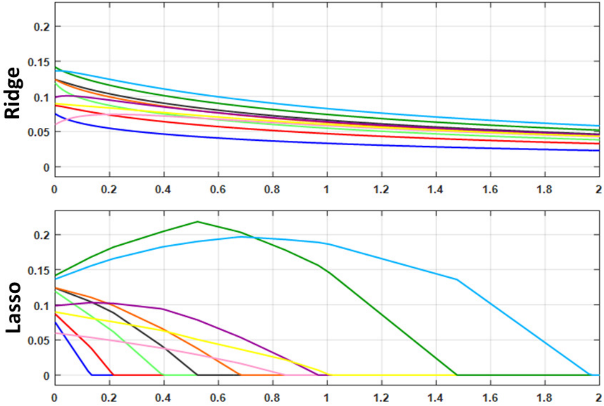


Ridge (L_2):
$$\sum_{j=1}^{n} w_j^2 \leqslant \varkappa$$

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

Сравнение L_2 (Ridge) и L_1 (LASSO) регуляризации

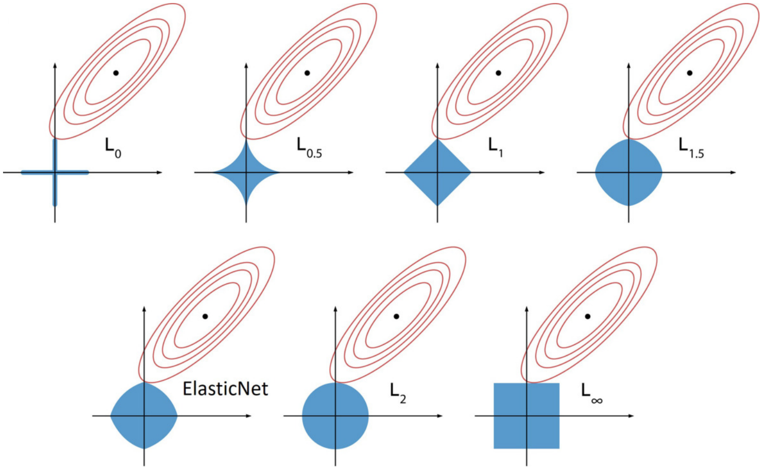
Типичный вид зависимости весов w_j от селективности μ



В LASSO с увеличением μ усиливается отбор признаков

Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным L_p-нормам:



Двойной регуляризатор $L_1 + L_2$ (Elastic Net)

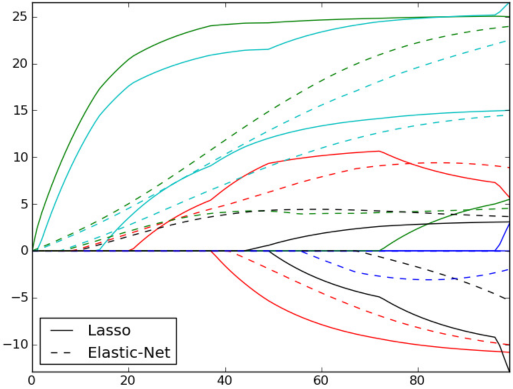
$$\frac{1}{2}\|Fw - y\|^2 + \mu \sum_{j=1}^{n} |w_j| + \frac{\tau}{2} \sum_{j=1}^{n} w_j^2 \;\to\; \min_{w}$$

- ⊕ *Параметр селективности* μ управляет отбором признаков: чем больше μ, тем меньше признаков останется
- ⊕ Есть *эффект группировки* (grouping effect): значимые зависимые признаки отбираются вместе и имеют примерно равные веса w_j
- ⊖ Приходится подбирать два параметра регуляризации μ, τ (есть специальные методы — regularization path)
- ⊖ Шумовые признаки также группируются вместе; по мере увеличения μ группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine. 2006

Двойной регуляризатор $L_1 + L_2$ (Elastic Net)

Elastic Net менее жёстко отбирает признаки, чем LASSO.
Зависимости весов w_j от коэффициента $\log \frac{1}{\mu}$:

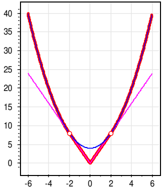


scikit-learn.org/0.5/auto_examples/glm/plot_lasso_coordinate_descent_path.html

Двойной регуляризатор: склейка L_1 и L_2

$$\frac{1}{2}\|Fw - y\|^2 + \sum_{j=1}^{n} R_\mu(w_j) \;\to\; \min_w$$

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



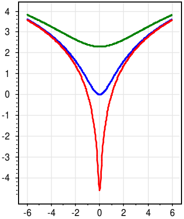
- Только один параметр регуляризации μ
- Отбор признаков с *параметром селективности* μ
- *Эффект группировки*: значимые зависимые признаки ($|w_j| > \mu$) входят в решение совместно (как в Elastic Net)
- Шумовые признаки ($|w_j| < \mu$) не группируются и подавляются независимо друг от друга (как в LASSO)

Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. 2010.

Невыпуклый регуляризатор с эффектом отбора признаков

$$\frac{1}{2}\|Fw - y\|^2 + \sum_{j=1}^{n} \ln\left(w_j^2 + \tfrac{1}{\mu}\right) \to \min_w$$

$R(w) = \ln\left(w^2 + \frac{1}{\mu}\right)$ при $\mu = 0.1,\ 1,\ 100$



- ⊕ Только один параметр регуляризации μ
- ⊕ Отбор признаков с параметром *селективности* μ
- ⊕ Есть эффект группировки
- ⊕ Лучше отбирает набор значимых признаков, когда они только совместно обеспечивают хорошее решение

Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines. 2008.

Резюме в конце лекции

- Многомерная линейная регрессия
  — через *сингулярное разложение*
- Три приёма против мультиколлинеарности и переобучения:
  — регуляризация
  — отбор признаков
  — преобразование признаков
- L_2-регуляризация, она же гребневая регрессия
  — тоже через *сингулярное разложение*
- L_1-регуляризация (LASSO) и др. негладкие регуляризаторы
  — регулируемый отбор признаков
- Преобразование признаков: *метод главных компонент* и другие *матричные разложения* — в следующем семестре