Линейная регрессия

Зинина Анастасия

Содержание

- Задана выборка $\{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^M\}$ и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ значений зависимой переменной.
- Объекты отождествляются со своим признаковым описанием, т.е. задана матрица объектов-признаков:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}$$

- Задана регрессионная модель

$$y = f(\alpha, \mathbf{x}) + \varepsilon,$$

где f — функция регрессионной зависимости, а ε — случайная величина с нулевым матожиданием.

- Модель называется линейной регрессией, если $f(\alpha, \mathbf{x})$ линейна по коэффициентам α:

$$f(\alpha, \mathbf{x}) = \alpha_0 + \sum_{j=1}^{n} \alpha_j f_j(x)$$

- Требуется найти наиболее вероятные параметры $\hat{\alpha}$:

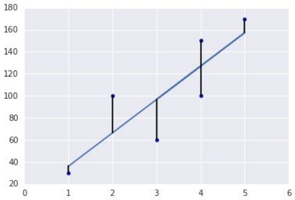
$$\hat{\alpha} = \underset{\alpha \in R^M}{argmax}\ p(y|x, \alpha, f).$$

Функционал ошибки

- Как подобрать подходящие параметры?
- Зададим функционал ошибки, зависящий от параметров, и найдём его минимум
- Рассмотрим в качестве функционала ошибки MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\alpha, \mathbf{x}_i))^2$$

- Восстановление регрессии с данным функционалом ошибки - **метод наименьших квадратов**



МНК

- Обозначения: матрица объектов-признаков F (где все признаки $f_{:X \to R}$ числовые), целевой вектор y, вектор параметров α:

$$F = f_1 \ \ldots \ f_n \ ,$$

$$f = \begin{pmatrix} f_{(x_1)} \\ \vdots \\ f_{(x_l)} \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

- Алгоритм:

$$a(x) \ = \ \sum_{j=1}^{n} \alpha_j f_j(x) \ = \ F\alpha.$$

- Оценим качество его работы на выборке $(x_i, \ y_i)_{i=1}^{l}$ с помощью МНК:

$$Q(\alpha, X^l) \ = \ \sum_{i=1}^{l} (a(x_i) - y_i)^2 \to \min_{\alpha \in \mathbb{R}^n},$$

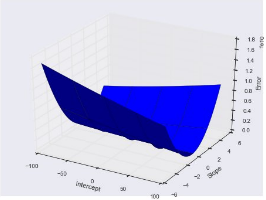
$$Q(\alpha) \ = \| \ F\alpha \ - \ y \ \|^2 \to \min_{\alpha \in \mathbb{R}^n} .$$

МНК

- Найдём минимум $Q(\alpha)$ по α:

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

$\Rightarrow (F^TF)\alpha = F^Ty$—нормальное уравнение МНК



- Если $rank(F^TF) = n$, то можно обращать матрицу F^TF:

$$\alpha^* = (F^TF)^{-1}F^Ty = F^+y,$$

где $F^+ = (F^TF)^{-1}F^T$ - псевдообратная матрица.

Мультиколлинеарность

- Основной проблемой многомерной линейной регресии является мультиколлинеарность матрицы $F^T F$, которую приходится обращать.
- Подобные проблемы возникают, когда среди признаков $f_j(x)$ есть почти линейно зависимые.
- Мультиколлинеарность матрицы определяется её числом обусловленности:

$$\mu(F^T F) \ = \| \, F^T F \, \| * \| \, (F^T F)^{-1} \, \| = \ \frac{\lambda_{max}}{\lambda_{min}},$$

  где λ — собственные значения матрицы $F^T F$.
- Чем больше число обусловленности, тем ближе матрица $F^T F$ к вырожденной и тем неустойчивее обратная к ней матрица.

SVD

Для получения обращения, устойчивого к малым изменениям значений матрицы F, используется SVD.

- Рассмотрим сингулярное разложение матрицы F:

$$F = VDU^T.$$

- В таких обозначениях:

$$F^+ = (F^T F)^{-1} F^T = (UDV^T VDU^T)^{-1} UDV^T = (UDDU^T)^{-1} UDV^T$$
$$= U^{-T} D^{-2} U^{-1} UDV^T = U^{-T} D^{-2} DV^T,$$

  а так как $U^{-1} = U^T$, то

$$F^+ = UD^{-1}V^T = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^T$$

  в силу диагональности матрицы D.

- А решение метода наименьших квадратов запишется в следующем виде:

$$\alpha^* = F^+ y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T,\ y);$$

Градиентный спуск

- Выбор начального приближения для вектора весов $\vec{\alpha}^{[0]}$.
- После вычисляется приблизительное $\vec{\alpha}^{[1]}$, а затем $\vec{\alpha}^{[2]}$ и так далее, согласно итерационной формуле:

$$\vec{\alpha}^{[j+1]} = \vec{\alpha}^{[j]} - \gamma^{[j]} \nabla F(\vec{\alpha}^{[j]}),$$

где $\gamma^{[j]}$ — шаг градиентного спуска.

- $\gamma^{[k]}$ выбирается:
  -постоянной, в этом случае метод может расходиться;
  -дробным шагом;
  -наискорейшим спуском: $\gamma^{[k]} = \arg\min_{\gamma} \; f(\alpha^{[k]} - \gamma \nabla f(\alpha^{[k]}))$.

SGD

- Градиентный шаг для весов будет выглядеть следующим образом:

$$\alpha_0 \leftarrow \alpha_0 + \frac{2\eta}{\ell} \sum_{i=1}^{\ell} (y_i - (\alpha_0 + \alpha_1 f_1(x_i) + \alpha_2 f_2(x_i) + \dots + \alpha_n f_n(x_i)))$$

$$\alpha_j \leftarrow \alpha_j + \frac{2\eta}{\ell} \sum_{i=1}^{\ell} f_j(x_i)(y_i - (\alpha_0 + \alpha_1 f_1(x_i) + \alpha_2 f_2(x_i) + \dots + \alpha_n f_n(x_i))),$$

$$j \in \{1, 2, ..., n\}$$

  Т.е. обновляем веса для улучшения качества на всей выборке

- В SGD поправки для весов вычисляются только с учетом одного случайно взятого объекта обучающей выборки:

$$\alpha_0 \leftarrow \alpha_0 + \frac{2\eta}{\ell} (y_k - (\alpha_0 + \alpha_1 f_1(x_k) + \alpha_2 f_2(x_k) + \dots + \alpha_n f_n(x_k)))$$

$$\alpha_j \leftarrow \alpha_j + \frac{2\eta}{\ell} f_j(x_k)(y_k - (\alpha_0 + \alpha_1 f_1(x_k) + \alpha_2 f_2(x_k) + \dots + \alpha_n f_n(x_k))),$$

$$j \in \{1, 2, ..., n\}$$

  где k - случайный индекс, $k \in \{1, \dots, \ell\}$.

Статистические свойства МНК-оценок

- Предположения:
  1. $y = F\alpha + \varepsilon$
  2. наблюдения (x_i, y_i) независимы
  3. $rank\ F = n$
  4. $E(\varepsilon_i) = 0$
  5. $D(\varepsilon_i) = \sigma^2$
  6. $\varepsilon \sim N(0, \sigma^2 I)$
- Параметры α могут быть оценены с помощью ММП: Функция правдоподобия:

$$L(S, \alpha_0, ..., \alpha_n) = \prod \frac{1}{\sqrt{2\pi\sigma}} exp \frac{-(y_j - \alpha_0 - \sum \alpha_j f_j(x_i))^2}{2\sigma}$$

- Точка максимума L совпадает с точкой минимума

$$-\ln L(S, \alpha_0, ..., \alpha_n) = \sum [\frac{1}{2} \ln 2\pi + \ln \sigma + \frac{1}{2\sigma}(y_j - \alpha_0 - \sum \alpha_i f_j(x_i))^2]$$

- А точка минимума $-\ln[L(S, \alpha_0, ..., \alpha_n)]$ совпадает с точкой минимума MSE.
- То есть ММП и МНК эквивалентны.

Статистические свойства МНК-оценок

α - истинный вектор параметров, $\hat{\alpha}$ - коэффициенты, полученные с помощью МНК

1. Несмещенность $M\hat{\alpha}_j = \alpha_j$

2. Состоятельность $\gamma > 0 \quad \lim P(\mid \alpha_j - \hat{\alpha}_j \mid < \gamma) = 1, j = 0, ..., k$

3. Нормальное распределение $\hat{\alpha} \sim N(\alpha, \sigma^2 (F^T F)^{-1})$

4. Эффективность: наименьшая дисперсия оценок $\hat{\alpha}$ в классе оценок, линейных по y (теорема Гаусса-Маркова).

Борьба с переобучением

- Отбор признаков
- Преобразование признаков
- Регуляризация

Регуляризация: Ridge-регрессия

- Вводится модифицированный функционал

$$Q_\tau = ||y - F\alpha||^2 + \tau||\alpha||^2 \to \min_\alpha$$

где $\tau \in (0,1)$ - коэффициент регуляризации.

- МНК (регуляризованное) решение получается таким

$$\hat{Q}_\tau = (F^T F + \tau I_k)^{-1} F^T y$$

- У матриц $F^T F$ и $F^T F + \tau I_k$ собственные вектора совпадают, а собственным значением различаются на τ. Поэтому число обусловленности для матрицы $F^T F + \tau I$ равно

$$\mu(F^T F + \tau I) = \frac{\lambda_{max} + \tau}{\lambda_{min} + \tau}.$$

Регуляризация: Lasso-регрессия

- Вводится ограничение-неравенство, запрещающее слишком большие абсолютные значения коэффициентов:

$$\begin{cases} Q(a) = \| F\alpha - y \|^2 \to \min_\alpha \\ \sum | \alpha_j | \leqslant \tau \end{cases}$$

  где τ — параметр регуляризации.

- По теореме Куна — Таккера решение задачи оптимизации с ограничениями-неравенствами можно записать в эквивалентном виде

$$\| F\alpha - y \|^2 + \lambda \sum | \alpha_j | \to \min_\alpha$$

Регуляризация: Lasso-регрессия

- Лассо осуществляет отбор информативных признаков. Введём:

$$\alpha_j^+ = 1/2(\mid \alpha_j \mid + \alpha_j) \geqslant 0, \quad \alpha_j^- = 1/2(\mid \alpha_j \mid - \alpha_j) \geqslant 0,$$

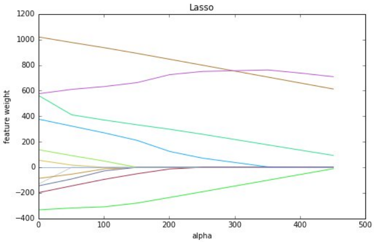
$$\alpha_j = \alpha_j^+ + \alpha_j^- .$$

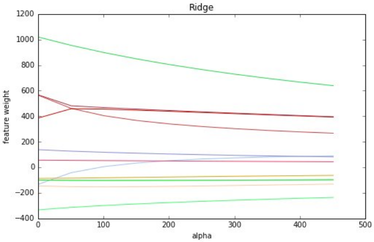
- Тогда минимизируемый функционал останется квадратичным по новым переменным, но станет гладким, а ограничение примет линейный вид:

$$\alpha_j^+ + \alpha_j^- \leqslant \tau$$

Lasso vs Ridge

Почувствуйте разницу:





Список литературы

- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. — New York: Springer, 2001.
- *Воронцов К.В.* Курс лекций "Вычислительные методы обучения по прецедентам"
- *Дрейпер Н.Р., Смит Г.* прикладной регрессионный анализ. - М.:Издательский дом "Вильямс 2007.
- *Кобзарь А.И.* Прикладная математическая статистика - М.: Физматлит, 2006.
- *Tibshirani R. J.* Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. Series B (Methodological). — 1996. — Vol. 58, no. 1. — Pp. 267–288
- *Лагутин М.Б.* Наглядная математическая статистика - М.:БИНОМ.Лаборатория знаний, 2009.
- *Andrew Ng* Курс "Machine Learning"