

Введение в искусственный интеллект.

Машинное обучение

Лекция 4. Линейная регрессия

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

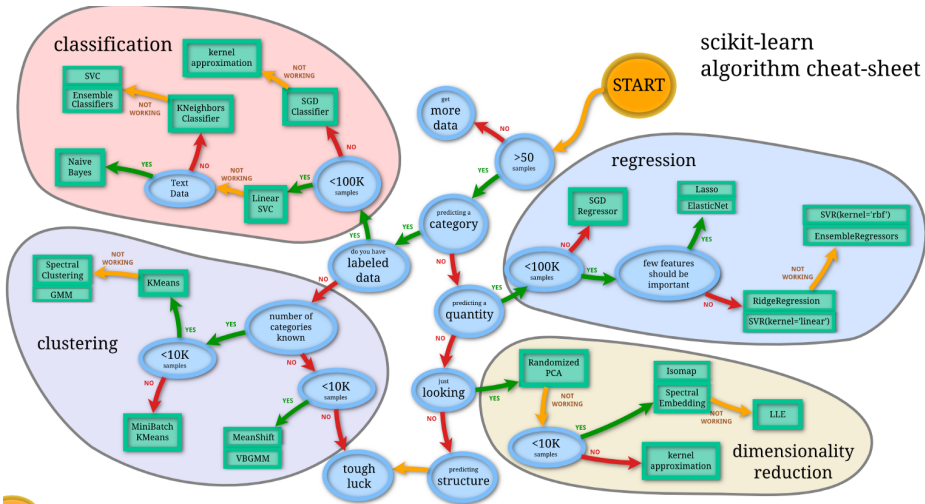
27 октября 2020г.



- 1 Регрессия
- 2 Оценка качества регрессии

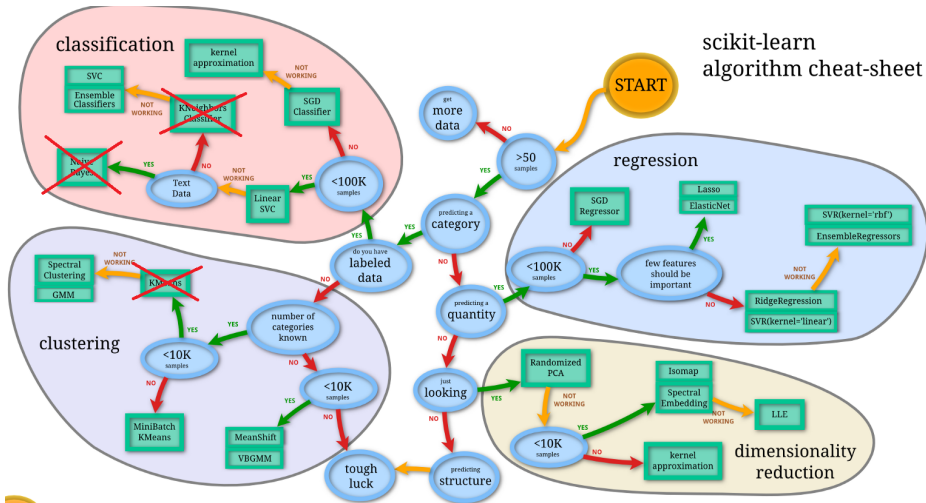


Дорожная карта Scikit-Learn¹



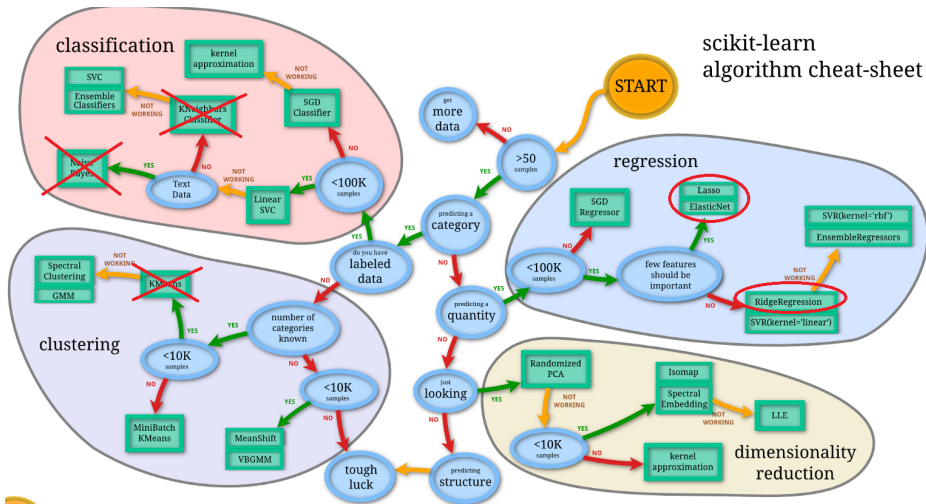
¹https://scikit-learn.org/stable/tutorial/machine_learning_map/

Дорожная карта Scikit-Learn¹



¹https://scikit-learn.org/stable/tutorial/machine_learning_map/

Дорожная карта Scikit-Learn¹



¹https://scikit-learn.org/stable/tutorial/machine_learning_map/

Постановка задачи

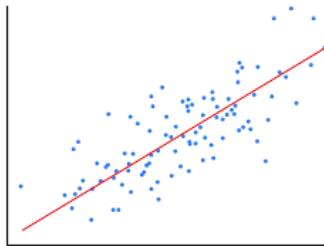
Дано

$$p(y_i|x_i) = w^T x_i + \varepsilon_i,$$

для $i = 1.., \ell$, где $w \in \mathbf{R}^{n+1}$, $\varepsilon_i \sim N(0, \sigma^2)$

Задача

Найти w



Напоминание: два вида оценивания параметров

Принцип максимального правдоподобия

$$w_{ML} = \arg \max_w p(y|w, x)$$

Принцип максимума апостериорной вероятности

$$w_{MAP} = \arg \max_w p(w|x, y)$$



Оценка максимального правдоподобия

$$w_{ML} = \arg \max_w p(y|w, x)$$

$$w_{ML} = \arg \max_w \prod_i p(y_i|w, x_i)$$

$$p(y_i|w, x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

$$w_{ML} = \arg \max_w \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) = \arg \max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2}$$

$$w_{ML} = \arg \min_w \sum_i (y_i - w^T x_i)^2$$



Метод наименьших квадратов

Постановка задачи и допущения

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_w(x) = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$, где $w = (w_0, w_1, \dots, w_n)^T \in \mathbb{R}^{n+1}$ — параметры модели.
- Удобно писать в векторном виде

$$a(x) = w^T \cdot x,$$

где $x = (1, x^1, \dots, x^n)^T \in \mathbb{R}^{n+1}$.

Метод наименьших квадратов

- $L(w, X_{train}) = MSE(w, X_{train}) = \frac{1}{\ell} \sum_i (w^T \cdot x^{(i)} - y_i)^2$ — функция потерь
- Задача найти $\hat{w} = \arg \min_w (L(w, X_{train}))$

Теорема

Решением задачи $\arg \min_w \left(\sum_{i=1}^{\ell} (w^T \cdot x_i - y_i)^2 \right)$ является $\hat{w} = (X^T X)^{-1} \cdot X^T \cdot y$, где $X_{i,j} = x_i^j$, $y = (y_1, \dots, y_{\ell})$.

Доказательство

Запишем задачу в векторном виде $\|Xw - y\|^2 \rightarrow \min_w$. Необходимое условие минимума в матричном виде имеет вид:

$$\begin{aligned} \frac{\partial}{\partial w} \|Xw - y\|^2 &= \frac{\partial}{\partial w} \left((Xw - y)^T \cdot (Xw - y) \right) = \frac{\partial}{\partial w} \left((Xw)^T Xw - (Xw)^T y - y^T Xw + y^T y \right) = \\ &= \frac{\partial}{\partial w} \left(w^T X^T Xw - w^T X^T y - y^T Xw + y^T y \right) = \frac{\partial}{\partial w} w^T (X^T X)w - 2 \frac{\partial}{\partial w} (X^T y)^T w = 0 \end{aligned}$$

Определение

Пусть $w = (w_1, \dots, w_n)$ — вектор столбец, а $z = z(w_1, \dots, w_n)$. Тогда определим

$$\frac{\partial z}{\partial w} := \left(\frac{\partial z}{\partial w_1}, \dots, \frac{\partial z}{\partial w_n} \right)^T$$

Лемма 1

$$\frac{\partial}{\partial x} x^T a = a$$

Лемма 2

$$\frac{\partial}{\partial x} x^T A x = (A + A^T) x$$

Теорема

Решением задачи $\arg \min_w \left(\sum_{i=1}^{\ell} (w^T \cdot x_i - y_i)^2 \right)$ является $\hat{w} = (X^T X)^{-1} \cdot X^T \cdot y$, где $X_{i,j} = x_i^j$, $y = (y_1, \dots, y_{\ell})$.

Продолжение доказательства

Необходимое условие минимума в матричном виде имеет вид:

$$\frac{\partial}{\partial w} \|Xw - y\|^2 = \frac{\partial}{\partial w} w^T (X^T X) w - 2 \frac{\partial}{\partial w} (X^T y)^T w =$$

Далее применяем леммы и приравниваем к нулю:

$$= 2X^T Xw - 2X^T y = 0,$$

откуда получаем $w = (X^T X)^{-1} \cdot X^T \cdot y$, что и требовалось доказать.

Полиномиальная регрессия

Идея

Можно генерировать новые признаки на основе уже имеющихся, применяя нелинейные функции

Примеры преобразований

- Возведение в степень
- Парные произведения
- Квадратный корень
- Логарифм
- Экспонента



Преимущества и недостатки линейной регрессии

Преимущества

- Простой алгоритм, вычислительно не сложный
- Линейная регрессия хорошо интерпретируемая модель
- Несмотря на свою простоту может описывать довольно сложные зависимости (например, полиномиальные)

Недостатки

- Алгоритм предполагает, что все признаки числовые
- Алгоритм предполагает, что данные распределены нормально, что не всегда так
- Алгоритм сильно чувствителен к выбросам





Метод максимизации апостериорной вероятности

$$w_{MAP} = \arg \max_w p(w|x_1, \dots, x_\ell, y_1, \dots, y_\ell)$$

$$w_{MAP} = \arg \max_w \prod_i p(y_i|x_i, w)p(w)$$

$$w_{MAP} = \arg \max_w \sum_i \ln p(y_i|x_i, w) + \ln p(w)$$

$$w_{MAP} = \arg \max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2} + \ell \ln p(w)$$

$$w_{MAP} = \arg \min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - \ell \ln p(w)$$

В задаче минимизации появилось дополнительное слагаемое, которое зависит только от априорного распределения на веса w

Метод максимизации апостериорной вероятности

$$w_{MAP} = \arg \min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - \ell \ln p(w)$$

Предположим, что $p(w) \sim N(0, \tau^2)$

$$w_{MAP} = \arg \min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\ell w^T w}{2\tau^2}$$

$$w_{MAP} = \arg \min_w \frac{1}{\ell} \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2\tau^2} \|w\|^2$$



L2-регуляризация

- $L(w, X_{train}) = MSE(w, X_{train}) + \frac{\alpha}{2} \sum_{i=0}^n w_i^2 = \frac{1}{\ell} \sum_i (w^T \cdot x^{(i)} - y_i)^2 + \frac{\alpha}{2} \sum_{i=0}^n w_i^2$ — функция потерь
- Задача найти $\hat{w} = \arg \min_w (L(w, X_{train}))$

Теорема

Решением задачи $\arg \min_w (\sum_{i=1}^{\ell} (w^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^n w_i^2)$ является

$\hat{w} = (X^T X + \alpha I_{n+1})^{-1} \cdot X^T \cdot y$, где $X_{i,j} = x_i^j$, $y = (y_1, \dots, y_{\ell})$, I_{n+1} — единичная матрица.



Доказательство теоремы

Лемма 3

$$\frac{\partial}{\partial x} x^T x = 2x$$

Теорема

Решением задачи $\arg \min_w \left(\sum_{i=1}^{\ell} (w^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^n w_i^2 \right)$ является

$\hat{w} = (X^T X + \alpha I_{n+1})^{-1} \cdot X^T \cdot y$, где $X_{i,j} = x_i^j$, $y = (y_1, \dots, y_{\ell})$, I_{n+1} — единичная матрица.

Доказательство

Запишем задачу в векторном виде $\|Xw - y\|^2 + \alpha \|w\|^2 \rightarrow \min_w$. Необходимое условие минимума в матричном виде имеет вид:

$$\frac{\partial}{\partial w} \left((Xw - y)^T \cdot (Xw - y) + \alpha w^T w \right) = 2X^T Xw - 2X^T y + 2\alpha w = 0$$

Свойства гребневой регрессии

- Регуляризация не даёт параметрам модели быть слишком большими
- Как правило регуляризация обеспечивает большую обобщающую способность
- Более устойчива к выбросам
- Появился параметр, который можно настроить при помощи кросс-валидации

Вероятностный смысл параметра α

$\alpha = \frac{1}{\tau^2}$, где τ — среднеквадратическое отклонение априорного распределения на w



L1-регуляризация

- $L(w, X_{train}) = MSE(w, X_{train}) + \alpha \sum_{i=0}^n |w_i| = \sum_i (w^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^n |w_i|$ — функция потерь
- Задача найти $\hat{w} = \arg \min_w (L(w, X_{train}))$



L1-регуляризация

- $L(w, X_{train}) = MSE(w, X_{train}) + \alpha \sum_{i=0}^n |w_i| = \sum_i (w^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^n |w_i|$ — функция потерь
- Задача найти $\hat{w} = \arg \min_w (L(w, X_{train}))$

Свойства

- Эта регуляризация обеспечивает отбор признаков
- Нет аналитического решения



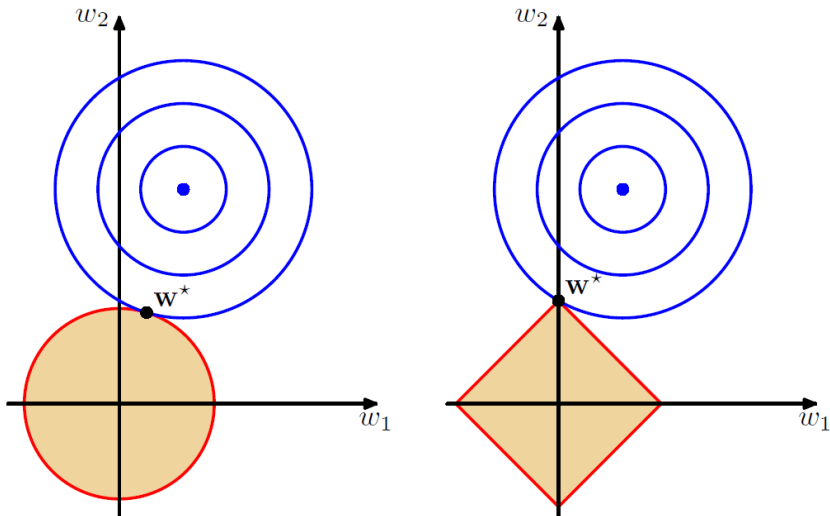
Вероятностный смысл параметра α

Параметр α — обратно пропорционален среднеквадратичному отклонению априорного распределения на w . В данном случае это распределение Лапласа

$$p(w) = \frac{1}{\tau} \exp\left(-\frac{\|w\|}{2\tau}\right)$$



Отбор признаков при L1-регуляризации



L1-регуляризация и L2-регуляризация

- $L(w, X_{train}) = MSE(w, X_{train}) + r\alpha \sum_{i=0}^n |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^n w_i^2 =$
 $\sum_i (w^T \cdot x^{(i)} - y_i)^2 + r\alpha \sum_{i=0}^n |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^n w_i^2$ — функция потерь
- Задача найти $\hat{w} = \arg \min_w (L(w, X_{train}))$

Свойства

- Нет аналитического решения
- Совмещает положительные свойства гребневой регрессии и LASSO.





Мотивация

- Постановка задачи машинного обучения обычно начинается с определения метрики и фиксирования тестового датасета, на котором эта метрика будет считаться
- Неправильно выбранная метрика может затруднить использование модели машинного обучения в жизни и свести на нет усилия команды, разрабатывающей алгоритм машинного обучения.
- Как правило заказчик не мыслит в терминах метрик и может объяснить проблему, которую он хочет решить, только бизнес языком
- Понимание влияния выбора той или иной метрики на бизнес — это ключ к успешной постановки задачи



Метрики качества для задачи регрессии

Mean Square Error

$$MSE = \frac{1}{\ell} \sum_i (y_i - a(x_i))^2$$

Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{\ell} \sum_i (y_i - a(x_i))^2}$$

Mean Absolute Error

$$MAE = \frac{1}{\ell} \sum_i |y_i - a(x_i)|$$

Метрики качества для задачи регрессии

Max Error

$$ME = \max(|y_i - a(x_i)|)$$

Mean Squared Logarithmic Error

$$MSLE = \frac{1}{\ell} \sum_i (\ln y_i - \ln a(x_i))^2$$

R^2 score

$$R^2 = 1 - \frac{\sum_i (y_i - a(x_i))^2}{\sum_i (y_i - \bar{y})^2},$$

где $\bar{y} = \frac{1}{\ell} \sum_i y_i$

- Линейная регрессия — простая, хорошо интерпретируемая модель, не устойчивая к выбросам
- Имеет наглядную вероятностную интерпретацию
- Регуляризация — отличный способ борьбы с переобучением и шумом в данных



Дорожная карта Scikit-Learn²

