ТЕХНОСФЕРА

Лекция 7
Линейные модели

Кристина Федоренко

14 ноября 2016 г.

План занятия

Линейная регрессия

Логистическая регрессия

Постановка задачи

Дано. Признаковые описания N объектов $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}$, образующие тренировочный набор данных X, и значения целевой переменной $y = f(\mathbf{x}) \in \mathcal{Y}$ для каждого объекта из X.

Найти. Для семейства параметрических функций

$$H = \{h(\mathbf{x}, \theta) = y : \mathcal{X} \times \Theta \to \mathcal{Y}\},$$

найти значение вектора параметров θ^*, такое что $h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$ наилучшим образом приближает целевую функцию.

$$\begin{aligned} Y &\in \{C_1, C_2, \ldots, C_K\} \text{ – задача классификации} \\ Y &\in [a, b] \subset \mathcal{R} \text{ – задача регрессии} \end{aligned}$$

Линейная регрессия

Дано

Обучающая выборка $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, где $x_i \in R^d$, $y \in R$.

Найти

Функцию $h : R^d \to R$, которая предсказывает y по новым x

Гипотеза

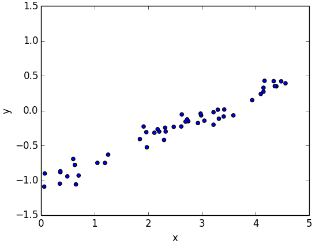
$h(\mathbf{x}, \theta) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$

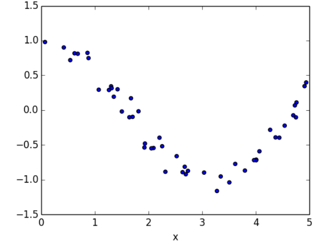
Базисные функции

Введем функции
$\phi_j : R^d \to R^m$.
$\phi_j(\mathbf{x}) = (\phi_1(x), \ldots, \phi_m(x))$.

Тогда,
$h(\mathbf{x}, \theta) = \sum_{j=0}^{m} \theta_j \phi_j(\mathbf{x}) = \theta^T \phi(\mathbf{x})$,
$\phi_j(\mathbf{x})$ – базисные функции, $\phi_0(\mathbf{x}) = 1$

Пример. Полиномиальные базисные функции
Пусть $d = 2$, тогда
$h(\mathbf{x}, \theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 + \theta_5 x_1 x_2$
$\phi(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$





Гауссовская линейная регрессия

Хотим оценить $p(y|\mathbf{x})$.
Предположим, что $p(y|\mathbf{x}) \in p_\phi(y|\mathbf{x})$, где $\phi \in \Phi$.
Задача оценить ϕ по данным $\mathcal{D}$

Рассмотрим семейство гауссовских функций
$p_\phi(y|\mathbf{x}) = \mathcal{N}(y|\mu(x), \sigma^2(x))$. $\phi = (\mu, \sigma^2)$.

Гауссовская линейная регрессия
$p_\phi(y|\mathbf{x}) = \mathcal{N}(y|\theta^T x, \sigma^2)$. $\phi = (\theta, \sigma^2)$, где $\theta \in R^d, \sigma^2 \in R^+$

ML – функция правдоподобия

Дана обучающая выборка $\mathcal{D} = (X, Y)$ из N объектов $(\mathbf{x_n}, y_n)$

Функция правдоподобия

$$\log p(Y|\mathbf{x}, \theta, \sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}(y|\theta^T \mathbf{x_n}, \sigma^2) =$$

$$= \sum_{n=1}^{N} \log(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_n - \theta^T x_n)^2)) =$$

$$= -\frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \theta^T \mathbf{x}_n)^2 \to \max_{\theta, \sigma}$$

Квадратичная функция потерь (cost function)

$$J(\theta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta^T \mathbf{x}_n)^2$$

ML – решение

$$\log p(Y|X,\theta,\sigma^2) = -\frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \theta^T \mathbf{x}_n)^2 \to \max_{\theta,\sigma}$$

Градиент

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \theta^T \mathbf{x}_n)\mathbf{x_n}^T = 0$$

Решение

$$\theta_{ML} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \theta_{ML}^T \mathbf{x}_n)^2),$$

где
$\mathbf{X} \in R^{n\times d}$ – матрица признаков.

Метод градиентного спуска

Гипотеза
$h(\mathbf{x}, \theta) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$

Параметры
$\theta = (\theta_1, \ldots, \theta_d)$

Функция потерь $\mathrm{J}(\theta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta^T \mathbf{x}_n)^2$

Градиентный спуск
Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

Градиентный спуск

```
function gd(J, alpha, epsilon):
    initialise theta
    do:
        theta = new_theta
        new_theta = theta - alpha*grad(theta)
    until dist(new_theta, theta) < epsilon
    return theta
```

Если нужно быстрее, то Stochastic Gradient Descent(SGD)

Логистическая регрессия

Гипотеза

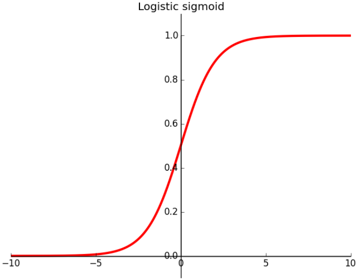
$h(\mathbf{x}, \theta) = g(\theta^T \mathbf{x})$

Сигмоида

$g(z) = \frac{1}{1+\exp(-z)}$.

Тогда,

$$\hat{p}(y = 1 | \mathbf{x}, \theta) = h(\mathbf{x}, \theta) = \frac{1}{1 + \exp(-\theta^T \mathbf{x})}$$



Упражнение

X_1	X_2
0	1
1	0
1	1
2	2
2	3
3	2

Определите к какому классу относятся наблюдения, если вектор параметров имеет вид $(\theta_0 = 0.2, \theta_1 = 0.1, \theta_2 = 0.1)$

Функция потерь для логистической регрессии

Обучающая выборка

$\mathcal{D} = (X, Y)$ из N объектов $(\mathbf{x_n}, y_n)$, где $x_i \in R^d$, $y \in \{0, 1\}$.

Гипотеза

$$h(\mathbf{x}, \theta) = \frac{1}{1 + \exp(-\theta^T \mathbf{x})}$$

Функция потерь для одного наблюдения

$$cost(y, h(\mathbf{x}, \theta)) = -y \log(h(\mathbf{x}, \theta)) - (1 - y) log(1 - h(\mathbf{x}, \theta))$$

Функция потерь для логистической регрессии

Функция потерь для выборки

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N} cost(y_n, h(\mathbf{x}_n, \theta)) = \frac{1}{N}\sum_{n=1}^{N} -y_n \log(h(\mathbf{x_n}, \theta)) - (1 - y_n) log(1 - h(\mathbf{x_n}, \theta))$$

Частная производная по θ_j

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{N}\sum_{n=1}^{N} (h(\mathbf{x}_n, \theta) - y_n) x_n^j$$

Градиентный спуск для логистической регрессии

Частная производная по θ_j

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{N} \sum_{n=1}^{N} (h(\mathbf{x}_n, \theta) - y_n) x_n^j$$

Градиентный спуск

Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) = \theta_j - \alpha \frac{1}{N} \sum_{n=1}^{N} (h(\mathbf{x}_n, \theta) - y_n) x_n^j$$

}

Регуляризация

Функция потерь

$$J_R(\theta, \lambda) = J(\mathbf{w}) + \lambda \sum_{n=1}^{N} |\mathbf{w}_n|^q,$$

где для линейной регрессии

$$J(\theta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta^T \mathbf{x}_n)^2,$$

для логистической регрессии

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} -y_n \log(h(\mathbf{x_n}, \theta)) - (1 - y_n) log(1 - h(\mathbf{x_n}, \theta))$$

Виды регуляризации

- $q = 1$ – Lasso
- $q = 2$
- $\rho \sum_{n=1}^{N} |\mathbf{w}_n| + (1 - \rho) \sum_{n=1}^{N} \mathbf{w}_n^2$ – Elastic Net

Мультикласс классификация

- one-vs-rest
  Строим K моделей, каждая соответствует одному классу
- one-vs-one
  Строим $K(K-1)/2$ моделей, каждая соответствует паре классов

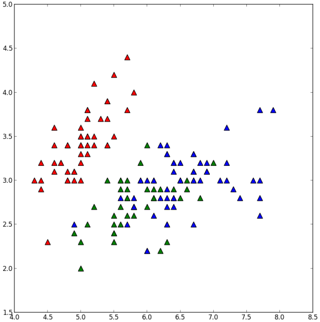
Ирисы Фишера

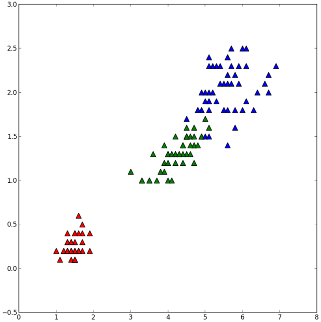
Setosa Versicolor Virginica

Задача

Определить вид ириса на основании длины чашелистика, ширины чашелистика, длины лепестка и ширины лепестка.

Ирисы Фишера





Вопросы