ТЕХНОСФЕРА

Лекция 6
Линейные модели для классификации и регрессии

Николай Анохин

30 октября 2014 г.

План занятия

Линейная регрессия

Логистическая регрессия

Обобщенные линейные модели

Постановка задачи

Пусть дан набор объектов $\mathcal{D} = \{(\mathbf{x}_i, y_i)\},\ \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y},\ i \in 1, \dots, N$, полученный из неизвестной закономерности $y = f(\mathbf{x})$. Необходимо выбрать из семейства параметрических функций

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

такую $h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$, которая наиболее точно апроксимирует $f(\mathbf{x})$.

Задачи

- Регрессия: $\mathcal{Y} = [a, b] \subset \mathbb{R}$
- Классификация: $|\mathcal{Y}| < C$

Линейная регрессия

Модель

$$y = h(\mathbf{x}, \theta) + \epsilon,$$

где ϵ – гауссовский шум

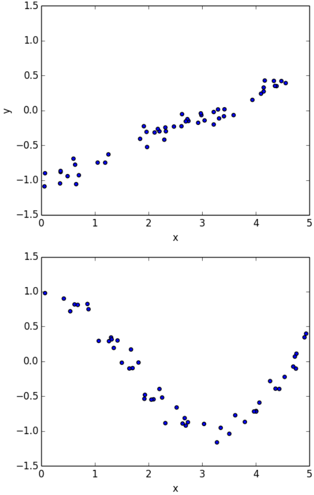
$$p(\epsilon) = \mathcal{N}(\epsilon | 0, \beta^{-1}),$$

откуда

$$p(y|\mathbf{x}, \theta, \beta) = \mathcal{N}(y | h(\mathbf{x}, \theta), \beta^{-1}).$$

Предсказание

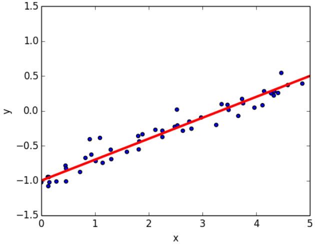
$$E[y|\mathbf{x}] = \int y p(y|\mathbf{x}) dy = h(\mathbf{x}, \theta).$$



Линейная модель

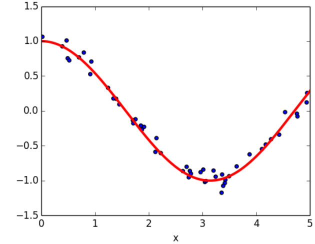
простейшая модель

$$h(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_M x_M = \sum_{j=0}^{M} w_j x_j$$



улучшенная модель

$$h(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}),$$



$\phi_j(\mathbf{x})$ – базисные функции, $\phi_0(\mathbf{x}) = 1$

примеры

$$\varphi_j(x) = x^j, \quad \varphi_j(x) = \exp\left\{-\frac{(x - \mu_j)^2}{2s^2}\right\}$$

ML – функция правдоподобия

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

Функция правдоподобия

$$\log p(Y|X, \mathbf{w}, \beta) = \sum_{n=1}^{N} \log \mathcal{N}(y|\mathbf{w}^T \phi(\mathbf{x_n}), \beta^{-1}) =$$

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

Квадратичная функция потерь

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \to \min_{\mathbf{w}}$$

ML – решение

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

Градиент

$$\beta\sum_{n=1}^{N}\{y_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}\phi(\mathbf{x_n})^T = 0$$

Решение

$$\mathbf{w}_{ML} = \mathbf{\Phi}^\dagger Y = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T Y, \quad \frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y_n - \mathbf{w}_{ML}^T\phi(\mathbf{x}_n)\}^2,$$

где

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x_1}) & \dots & \phi_M(\mathbf{x_1}) \\ \phi_0(\mathbf{x_2}) & \dots & \phi_M(\mathbf{x_2}) \\ \dots & \dots & \dots \\ \phi_0(\mathbf{x_N}) & \dots & \phi_M(\mathbf{x_N}) \end{pmatrix}$$

Регуляризация

Функция потерь

$$E(\mathbf{w}, \lambda) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}),$$

где (как и раньше)

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \to \min_{\mathbf{w}},$$

плюс регуляризация

$$E_W(\mathbf{w}) = E_q(\mathbf{w}) = \sum_{j=1}^{M} |\mathbf{w}_j|^q$$

Зоопарк

- $q = 1$ – Lasso
- $q = 2$ – Ridge (байесовский вывод: $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$)
- $E_W(\mathbf{w}) = \rho E_1(\mathbf{w}) + (1-\rho) E_2(\mathbf{w})$ – Elastic Net

Логистическая регрессия

Ирисы Фишера

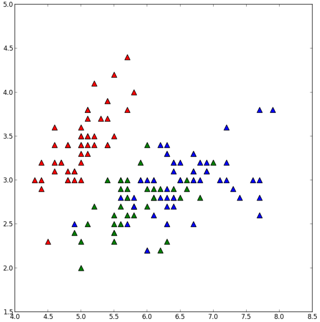
  

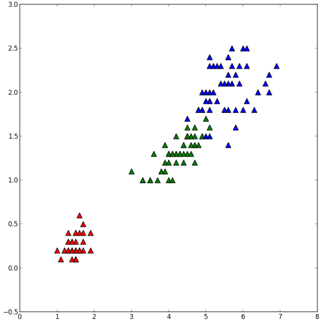
Setosa Versicolor Virginica

Задача

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

Ирисы Фишера





Многомерное нормальное распределение

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right\}$$

Параметры

D-мерный вектор средних

$$\mu = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$D \times D$-мерная матрица ковариации

$$\mathbf{\Sigma} = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$$

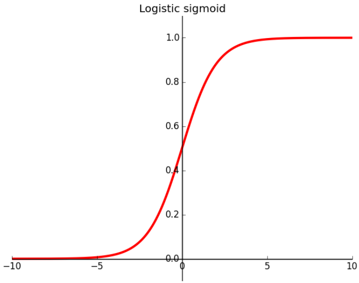
Генеративная модель

Рассматриваем 2 класса

$$p(y_1|x) = \frac{p(x|y_1)p(y_1)}{p(x|y_1)p(y_1) + p(x|y_2)p(y_2)} = \frac{1}{1 + e^{-a}} = \sigma(a)$$

$$a = \ln \frac{p(x|y_1)p(y_1)}{p(x|y_2)p(y_2)}$$

$\sigma(a)$ – сигмоид-функция, $a = \ln(\sigma/(1-\sigma))$



Случай нормальных распределений

Пусть

$$p(\mathbf{x}|y_k) = \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}),$$

тогда

$$p(y_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0),$$

где

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

$$w_0 = -\frac{1}{2}\mu_1^T\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2^T\mathbf{\Sigma}^{-1}\mu_2 + \ln\frac{p(y_1)}{p(y_2)}$$

Аналогичный результат для любых распределений из экспоненциального семейства

Maximum Likelihood

$$p(y_1, \mathbf{x}) = p(y_1)p(\mathbf{x}|y_1) = \pi\mathcal{N}(\mathbf{x}|\mu_1, \mathbf{\Sigma})$$

$$p(y_2, \mathbf{x}) = p(y_2)p(\mathbf{x}|y_2) = (1-\pi)\mathcal{N}(\mathbf{x}|\mu_2, \mathbf{\Sigma})$$

Функция правдоподобия

$$p(Y, X|\pi, \mu_1, \mu_2, \mathbf{\Sigma}) = \prod_{n=1}^{N} \left[\pi\mathcal{N}(\mathbf{x}|\mu_1, \mathbf{\Sigma})\right]^{y_n} \left[(1-\pi)\mathcal{N}(\mathbf{x}|\mu_2, \mathbf{\Sigma})\right]^{1-y_n}$$

Максимизируя $\log p(Y, X|\pi, \mu_1, \mu_2, \mathbf{\Sigma})$, имеем

$$\pi = \frac{1}{N}\sum_{n=1}^{N} y_n = \frac{N_1}{N_1 + N_2},$$

$$\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} y_n \mathbf{x}_n, \quad \mu_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1-y_n)\mathbf{x}_n,$$

аналогично для $\mathbf{\Sigma}$

Обобщенная линеная модель

Базисные функции $\phi_n(\mathbf{x})$

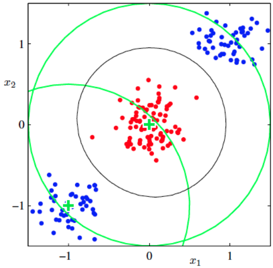
$$\phi_n(\mathbf{x}) = \exp\left[-\frac{(x-\mu_n)^2}{2s^2}\right]$$

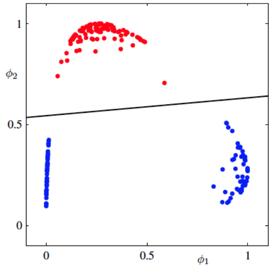
Функция активации $f(a)$

$$f(a) = \sigma(a)$$

(Совсем) обобщенная линейная модель

$$y(\mathbf{x}, \mathbf{w}) = f(\mathbf{w}^\top \phi(\mathbf{x}))$$





Логистическая регрессия

Дано.

$$\mathcal{D} = \{\phi_n = \phi(\mathbf{x}_n), y_n\},\ y_n \in \{0, 1\},\ n = 1 \dots N$$

Модель.

$$p(y = 1|\phi) = \sigma(\mathbf{w}^\top \phi)$$

функция правдоподобия (кросс-энтропия)

$$l(\mathbf{w}) = \log \left[\prod_{n=1}^{N} p^{y_n}(y = 1|\phi_n)(1 - p(y = 1|\phi_n))^{1-y_n} \right] =$$

$$= \sum_{n=1}^{N} y_n \log p(y = 1|\phi_n) + (1 - y_n) \log(1 - p(y = 1|\phi_n)) = -J_c(\mathbf{w}) \to \max_{\mathbf{w}}$$

Градиент

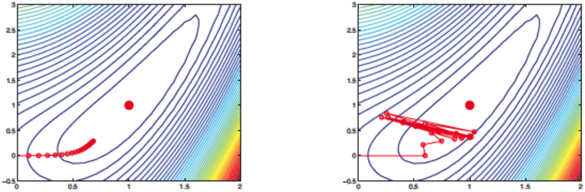
$$\nabla J_c(\mathbf{w}) = \sum_{n=1}^{N} (p(y = 1|\phi_n) - y_n)\phi_n$$

Гессиан

$$\nabla^2 J_c(\mathbf{w}) = \sum_{n=1}^{N} p(y = 1|\phi_n)(1 - p(y = 1|\phi_n))\phi_n \phi_n^T$$

Градиентный спуск

```
function gd(grad, a0, epsilon):
    initialise eta(k)
    k = 0
    a = a0
    do:
        k = k + 1
        a = a - eta(k) grad(a)
    until eta(k) grad(a) < epsilon
    return a
```



(a) (b)

Добавление момента: $\mathbf{a}_{k+1} = \mathbf{a}_k - \eta_k \nabla J(\mathbf{a}_k) + \mu_k(\mathbf{a}_k - \mathbf{a}_{k-1})$

Метод Ньютона

$$J(\mathbf{a}) \approx J(\mathbf{a}_k) + \nabla J(\mathbf{a}_k)^T(\mathbf{a} - \mathbf{a}_k) + \frac{1}{2}(\mathbf{a} - \mathbf{a}_k)^T \nabla^2 J(\mathbf{a}_k)(\mathbf{a} - \mathbf{a}_k) \to \min_{\mathbf{a}}$$

$$\mathbf{a} = \mathbf{a}_k - \nabla^2 J(\mathbf{a}_k)^{-1} \nabla J(\mathbf{a}_k)$$

```
function newton(grad, hessian, a0, epsilon):
    initialise eta(k)
    k = 0
    a = a0
    do:
        k = k + 1
        g = grad(a)
        H = hessian(a)
        d = solve(H * d = -g) # find d = - inv(H) * g
        a = a + eta(k) d
    until convergence
    return a
```

BFGS – использовать приближение $\nabla^2 J(\mathbf{a}_k)$ или $\nabla^2 J(\mathbf{a}_k)^{-1}$

Iterative Reweighted Least Squares

Градиент и Гессиан логистической регрессии в матричной форме

$$\nabla J_c(\mathbf{w}) = X^T(\sigma - Y)$$

$$\nabla^2 J_c(\mathbf{w}) = X^T S X = X^T \mathrm{diag}\{\sigma_n(1-\sigma_n)\} X$$

Обновление весов

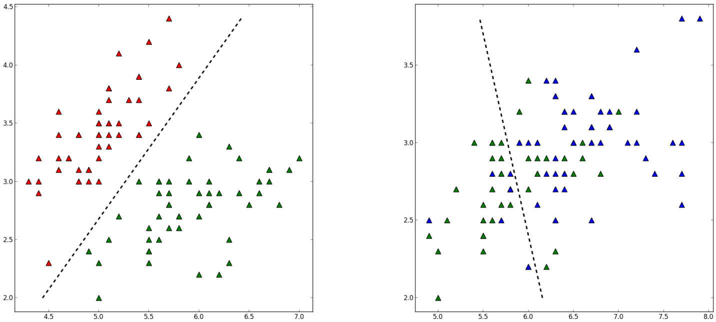
$$\mathbf{w}_{k+1} = \mathbf{w}_k - (X^T S_k X)^{-1} X^T S_k \mathbf{z}_k,$$

$$\mathbf{z}_k = X\mathbf{w}_k + S_k^{-1}(Y - \sigma_k)$$

Минимизация

$$\sum_{n=1}^{N} S_{kn}(z_{kn} - \mathbf{w}^T x_n)^2$$

Логистическая регрессия: результаты



Обобщенные линейные модели

Линейные модели

Рассматривается случай 2 классов

Функция принятия решения

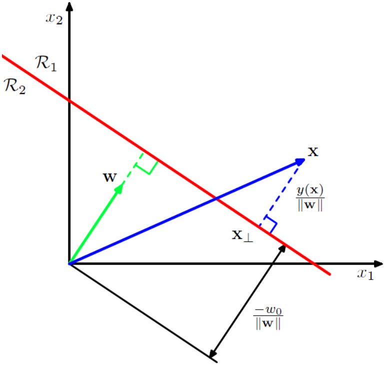
$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$$

Регионы принятия решения

$$R_1 = \{\mathbf{x} \,:\, y(\mathbf{x}) > 0\}$$

$$R_2 = \{\mathbf{x} \,:\, y(\mathbf{x}) < 0\}$$

Задача
найти параметры модели $\mathbf{w}$, w_0



Линейные модели: наблюдения

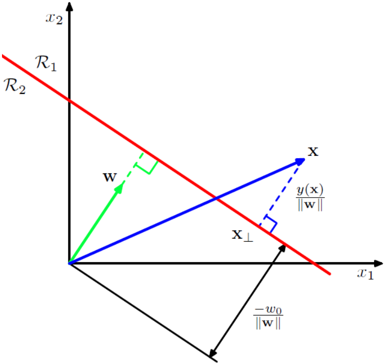
Разделяющая поверхность

$$\mathcal{D} = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + w_0 = 0\}$$

1. $\mathbf{w}$ – нормаль к $\mathcal{D}$
2. $d = -\frac{w_0}{\|\mathbf{w}\|}$ – расстояние от центра координат до $\mathcal{D}$
3. $r(\mathbf{x}) = \frac{y(x)}{\|\mathbf{w}\|}$ – расстояние от $\mathcal{D}$ до $\mathbf{x}$

Положим $x_0 \equiv 1$, получим модель

$$y(\tilde{\mathbf{x}}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}$$



Обобщенные линейные модели

Линейная модель

$$y(\mathbf{x}) = w_0 + \sum w_i x_i$$

Квадратичная модель

$$y(\mathbf{x}) = w_0 + \sum w_i x_i + \sum \sum w_{ij} x_i x_j$$

Обобщенная линейная модель

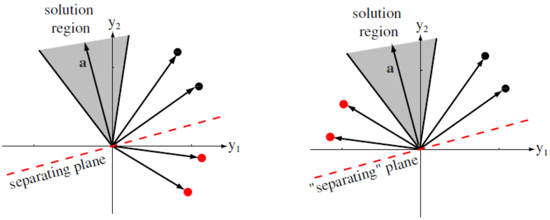
$$g(\mathbf{x}) = \sum a_i \phi_i(\mathbf{x}) = \mathbf{a}^\top \mathbf{y}$$

Случай линейно разделимых классов

Обобщенная линейная модель

$$g(\mathbf{x}) = \sum a_i \phi_i(\mathbf{x}) = \mathbf{a}^\top \mathbf{y}$$

Дана обучающая выборка $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$



Идея

Преобразовать объекты второго класса в обратные им и решать задачу оптимизации в области $a^T \mathbf{y}_i > 0, \ \forall i$

Задача оптимизации

Задача

Минимизируем критерий $J(a)$ при условиях $a^T \mathbf{y}_i > 0, \ \forall i$

Пусть $\mathcal{Y}$ – множество неправильно проклассифицированных объектов

- $J_e(a) = \sum_{\mathbf{y} \in \mathcal{Y}} 1$
- $J_p(a) = \sum_{\mathbf{y} \in \mathcal{Y}} -a^\top \mathbf{y}$
- $J_q(a) = \sum_{\mathbf{y} \in \mathcal{Y}} (a^\top \mathbf{y})^2$
- $J_r(a) = \sum_{\mathbf{y} \in \mathcal{Y}} \frac{(a^\top \mathbf{y})^2 - b}{\|\mathbf{y}\|}$

Улучшение: добавить отступы

Случай линейно неразделимых классов

- Использовать $\eta(k) \to 0$ при $k \to \infty$
- От системы неравенств перейти к системе линейных уравнений
- Линейное программирование

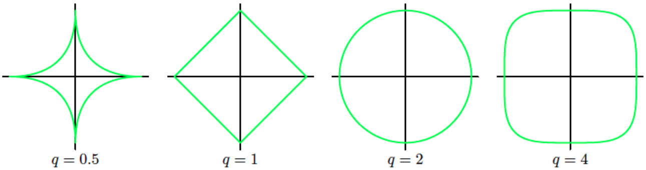
Снова переобучение

Оптимизируем критерий с регуляризацией

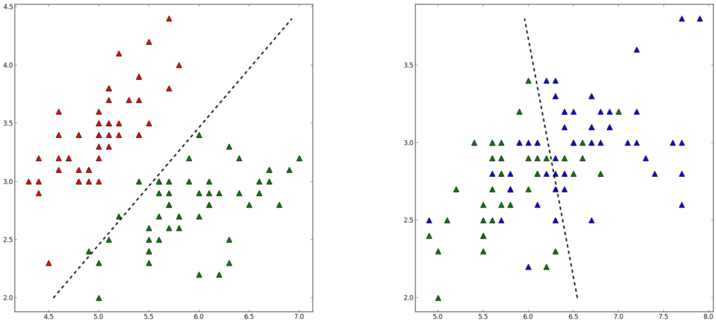
$$J_1(a) = J(a) + \lambda J_R(a)$$

λ – коэффициент регуляризации

$$J_R(a) = \sum |a_j|^q$$



Перцептрон: результаты



Задача: 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

Задача

- Скачать шаблон кода `http://bit.ly/1DvG6hh`
- Реализовать схему one-vs-one
  - Нарисовать раздляющие поверхности на графиках
  - Посчитать итоговую accuracy

Вопросы