МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 9

Логистическая регрессия

9.1. Определения

(Дискриминантный метод)

Рассмотрим задачу классификации на 2 класса: $\mathscr{Y} = \{0, 1\}$

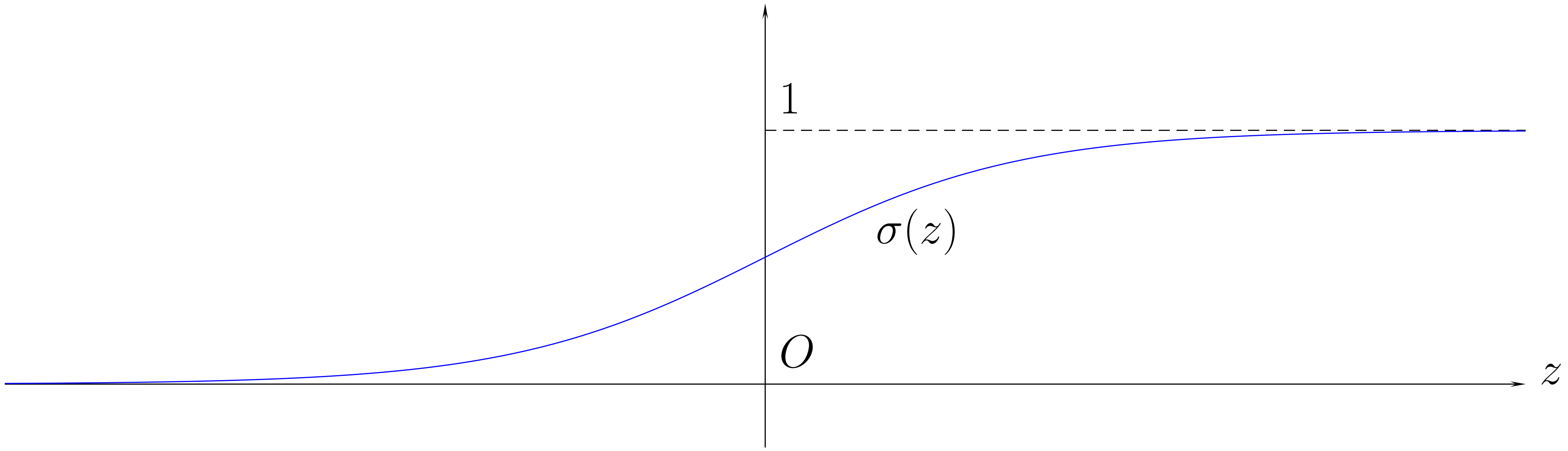
Пусть

$$\mathsf{Pr}(Y = 1 | X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d)}} = \sigma(\beta_0 + \beta^\top x),$$

где

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

— *логистическая функция* (элементарный *сигмоид* или *логит*-функция)



$$\mathrm{Pr}(Y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\dots+\beta_d x_d)}}=\sigma(\beta_0+\beta^\top x),$$

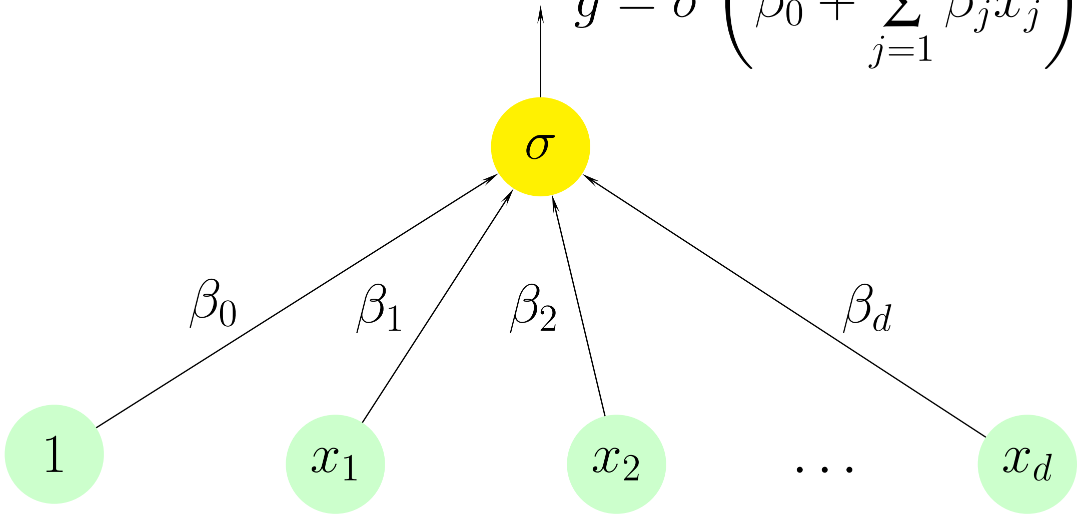
тогда

$$\mathrm{Pr}(Y=0|x)=1-\mathrm{Pr}(Y=1|x)=\frac{1}{1+e^{\beta_0+\beta_1 x_1+\dots+\beta_d x_d}}=\sigma(-\beta_0-\beta^\top x),$$

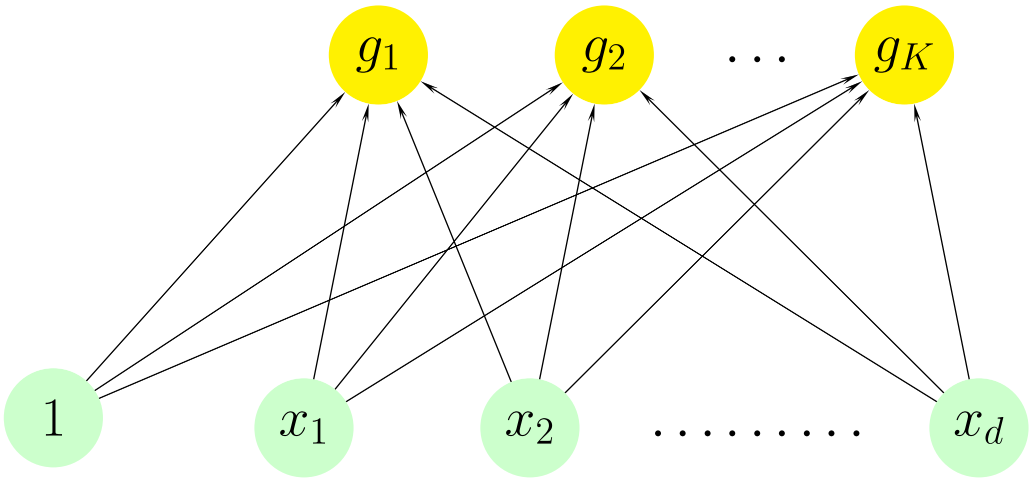
Разделяющая поверхность — линейная (гиперплоскость):

$$\mathrm{Pr}(Y=0|x)=\mathrm{Pr}(Y=1|x)=\frac{1}{2}\qquad\Leftrightarrow\qquad\beta_0+\beta_1 x_1+\dots+\beta_d x_d=0$$

$$g=\sigma\left(\beta_0+\sum_{j=1}^{d}\beta_j x_j\right)\approx\mathrm{Pr}(Y=1|X=x)$$



Случай K классов: $\mathscr{Y} = \{1, 2, \ldots, K\}$. Функция softmax:



$$g_k = \frac{\exp\left(\beta_{k0} + \sum\limits_{j=1}^{d} \beta_{kj} x_j\right)}{\sum\limits_{\ell=1}^{K} \exp\left(\beta_{\ell 0} + \sum\limits_{j=1}^{d} \beta_{\ell j} x_j\right)} \approx \mathsf{Pr}(k \,|\, x)$$

$$(k = 1, 2, \ldots, K)$$

Разделяющие поверхности (между каждой парой классов) снова линейные:

$$\mathsf{Pr}(Y = k \mid X = x) = \mathsf{Pr}(Y = k' \mid X = x) \Leftrightarrow \beta_{k0} + \beta_k^{\mathrm{T}} x = \beta_{k'0} + \beta_{k'}^{\mathrm{T}} x$$

Замечание 9.1 Можно преобразовать и переобозначить:

$$\mathsf{Pr}(Y = k \mid X = x) = \frac{e^{\beta_{k0} + \beta_k^{\mathrm{T}} x}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_{\ell 0} + \beta_\ell^{\mathrm{T}} x}} \qquad (k = 1, 2, \ldots, K-1),$$

$$\mathsf{Pr}(Y = K \mid X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} e^{\beta_{\ell 0} + \beta_\ell^{\mathrm{T}} x}}$$

9.1.1. Расчет параметров

Как найти $\beta_{10}, \beta_1, \beta_{20}, \beta_2, \ldots, \beta_{K0}, \beta_K$?

В логистической регрессии параметры обычно подбираются с помощью метода максимального правдоподобия.

Максимизации подвергается логарифмическая функция правдоподобия

$$\ell(\beta) = \sum_{i=1}^{N} \ln \mathsf{Pr}\left\{Y = y^{(i)} \,|\, X = x^{(i)}, \beta\right\} \to \max,$$

где

$$\beta = (\beta_{10}, \beta_1, \beta_{20}, \beta_2, \ldots, \beta_{K-1,0}, \beta_{K-1}), \quad \mathsf{Pr}(Y = k \,|\, X = x, \beta) = \mathsf{Pr}(Y = k \,|\, X = x).$$

Возможна регуляризация:

$$\sum_{i=1}^{N} \ln \mathsf{Pr}\left\{Y = y^{(i)} \,|\, X = x^{(i)}, \beta\right\} - \lambda \|\beta\|^2 \to \max$$

где вместо квадрата евклидовой нормы можно рассматривать 1-норму $\|\beta\|_1$

Вначале рассмотрим случай $K = 2$: $\mathscr{Y} = \{0, 1\}$.

$$g(x, \beta) = \mathsf{Pr}\,\{Y = 1 \,|\, X = x, \beta\} = \frac{1}{1 + e^{-\beta_0 - \beta^\top x}}.$$

Требуется максимизировать логарифмическую функцию правдоподобия

$$\ell(\beta) = \sum_{i=1}^{N} \Big(y^{(i)} \ln g(x^{(i)}, \beta) + (1 - y^{(i)}) \ln\big(1 - g(x^{(i)}, \beta)\big) \Big).$$

Упражнение 9.2 Докажите, что

$$\frac{\partial \ell(\beta)}{\partial \beta_0} = \sum_{i=1}^{N} \Big(y^{(i)} - g(x^{(i)}, \beta) \Big), \qquad \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} \Big(y^{(i)} - g(x^{(i)}, \beta) \Big)\, x_j^{(i)}.$$

Теперь можем воспользоваться методом градиентного спуска или более продвинутым методом оптимизации (сопряженных градиентов, BFGS, L-BFGS и др.).

Теперь рассмотрим случай $\mathscr{Y} = \{1, 2, \ldots, K\}$.

$$g(x, \beta) = \mathsf{Pr}\left\{Y = k \,|\, X = x, \beta\right\} = \frac{e^{\beta_{k0} + \beta_k^\top x}}{\sum_{\ell=1}^K e^{\beta_{\ell 0} + \beta_\ell^\top x}}.$$

Требуется максимизировать логарифмическую функцию правдоподобия

$$\ell(\beta) = \sum_{i=1}^N \sum_{k=1}^K I(y^{(i)} = k) \ln g_k(x^{(i)}, \beta) = \sum_{i=1}^N \sum_{k=1}^K I(y^{(i)} = k) \left(\beta_{k0} + \beta_k^\top x^{(i)} - \ln \sum_{\ell=1}^K e^{\beta_{\ell 0} + \beta_\ell^\top x^{(i)}}\right).$$

Упражнение 9.3 Докажите, что

$$\frac{\partial \ell(\beta)}{\partial \beta_{k0}} = \sum_{i=1}^N \left(I(y^{(i)} = k) - g(x^{(i)}, \beta)\right), \qquad \frac{\partial \ell(\beta)}{\partial \beta_k} = \sum_{i=1}^N x^{(i)} \left(I(y^{(i)} = k) - g(x^{(i)}, \beta)\right).$$

Замечание 9.4 Легко видеть, что максимизация логарифмической функции правдоподобия эквивалентна минимизации эмпирического риска

$$\widehat{R}(\beta) = -\frac{1}{N}\ell(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\Big(y^{(i)}\ln g(x^{(i)},\beta) + (1-y^{(i)})\ln\big(1-g\big(x^{(i)},\beta\big)\big)\Big),$$

если в качестве штрафной функции рассмотреть *кросс-энтропию* (или logloss-функцию):

$$L\big(g(x,\beta),y\big) = -y\ln g(x,\beta) - (1-y)\ln\big(1-g(x,\beta)\big).$$

Аналогично для случая K классов:

$$\widehat{R}(\beta) = -\frac{1}{N}\ell(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} I(y^{(i)}=k)\ln g_k(x^{(i)},\beta),$$

если в качестве штрафной функции рассмотреть *кросс-энтропию* (или logloss-функцию):

$$L\big(g(x,\beta),y\big) = -\sum_{k=1}^{K} I(y=k)\ln g_k(x,\beta).$$

9.1.2. Регуляризация

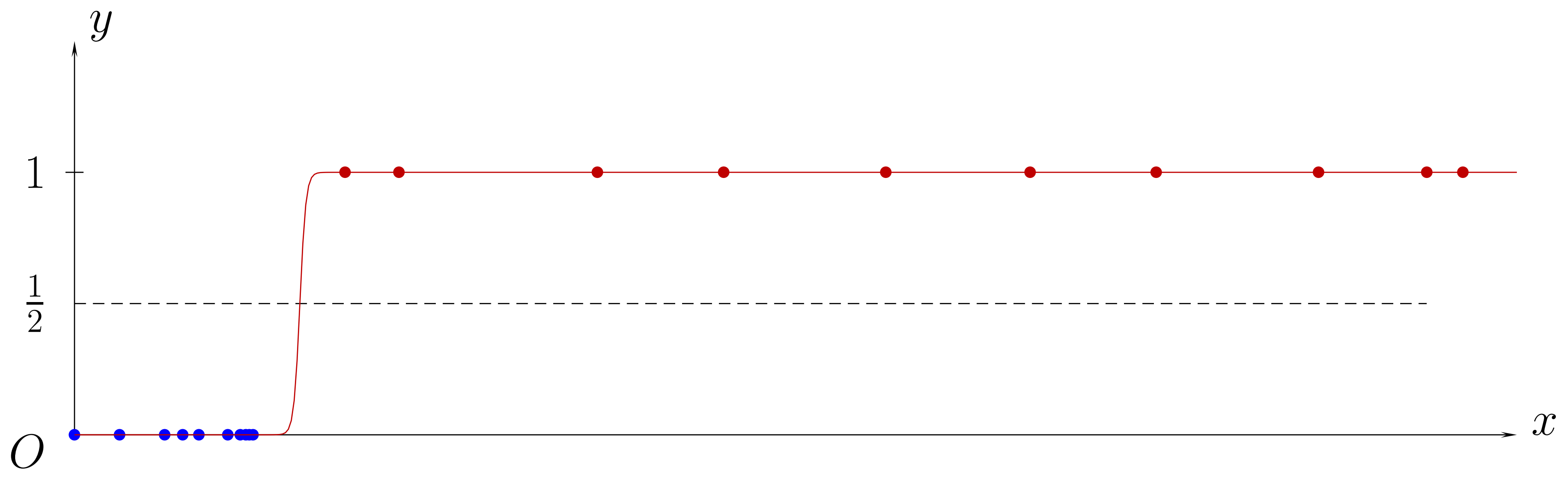
В L_2 норме:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} I(y^{(i)}=k)\ln g(x^{(i)},\beta)+\lambda\|\beta\|_2\to\min,$$

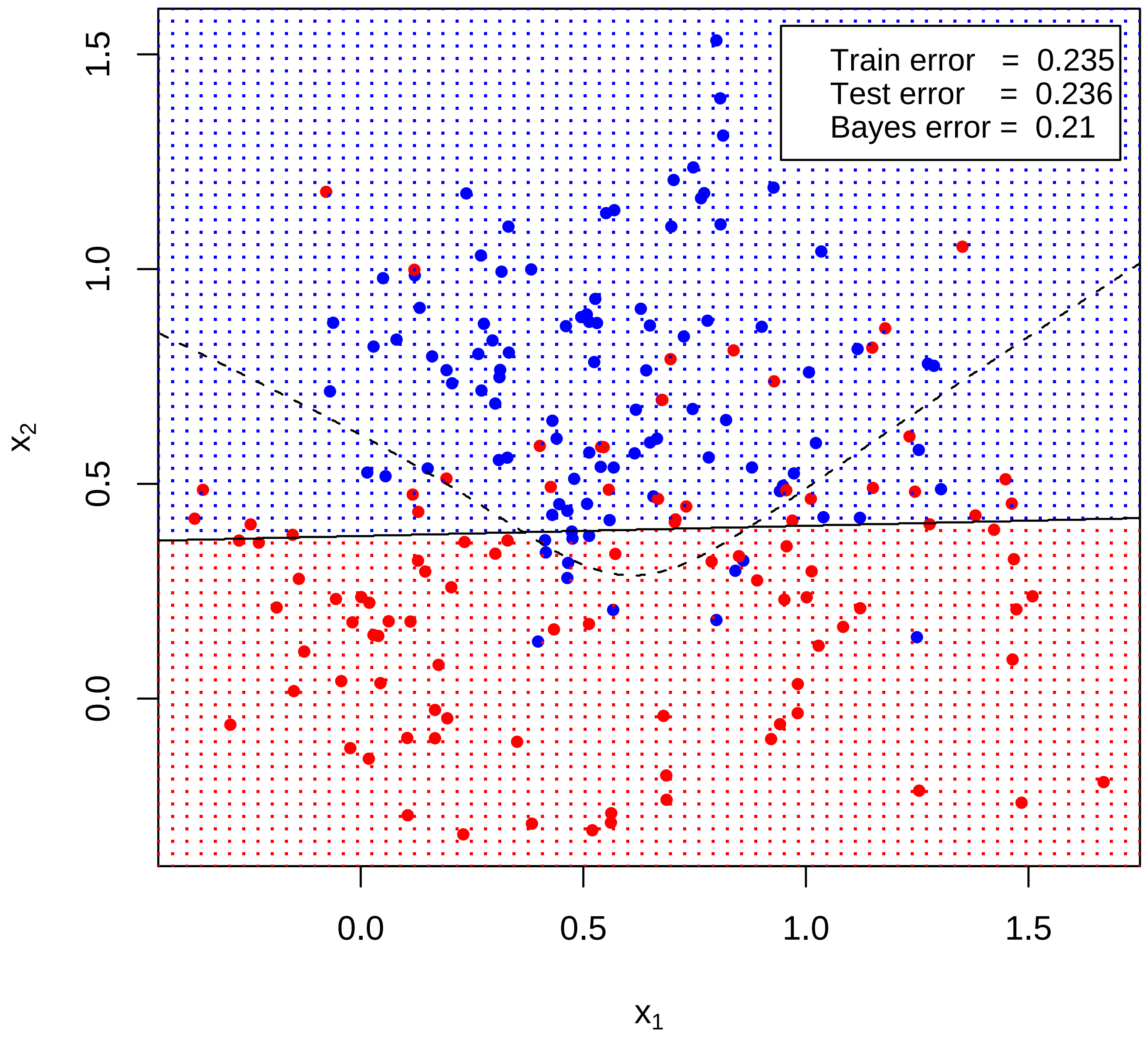
В L_1 норме:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} I(y^{(i)}=k)\ln g(x^{(i)},\beta)+\lambda\|\beta\|_1\to\min,$$

Без регуляризации, если классы линейно отделимы: $\beta\to-\infty$



Логистическая регрессия



Логистическая регрессия с признаками x_1, x_2, x_1^2, x_1x_2, x_2^2