Класификация

Доц. д-р Ивайло Пенев

Кат. „Компютърни науки и технологии“

Класификация

- При класификационните задачи предсказаните стойности за y принадлежат на дискретно множество с малко на брой елементи
- Практически много често се решават задачи, при които предсказаната стойност може да приема стойност 0 или 1 – двоична класификационна задача (binary classification problem)
- Пример – класификатор за spam email
  - $x^{(i)}$ - променливи от email
  - $y \in \{0,1\}$
  - $y = 1$ - email е spam – позитивен клас (означава се „+“)
  - $y = 0$ - email не е spam – негативен клас (означава се „-“)
  - При дадено $x^{(i)}$ съответната стойност $y^{(i)}$ се нар. „етикет“ (label) за съответния обучителен пример

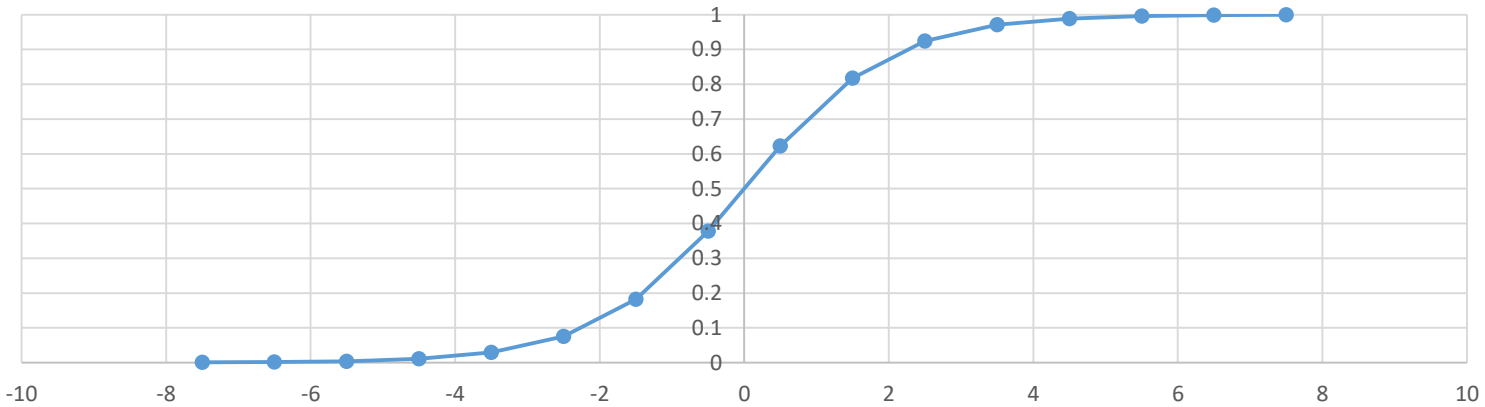
Представяне на хипотезата

- Възможно е да пренебрегнем факта, че y е дискретна стойност и да използваме линейна регресия за предсказване на y
- В много задачи горният подход не е удачен – не се интересуваме от стойности, при които y се отдалечава силно от 1 или от 0
- Променяме формата на хипотезата $h_\theta(x)$, така че да удовлетворява условието:

$$0 \leq h_\theta(x) \leq 1$$

Сигмоидна (Sigmoid function) или логистична (Logistic function) функция

- $h_\theta(x) = g(\theta^T x)$
- $z = \theta^T x$
- $g(z) = \frac{1}{1+e^{-z}}$



- Функцията $g(z)$ съпоставя на всяко реално число z число от интервала (0,1)
- При тази постановка хипотезата $h_\theta(x)$ показва каква е **вероятността** изходната стойност да бъде 1
- Напр. $h_\theta(x) = 0.7$ означава вероятност 70% за y=1
- Вероятността за y=0 е допълнение – т.е. 30%
- $h_\theta(x) = \mathrm{P}(\mathrm{y} = 1|x;\theta) = 1 - P(y = 0|x;\theta)$
- $P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1$

Определяне на изходната стойност y при класификацията

- $h_\theta(x) \geq 0.5 \rightarrow y = 1$
- $h_\theta(x) < 0.5 \rightarrow y = 0$
- Когато параметърът z на ф-та g има стойност по-голяма или равна на 0, то g има стойност по-голяма или равна на 0.5, т.е.:
- $g(z) \geq 0.5$ при $z \geq 0$
- Важни случаи:
  - При $\mathrm{z} = 0, e^0 = 1 \Rightarrow g(z) = 1/2$
  - При $\mathrm{z} \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z) = 1$
  - При $\mathrm{z} \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0$

Практическо значение на функцията g(z)

- За функцията g(z) с параметър $z = \theta^T x$ горното означава:

$$h_\theta(x) = g(\theta^T x) \geq 0.5 \text{ при } \theta^T x \geq 0$$

- Следователно:

$$\theta^T x \geq 0 \Rightarrow y = 1$$
$$\theta^T x < 0 \Rightarrow y = 0$$

- Практически такава функция на хипотезата създава разделителната линия между стойностите на x за y=1 и y=0
- Поради това хипотезата се нарича граница (decision boundary)

Пример

- Дадени е вектор с параметри:
- $\theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix}$
- $y = 1, \text{ако } 5 + (-1)x_1 + 0x_2 \geq 0$
- $\Rightarrow 5 - x_1 \geq 0 \Rightarrow x_1 \leq 5$
- Следователно границата е вертикална права линия, разположена при $x_1 = 5$
- Не е задължително ф-та $g(\theta^T x)$ да бъде права линия. Може да бъде например окръжност ($z = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2$) както и всяка друга повърхнина