3. Classification and logistic regression

Outline

Problem setup

- example: classifying emails as spam or not spam
- *features*: words in email, sender, etc.
- *response*: spam or not spam
- goal: learn a model that predicts spam from features
- *training data*: pairs of features and labels
- It is similar to linear regression, but the response variable is binary.

Outline

Logistic regression model

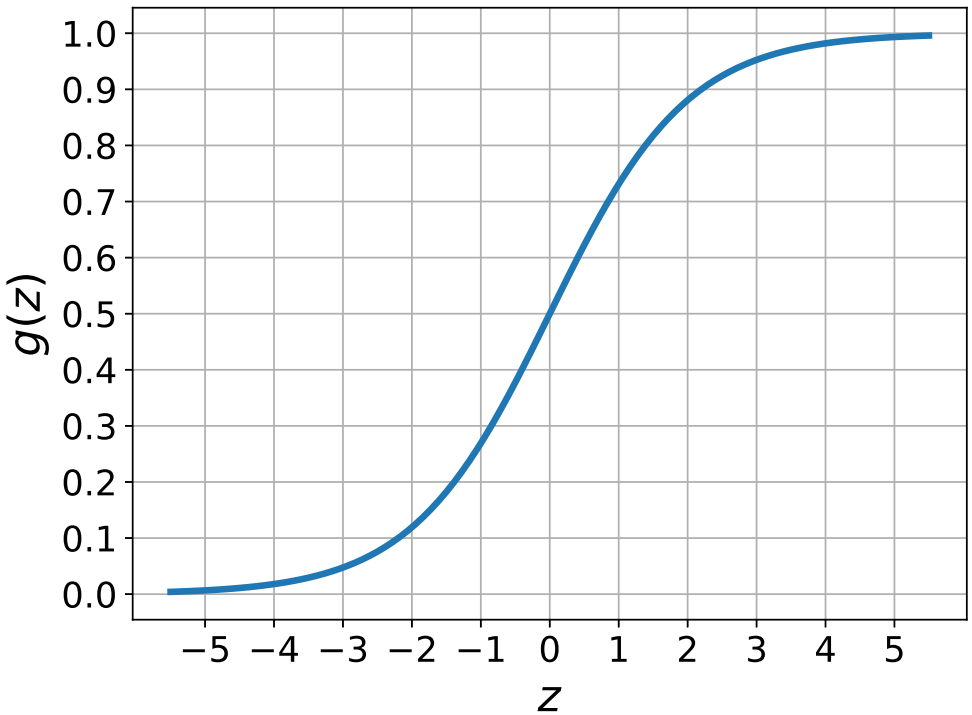
- data:
$$\{(x^{(i)}, y^{(i)}) \mid i = 1, \cdots, n\},$$
- $x^{(i)} \in \mathbf{R}^{d+1}$ is the *augmented feature* vector
- $y^{(i)} \in \{0, 1\}$ is the *label* variable
- *logistic* model (*hypothesis*):
$$P(y = 1|x; \theta) = h(x; \theta) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d)}} = \frac{1}{1 + e^{-\theta^T x}}$$

The logistic function

- also called the *sigmoid* functions:

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

- $g'(z) = g(z) \cdot (1 - g(z))$



Maximum likelihood estimation

- probability of observing y given x

$$P(y = 1|x; \theta) = h(x; \theta)$$
$$P(y = 0|x; \theta) = 1 - h(x; \theta)$$

- write it compactly

$$P(y|x; \theta) = (h(x; \theta))^y (1 - h(x; \theta))^{1-y}$$

Maximum likelihood estimation

- likelihood of all data

$$L(\theta) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{n} (h(x^{(i)};\theta))^{y^{(i)}} (1 - h(x^{(i)};\theta))^{1-y^{(i)}}$$

- log-likehood:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} y^{(i)} \log h(x^{(i)};\theta) + (1 - y^{(i)}) \log(1 - h(x^{(i)};\theta))$$

- gradient ascent to maximize $\frac{1}{n}\ell(\theta)$

$$\theta_j := \theta_j + \alpha \cdot \frac{1}{n} \frac{\partial \ell(\theta)}{\partial \theta_j} = \theta_j + \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h(x^{(i)}, \theta) \right) x_j^{(i)}$$

Outline

Multi-class classification

- data: $\{(x^{(i)}, y^{(i)}) \mid i = 1, \cdots, n\}$,
- $x^{(i)} \in \mathbf{R}^{d+1}$ is the *augmented feature* vector
- $y^{(i)} \in \{1, 2, \cdots, m\}$ is the *label* variable
- *softmax* model (*hypothesis*)

$$P(y = k|x; \theta) \propto e^{\theta_k^T x}$$

$$P(y = k|x; \theta) = \frac{e^{\theta_k^T x}}{\sum_{j=1}^{m} e^{\theta_j^T x}}$$

- $\theta_k \in \mathbf{R}^{d+1}$ is the parameter vector for class k

Softmax function

- generalization of the logistic function

$$P(y = k|x;\theta) = \frac{e^{\theta_k^T x}}{\sum_{j=1}^{m} e^{\theta_j^T x}}$$

- degeneracy in θ

$$P(y = k|x;\theta) = \frac{e^{(\theta_k - \theta_m)^T x}}{\sum_{j=1}^{m} e^{(\theta_j - \theta_m)^T x}}$$

- multiple choices of θ that give the same probability
- to remove the degeneracy, fix $\theta_m = 0$

Maximum likelihood estimation

- likelihood of a single data point

$$P(y^{(i)}|x^{(i)};\theta) = \prod_{k=1}^{m} (P(y=k|x^{(i)};\theta))^{1\{y^{(i)}=k\}}$$

- log-likelihood of a single data point

$$\ell(\theta) = \sum_{k=1}^{m} 1\{y^{(i)}=k\} \log P(y=k|x^{(i)};\theta)$$

- log-likelihood of all data

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{m} 1\{y^{(i)}=k\} \log P(y=k|x^{(i)};\theta)$$

- gradient ascent to maximize $\ell(\theta)/n$