MLE, Logistic Regression

Iran University of Science and Technology

M. S. Tahaei, PhD.

Fall 2024

Courtesy: slides are adopted partly from Dr. Soleymani, Sharif University

Outline

- Maximum-Likelihood (ML) estimation
- Logistic Regression

Parametric density estimation

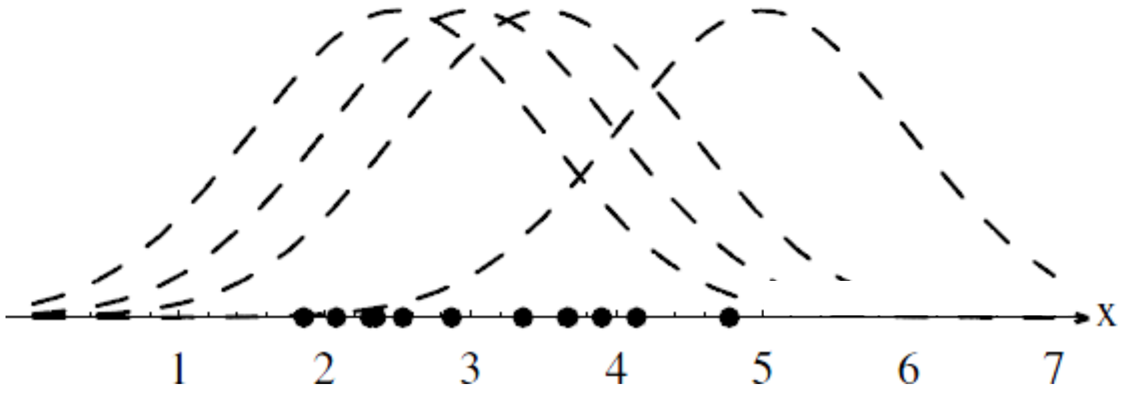
- Estimating the probability density function $p(\boldsymbol{x})$, given a set of data points $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$ drawn from it.
- Assume that $p(\boldsymbol{x})$ in terms of a specific functional form which has a number of adjustable parameters.
- Methods for parameter estimation
  - Maximum likelihood estimation
  - Maximum A Posteriori (MAP) estimation

Parametric density estimation

- Goal: estimate parameters of a distribution from a dataset $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$
  - $\mathcal{D}$ contains N independent, identically distributed (i.i.d.) training samples.

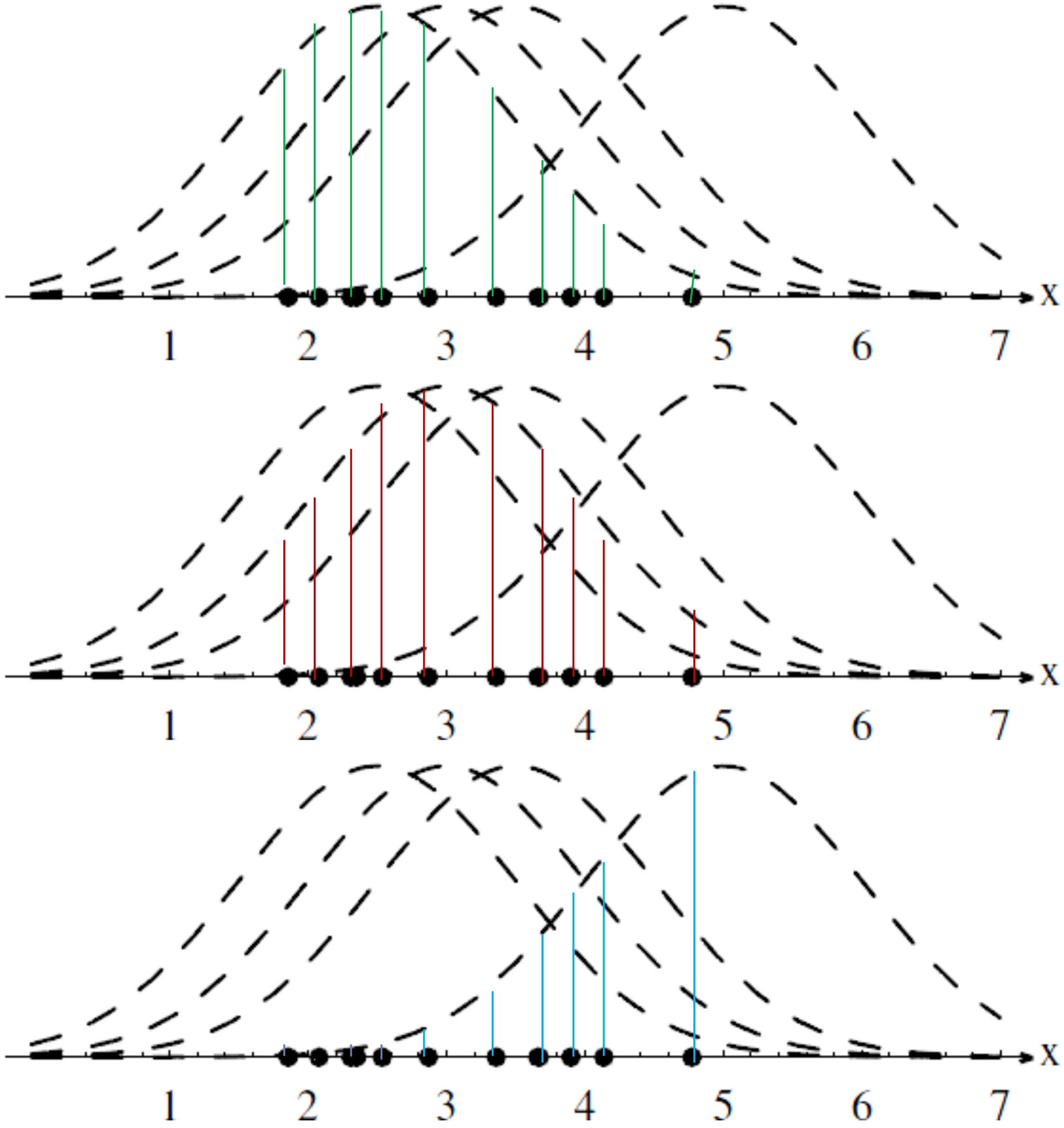
- We need to determine $\boldsymbol{\theta}$ given $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$
  - How to represent $\boldsymbol{\theta}$?
    - $\boldsymbol{\theta}^*$ or $p(\boldsymbol{\theta})$?

Example



$$P(x|\mu) = N(x|\mu, 1)$$

Example



Maximum Likelihood Estimation (MLE)

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given data.

- Likelihood is the conditional probability of observations $\mathcal{D} = \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\right\}$ given the value of parameters $\boldsymbol{\theta}$
  - Assuming i.i.d. observations:

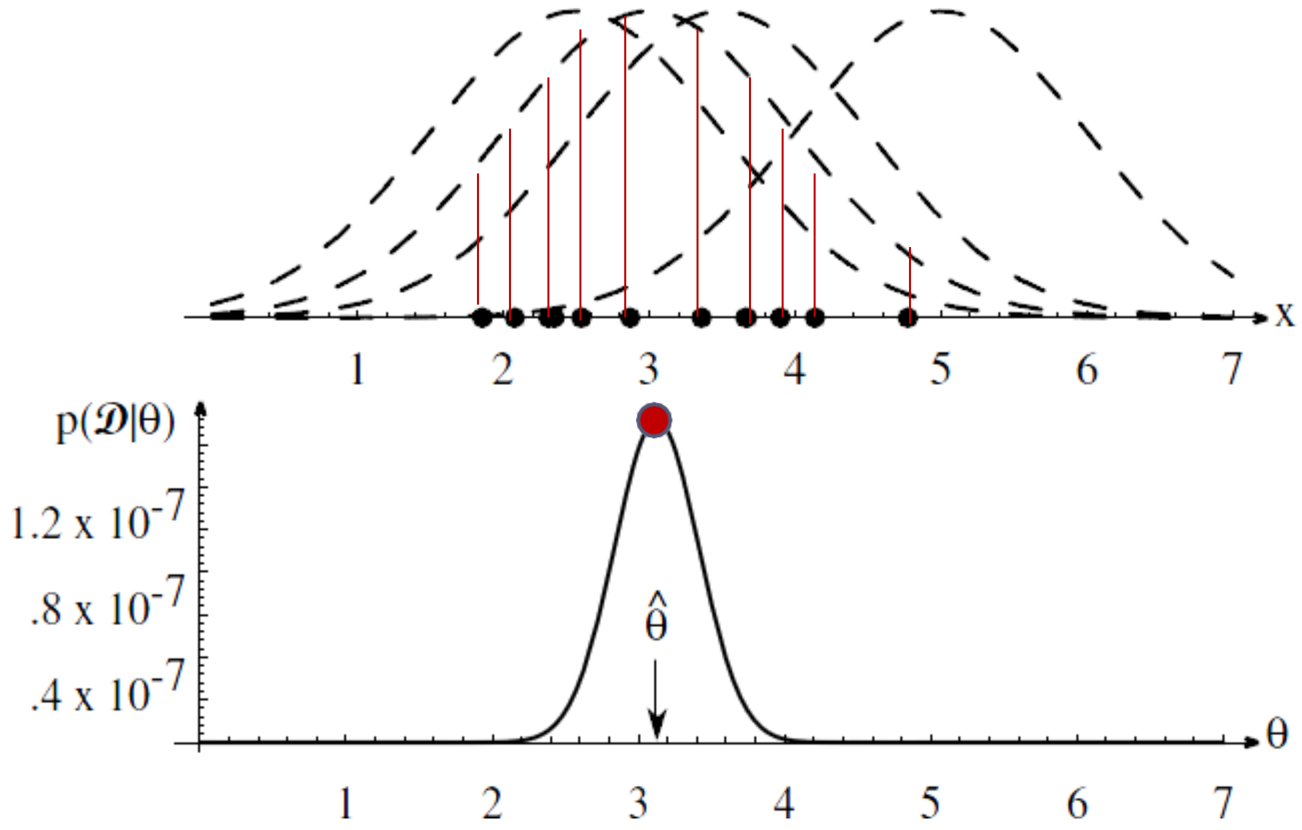
$$p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

likelihood of $\boldsymbol{\theta}$ w.r.t. the samples

- Maximum Likelihood estimation

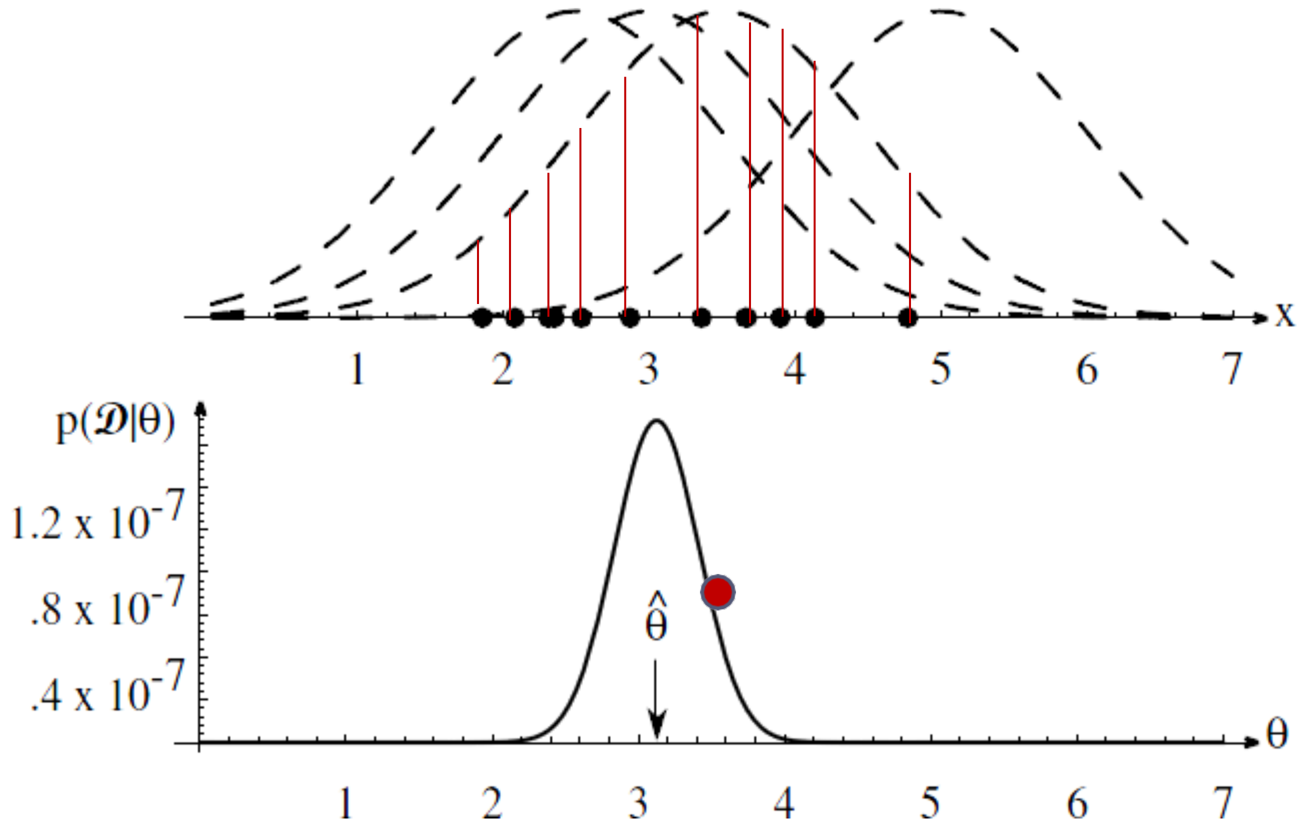
$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)



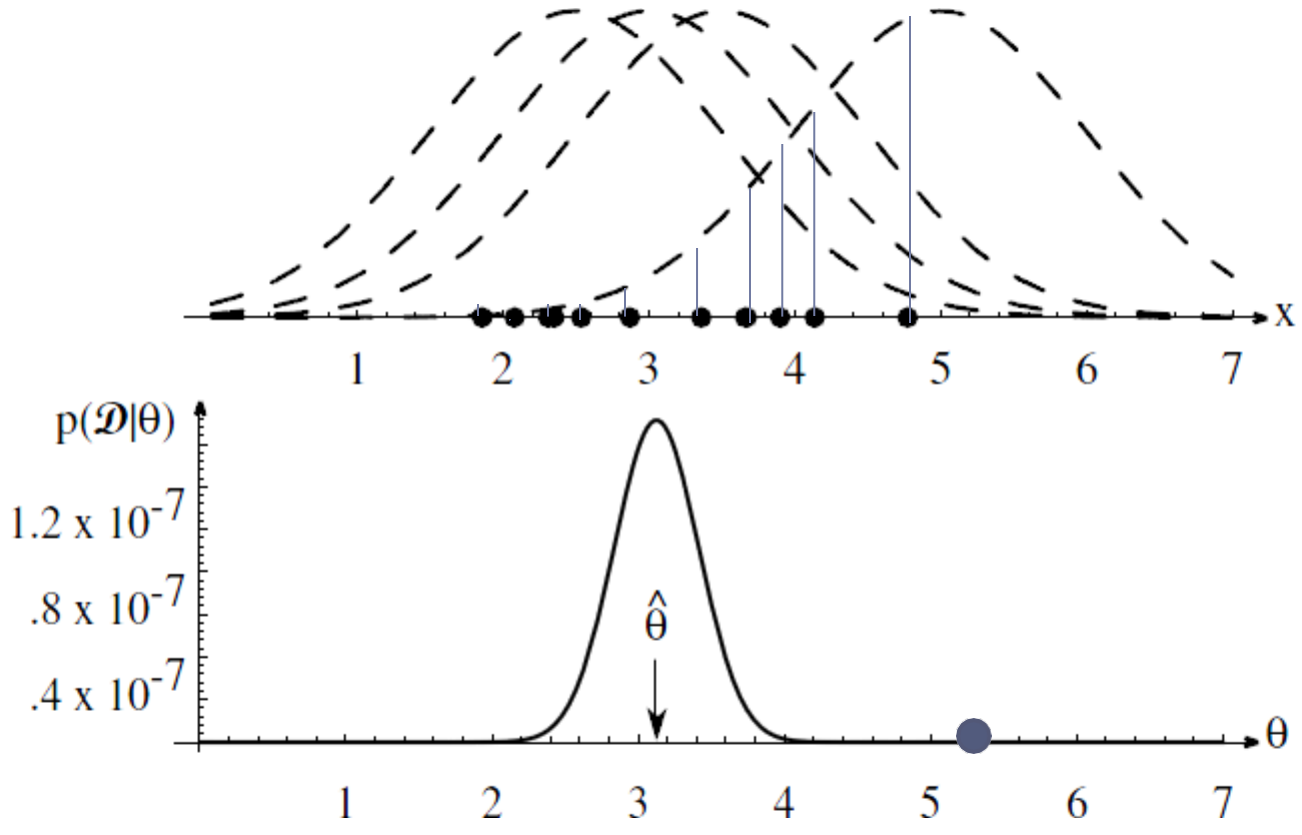
θ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



θ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



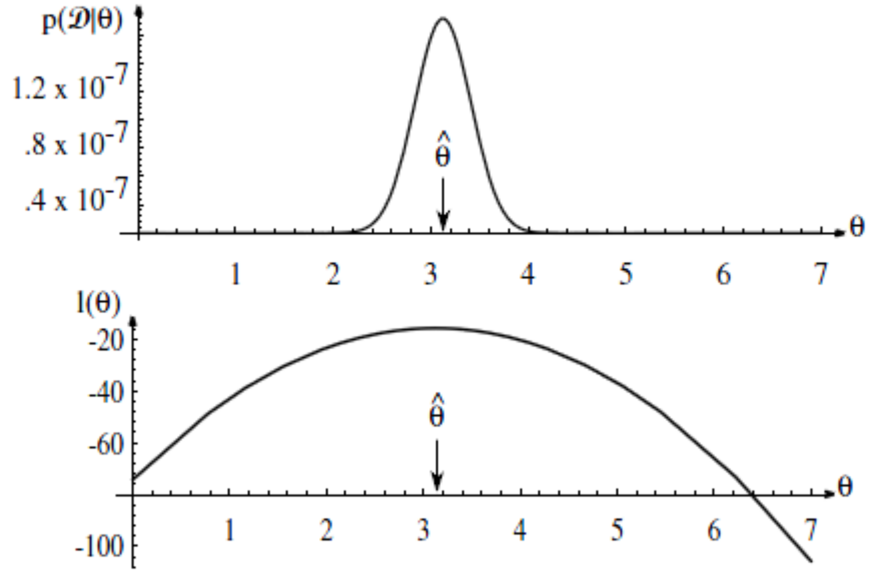
θ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{1=i}^{N} p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

- Thus, we solve $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$ to find global optimum



MLE. Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^{x}(1-\theta)^{1-x}$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}}(1-\theta)^{1-x^{(i)}}$$

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta) = \sum_{i=1}^{N} \{x^{(i)} \ln\theta + (1-x^{(i)}) \ln(1-\theta)\}$$

$$\frac{\partial \ln p(\mathcal{D}|\theta)}{\partial\theta} = 0 \Rightarrow \theta_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N} = \frac{m}{N}$$

MLE. Bernoulli: example

- Example: $\mathcal{D} = \{1,1,1\}, \theta_{ML} = \frac{3}{3} = 1$
  - Prediction: all future tosses will land heads up

- Overfitting to $\mathcal{D}$

Logistic regression

- More general than discriminant functions:
  - $f(\boldsymbol{x};\boldsymbol{w})$ predicts posterior probabilities $P(y = 1|\boldsymbol{x})$

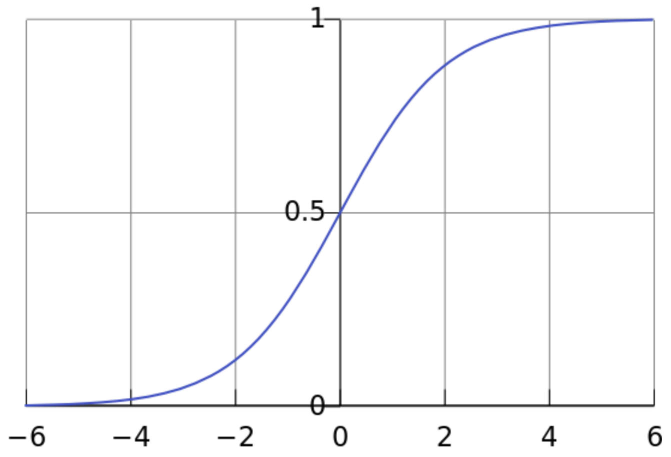
$$f(\boldsymbol{x};\boldsymbol{w}) = \sigma(\boldsymbol{w}^T\boldsymbol{x})$$

$$\boldsymbol{x} = [1, x_1, \dots, x_d]$$
$$\boldsymbol{w} = [w_0, w_1, \dots, w_d]$$

$\sigma(.)$ is an activation function

- Sigmoid (logistic) function
  - Activation function

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



Logistic regression

- $f(\boldsymbol{x};\boldsymbol{w})$: probability that $y = 1$ given $\boldsymbol{x}$ (parameterized by $\boldsymbol{w}$)

$$P(y = 1|\boldsymbol{x},\boldsymbol{w}) = f(\boldsymbol{x};\boldsymbol{w})$$

$K = 2$
$y \in \{0,1\}$

$$P(y = 0|\boldsymbol{x},\boldsymbol{w}) = 1 - f(\boldsymbol{x};\boldsymbol{w})$$

$f(\boldsymbol{x};\boldsymbol{w}) = \sigma(\boldsymbol{w}^T\boldsymbol{x})$
$0 \leq f(\boldsymbol{x};\boldsymbol{w}) \leq 1$
estimated probability of $y = 1$ on input x

- Example: Cancer (Malignant, Benign)
  - $f(\boldsymbol{x};\boldsymbol{w}) = 0.7$
  - 70% chance of tumor being malignant

Logistic regression: Decision surface

- Decision surface $f(\boldsymbol{x}; \boldsymbol{w}) = \text{constant}$
  - $f(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^T \boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{w}^T \boldsymbol{x})}} = 0.5$

- Decision surfaces are linear functions of $\boldsymbol{x}$

if $f(\boldsymbol{x}; \boldsymbol{w}) \geq 0.5$ then $y = 1$
else $y = 0$

Equivalent to

if $\boldsymbol{w}^T \boldsymbol{x} + w_0 \geq 0$ then $y = 1$
else $y = 0$

Logistic regression: ML estimation

- Maximum (conditional) log likelihood:

$$\boldsymbol{w} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ \operatorname{Log} \sum_{i=1}^{n} p\left(y^{(i)} \middle| \boldsymbol{w}, \boldsymbol{x}^{(i)}\right)$$

$$p\left(y^{(i)} \middle| \boldsymbol{w}, \boldsymbol{x}^{(i)}\right) = f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)^{y^{(i)}} \left(1 - f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right)^{(1-y^{(i)})}$$

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \sum_{i=1}^{n} \left[y^{(i)} \log \left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right) + (1 - y^{(i)}) \log \left(1 - f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right)\right]$$

Logistic regression: cost function

$$\boldsymbol{w} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, J(\boldsymbol{w})$$

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} \log p\left(y^{(i)} \middle| \boldsymbol{w}, \boldsymbol{x}^{(\boldsymbol{i})}\right)$$

$$= \sum_{i=1}^{n} -y^{(i)} \log\left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right) - (1 - y^{(i)}) \log\left(1 - f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right)$$

- No closed form solution for

$$\boldsymbol{\nabla}_{\boldsymbol{w}} J(\boldsymbol{w}) = 0$$

- However $J(\boldsymbol{w})$ is convex.

Logistic regression: Gradient descent

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} - \eta \nabla_{\boldsymbol{w}} J(\boldsymbol{w}^{t})$$

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = \sum_{i=1}^{n} \left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right) - y^{(i)} \right) \boldsymbol{x}^{(i)}$$

- Is it similar to gradient of SSE for linear regression?

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = \sum_{i=1}^{n} \left(\boldsymbol{w}^{T} \boldsymbol{x}^{(i)} - y^{(i)} \right) \boldsymbol{x}^{(i)}$$

Logistic regression: loss function

$$\text{Loss}\big(y, f(\boldsymbol{x};\boldsymbol{w})\big) = -y \times \log\big(f(\boldsymbol{x};\boldsymbol{w})\big) - (1-y) \times \log(1 - f(\boldsymbol{x};\boldsymbol{w}))$$

Since $y = 1$ or $y = 0$ $\Rightarrow$ $\text{Loss}\big(y, f(\boldsymbol{x};\boldsymbol{w})\big) = \begin{cases} -\log(f(\boldsymbol{x};\boldsymbol{w})) & \text{if } y = 1 \\ -\log(1 - f(\boldsymbol{x};\boldsymbol{w})) & \text{if } y = 0 \end{cases}$

How is it related to zero-one loss?

$$\text{Loss}(y, y) = \begin{cases} 1 & y \neq y \\ 0 & y = y \end{cases}$$

$$f(\boldsymbol{x};\boldsymbol{w}) = \frac{1}{1 + exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

Logistic regression: cost function

- Logistic Regression (LR) has a more proper cost function for classification than SSE and Perceptron

- Why is the cost function of LR also more suitable than?

$$J(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)\right)^2$$

  where $f(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^T \boldsymbol{x})$

  - The conditional distribution $p(y|\boldsymbol{x}, \boldsymbol{w})$ in the classification problem is not Gaussian (it is Bernoulli)
  - The cost function of LR is also convex

Logistic Regression (LR): summary

- LR is a linear classifier

- LR optimization problem is obtained by maximum likelihood
  - when assuming Bernoulli distribution for conditional probabilities whose mean is $\frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x})}}$

- No closed-form solution for its optimization problem
  - But convex cost function and global optimum can be found by gradient ascent

Resource

- Yaser S. Abu-Mostafa, MalikMaghdon-Ismail, and Hsuan Tien Lin,“**Learning from Data**”, 2012.
- C. Bishop, “Pattern Recognition and Machine Learning”.