LOGISTIC REGRESSION

Week06

Logistic Regression

Logistic Regression

$$f(x) = P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p}}$$

- Unknown parameters
  - $\beta_0, \beta_1, \dots, \beta_p$

- Logistic regression should estimate $\beta_0, \beta_1, \dots, \beta_p$ based on the given observations

Maximum Likelihood Estimation

- Maximum likelihood estimation
  - Method of estimating the parameters of statistical model
  - Given a statistical model, maximize likelihood

- Likelihood function
  - Suppose that data set $D = \{x_1, x_2, \dots, x_n\}$ consists of n independent and identically distributed(iid) samples coming from a distribution with an unknown probability density function $f(x)$
  - Assume $f(x)$ belongs to a certain type of distributions with parameters θ
  - Joint probability density function for all observations

$$f(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

because x_i is iid sample

  - Likelihood

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Likelihood

- Imagine the situation that a ball is drawn from the bag consisting of three blue balls and five white balls with replacement
  - Drawing is repeated five times and output is color of ball

	1	2	3	4	5
Case 1	blue	white	blue	white	white
Case 2	blue	blue	blue	blue	blue

Which case is more probable?

$$p(Case\ 1) = p(blue) \times p(white) \times p(blue) \times p(white) \times p(white)$$

$$p(Case\ 2) = p(blue) \times p(blue) \times p(blue) \times p(blue) \times p(blue)$$

Likelihood

- Imagine the situation that a ball is drawn from the bag consisting of three blue balls and five white balls with replacement
  - Drawing is repeated five times and output is color of ball

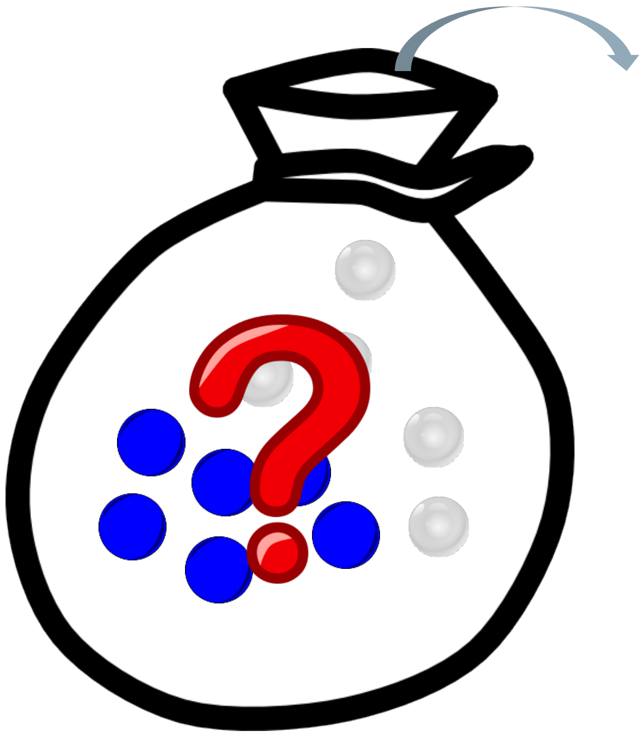
	1	2	3	4	5
Case 1	blue	white	blue	white	white
Case 2	blue	blue	blue	blue	blue

Which case is more probable?

⬇

Likelihood represents how much probable is observed data samples given statistical model

Example of Likelihood Function



Sampling with replacement

↓

3×

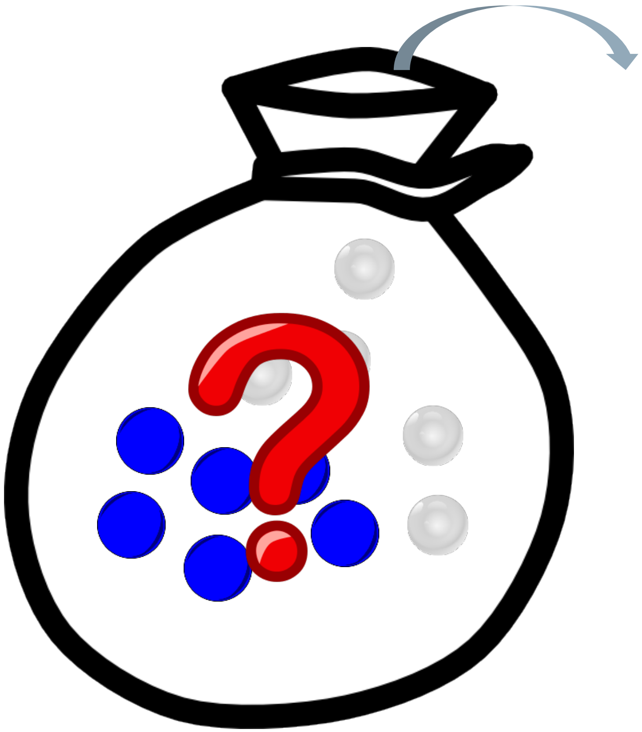
7×

- Want to estimate p_{blue} and p_{white} based on the sampling result

Example of Likelihood Function

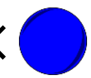
- There are only two outputs → Bernoulli distribution
- Bernoulli distribution: the probability distribution of a random variable which takes the value 1 with success probability of p and the value 0 with failure probability of $q = 1 - p$
  - For random variable following Bernoulli distribution,
    $$p(X = 1) = 1 - p(X = 0) = p = 1 - q$$
  - Probability mass function over possible outcomes y
    $$f(y; p) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}$$
    - This can also be expressed as
      $$f(y; p) = p^y (1 - p)^{1-y} \quad \text{for } y \in \{1,0\}$$
  - For Bernoulli distribution, p is θ
    - In this example, assume that blue ball is 1
      $$p = p_{blue}$$
      $$1 - p = p_{white}$$

Example of Likelihood Function



Sampling with replacement

↓

3× 

7× 

- Likelihood function
  - If blue ball, $f(1;p) = p$
  - If white ball, $f(0;p) = 1 - p$

$$\mathcal{L} = \prod_{i=1}^{10} f(y_i;p) = p^3(1-p)^7$$

  - Maximize $\mathcal{L}$ with respect to p

Example of Likelihood Function

- 1D data samples from Gaussian distribution with $\sigma = 1$

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12

- Likelihood function is function of parameter θ

$$\mathcal{L}(\theta;\mathbf{x}) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}$$

  - If $\theta = 2, \mathcal{L}(2) \approx 0.33 \times 0.09 \times 0.29 \times 0.03 \times 0.21 = 0.0000542619$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x;\theta)$	0.33	0.09	0.29	0.03	0.21

- Maximum likelihood estimation is method to find parameter to maximize likelihood function with given data samples

Maximum Likelihood Estimation

- Compare likelihood with different parameters
  - If $\theta = 2, \mathcal{L}(2) \approx 0.33 \times 0.09 \times 0.29 \times 0.03 \times 0.21 = 0.0000542619$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x;\theta)$	0.33	0.09	0.29	0.03	0.21

  - If $\theta = 3, \mathcal{L}(3) \approx 0.37 \times 0.31 \times 0.39 \times 0.17 \times 0.40 = 0.003041844$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x;\theta)$	0.37	0.31	0.39	0.17	0.40

  - If $\theta = 4, \mathcal{L}(4) \approx 0.15 \times 0.38 \times 0.19 \times 0.38 \times 0.27 = 0.001111158$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x;\theta)$	0.15	0.38	0.19	0.38	0.27

Maximum Likelihood Estimation

- Likelihood function

$$\mathcal{L}(\theta;\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(\theta^2-2x_i\theta+x_i^2)}{2}\right)$$

$$\propto \exp\left(-\sum_{i=1}^{n}(\theta^2 - 2x_i\theta + x_i^2)\right)$$

- When $\sum_{i=1}^{n}(\theta^2 - 2x_i\theta + x_i^2)$ is minimum, $\mathcal{L}(\theta;\mathbf{x})$ is maximized

$$n\theta^2 - 2\left(\sum_{i=1}^{n} x_i\right)\theta + \sum_{i=1}^{n} x_i^2$$

  - Second order equation of θ → There is a solution to minimize equation
- Example
  - https://www.geogebra.org/m/zOmGcvXq

※ Gaussian (Normal) Distribution

- The Gaussian distribution is a continuous probability distribution
  - probability density function

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

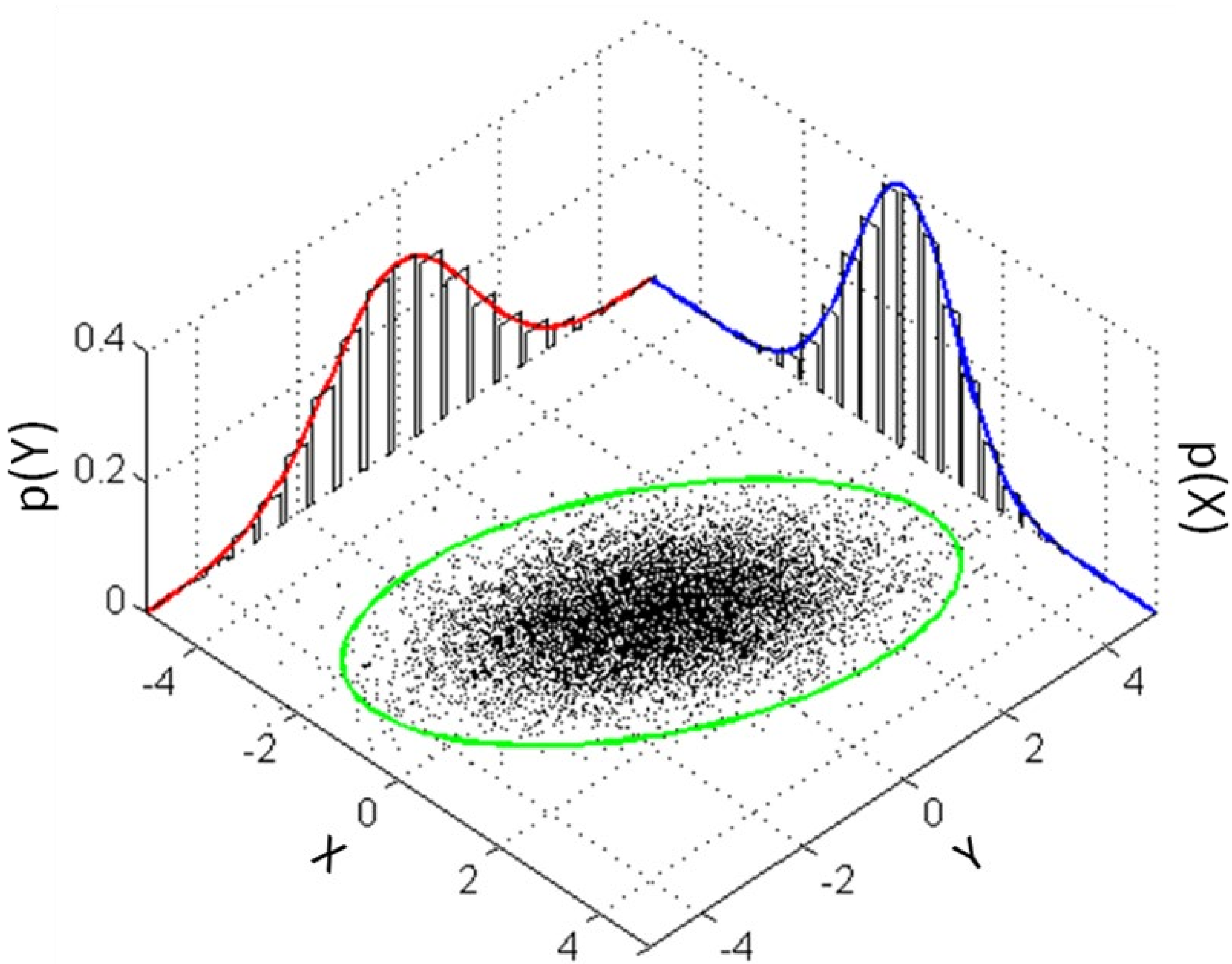
    - μ: mean or expectation of the distribution
    - σ: standard deviation
  - When $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution
- Multivariate normal distribution is a generalization of the 1D normal distribution
  - probability density function

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

    - p: dimensionality
    - $\boldsymbol{\mu}$: mean vector
    - $\boldsymbol{\Sigma}$: covariance matrix

※ Gaussian (Normal) Distribution

- Two dimensional normal distribution



How to Find Parameters for Logistic Regression?

- Output is 0 or 1 → Output follows Bernoulli distribution with parameter p

- Each sample has different p deepening on input

$$y_i \sim Bernoulli(P_i)$$

  - P_i is the probability that output value is 1 for i-th sample
  - Output of each sample follows Bernoulli distribution with parameter P_i

$$f(y_i) = P\{Y = y_i\} = {P_i}^{y_i}(1 - P_i)^{1-y_i}, \qquad y_i \in \{0,1\}$$

How to Find Parameters for Logistic Regression?

- Likelihood function of logistic regression model

$$\mathcal{L} = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}$$

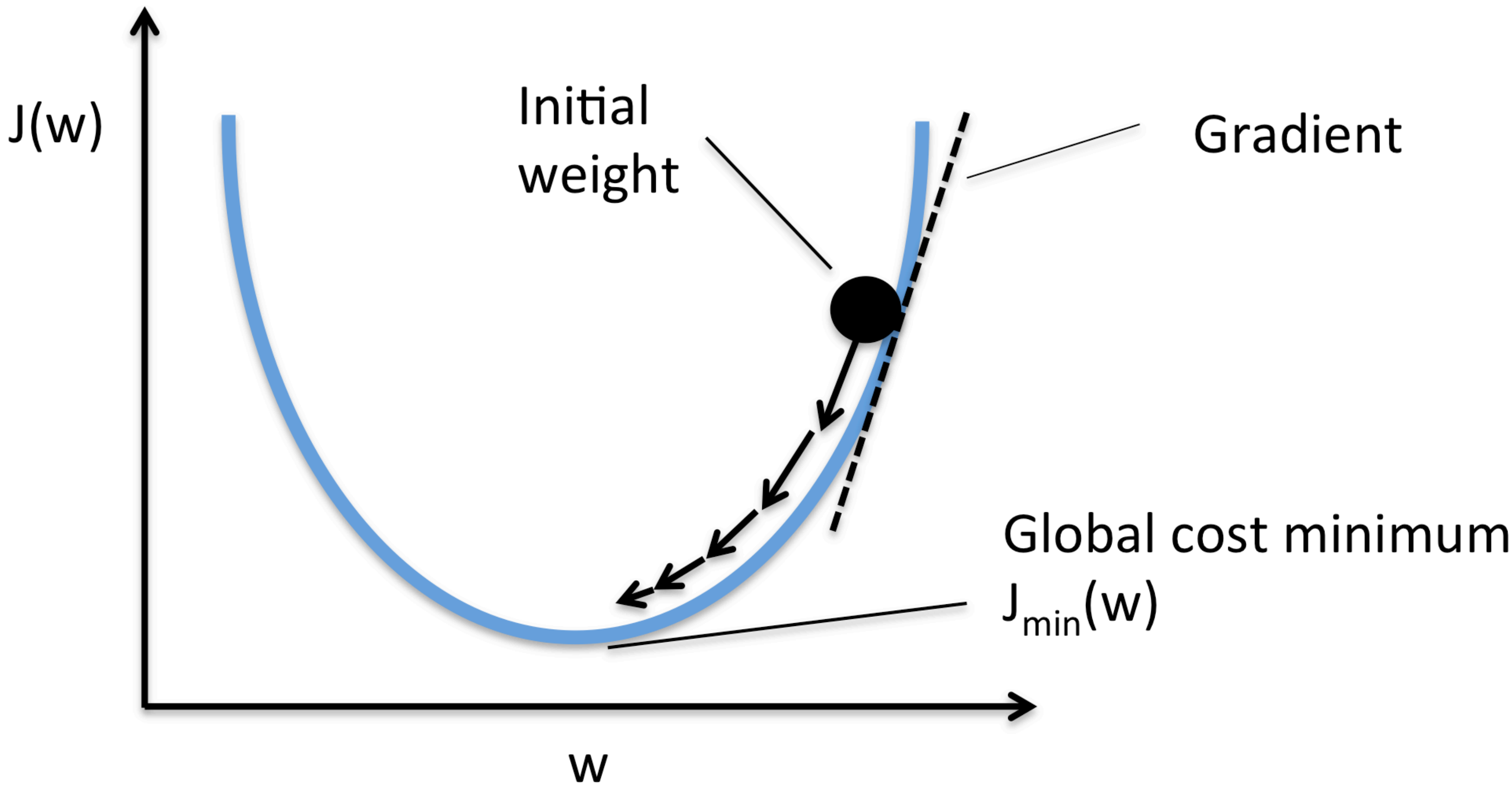
  - $P_i = P(y = 1|\mathbf{x}) = \frac{1}{1+e^{-\beta_0-\beta_1 x_1-\beta_2 x_2-\cdots-\beta_p x_p}}$

- Log-likelihood function

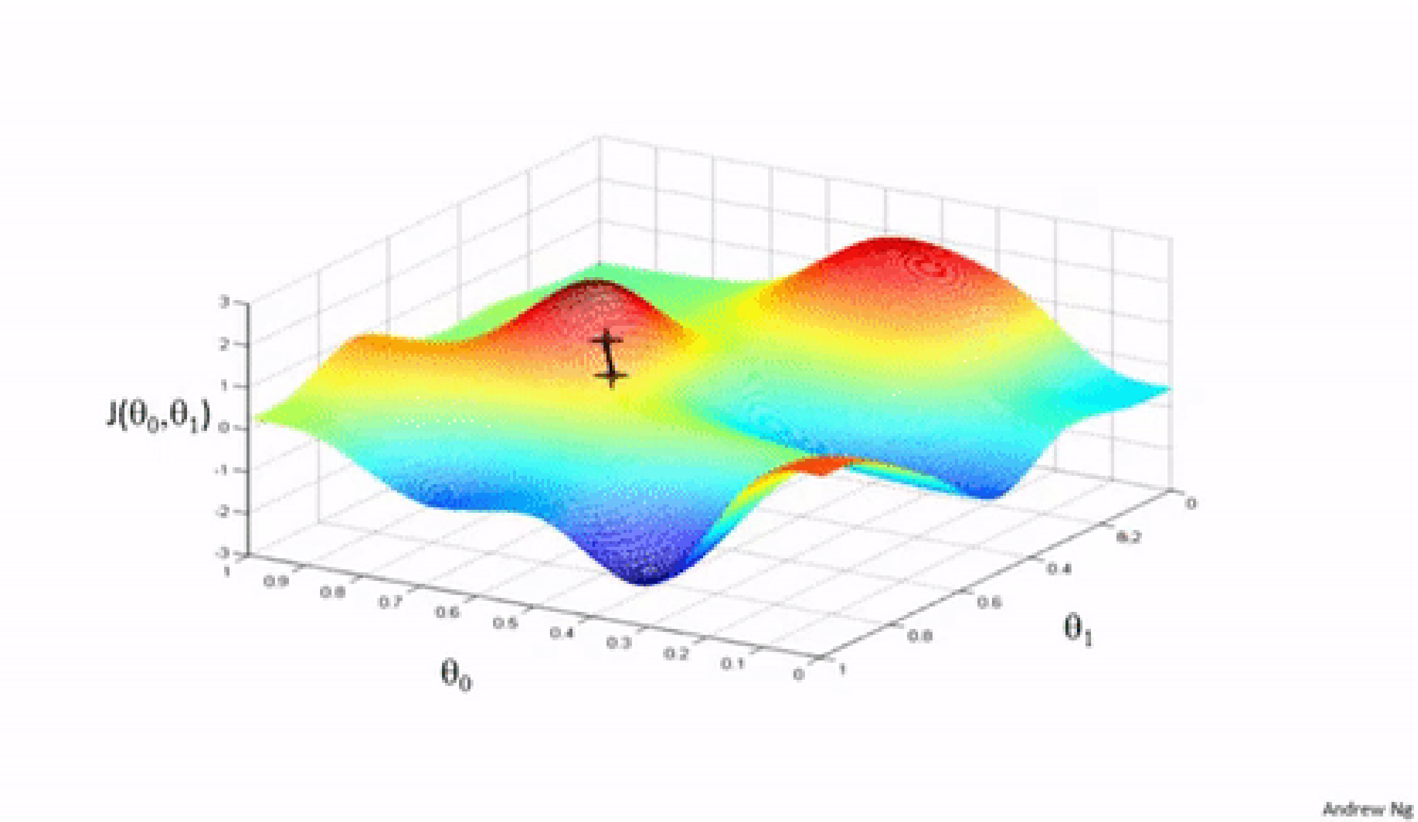
$$\log \mathcal{L} = \sum_{i=1}^{n} y_i \log P_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - P_i)$$

  - Find parameters $\beta_0, \beta_1, \dots, \beta_p$ to maximize $\log \mathcal{L}$

※ Gradient Descent



※ Gradient Descent



Odds and Odds Ratio

- Odds reflect the likelihood that the event will take place
  - In gambling, odds represent the ratio between the amounts staked by parties to a wager or bet

$$\frac{P(Wins)}{P(Losses)}$$

  - In logistic regression, odds represent the ratio between $P(y=1)$ and $P(y=0)$

$$\text{odds} = \frac{P(y=1)}{P(y=0)} = \frac{\dfrac{1}{1+e^{-\beta_0-\beta_1 x_1-\beta_2 x_2-\cdots-\beta_p x_p}}}{1-\dfrac{1}{1+e^{-\beta_0-\beta_1 x_1-\beta_2 x_2-\cdots-\beta_p x_p}}} = \exp\left(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\right)$$

- Odds ratio is the ratio between odds when unit increment of a variable

$$\text{odds ratio} = \frac{\text{odds when input is } x_1 = x+1}{\text{odds when input is } x_1 = x} = \frac{\exp(\beta_0+\beta_1(x+1)+\beta_2 x_2+\cdots+\beta_p x_p)}{\exp(\beta_0+\beta_1 x+\beta_2 x_2+\cdots+\beta_p x_p)} = e^{\beta_1}$$

  - Odd increases e^{β_1} times for every 1-unit increase in x_1

Logistic Regression: Odds

- A logistic model is one where the log-odds of the probability (logit function) of an event is a linear combination of independent or predictor variables (binary case)
- Logistic model

Logit function

$$\ln(\text{odds}) = \ln\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  - Let $P = P(y=1)$

$$\frac{P(y=1)}{P(y=0)} = \frac{P}{1-P}$$

$$g(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Link function

We can use other link functions

- Gompertz function

$$P = 1 - \exp(-\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p))$$
$$g(P) = \ln(-\ln(1 - P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Probit model

$$g(P) = F^{-1}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  - Normit model

$$g(P) = \Phi^{-1}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

    - $\Phi^{-1}(x)$ is inverse cumulative density function of normal distribution

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

Logistic Regression for Multi-class

- For K classes, $P(y_i = k)$ is the probability that ith data point belong to class k ($k \in \{1,2,3,\cdots,K\}$)
  - It is reasonable to select class k whose probability is the highest

How to extend logistic regression to multi-class classification problems?

Multinomial Logistic Regression

- Multinomial logistic regression assumes that log ratio between probabilities of two different classes is linear
  - Log linear model

$$
\begin{gathered}
\ln p(y_i = 1) = \boldsymbol{\beta}_1 \cdot \mathbf{x}_i - \ln Z \\
\ln p(y_i = 2) = \boldsymbol{\beta}_2 \cdot \mathbf{x}_i - \ln Z \\
\vdots \\
\ln p(y_i = K) = \boldsymbol{\beta}_K \cdot \mathbf{x}_i - \ln Z
\end{gathered}
$$

    - $\mathbf{x}_i = \left(1, x_{i1}, x_{i2,}, \ldots, x_{ip}\right)$
    - $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})$
    - $\boldsymbol{\beta}_k \cdot \mathbf{x}_i = \beta_{k0} + \beta_{k1} x_{i1} + \cdots + \beta_{kp} x_{ip}$

$$\Downarrow$$

$$
p(y_i = k) = \frac{1}{Z} e^{\boldsymbol{\beta}_k \cdot \mathbf{x}_i}
$$

$$
Z = \sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{x}_i}
$$

※ Multinomial distribution

- Multinomial distribution is a generalization of the binomial distribution
  - Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments with success probability p

$$p(k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

  - Example of binomial distribution is the distribution of the number of head when flipping a coin n times (in this case, $p = 0.5$)
    - Probability that k times head occur among n trials

$$p(k) = \frac{n!}{k!\,(n-k)!} 0.5^k 0.5^{n-k} = \frac{n!}{k!\,(n-k)!} 0.5^n$$

  - In multinomial distribution, possible outcome is more than two and each outcome has its own probability to occur, $(p_1, \ldots, p_d)$
    - $p_1 + \cdots + p_d = 1$
    - d is the number of possible outcomes
    - $n_{\mathbf{x}} = \sum_{i=1}^{d} x_i$

$$p(\mathbf{x} = (x_1, x_2, \ldots, x_d)) = \frac{n_{\mathbf{x}}!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d}$$

Likelihood Function

- Likelihood function

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{k=1}^{K} P_{ik}^{v_{ik}} \quad , \qquad v_{ik} = \begin{cases} 1, y_i = k \\ 0, y_i \neq k \end{cases}$$

  - $P_{ik} = p(y_i = k)$

- Log likelihood function

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{k=1}^{K} v_{ik} \log P_{ik}$$

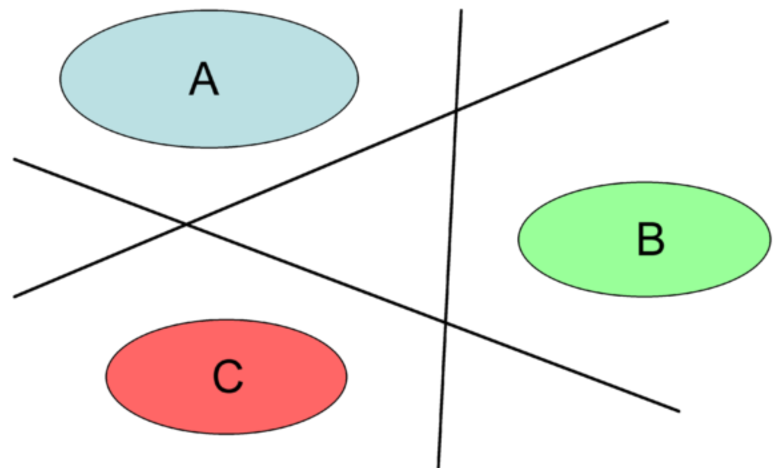
  - Through maximum likelihood estimation, determine $\boldsymbol{\beta}_k$ as the same as in binary logistic regression

Multiclass Classification Using Binary Classifiers

- There are other ways to get multi-class classifiers by combining binary classifiers
  - For multiclass classification commonly used approach is to construct K separate binary classifiers
    - Each model is trained using the data from class C_k as the positive examples and the data from the remaining $K-1$ classes as the negative examples

$$y(\mathbf{x}) = \max_k y_k(\mathbf{x})$$

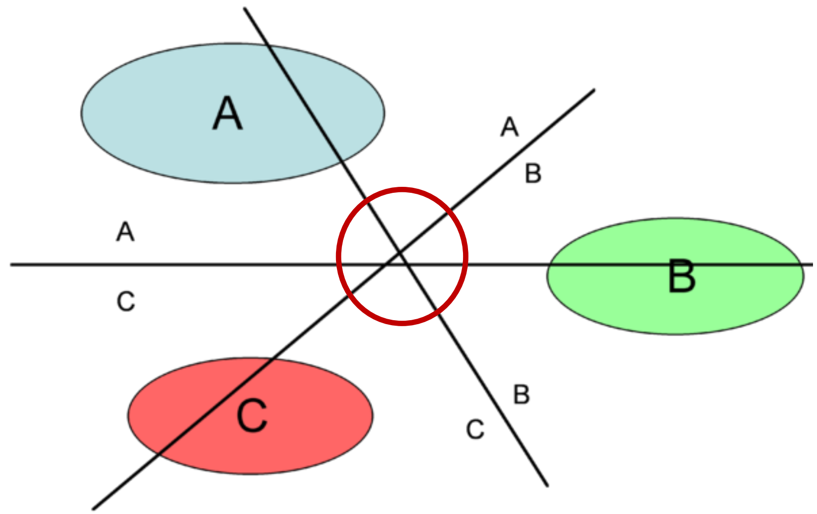
$\rightarrow$ One-versus-the rest approach



- Problems of one-versus-the rest approach
  - Because each classifier was trained on different task, there is no guarantee that the real-values quantities $y_k(\mathbf{x})$ will have appropriate scales
  - Imbalance of data on training

Multiclass Classification Using Binary Classifiers

- Another approach is to train $K(K-1)/2$ different 2-class classifiers on all possible pairs of classes
  - Classify test points according to which class has the highest number of votes
  - → one-versus-one approach



- Problems of one-versus-one approach
  - It can lead to ambiguities in the resulting classification
  - For large K, it requires significantly more training time