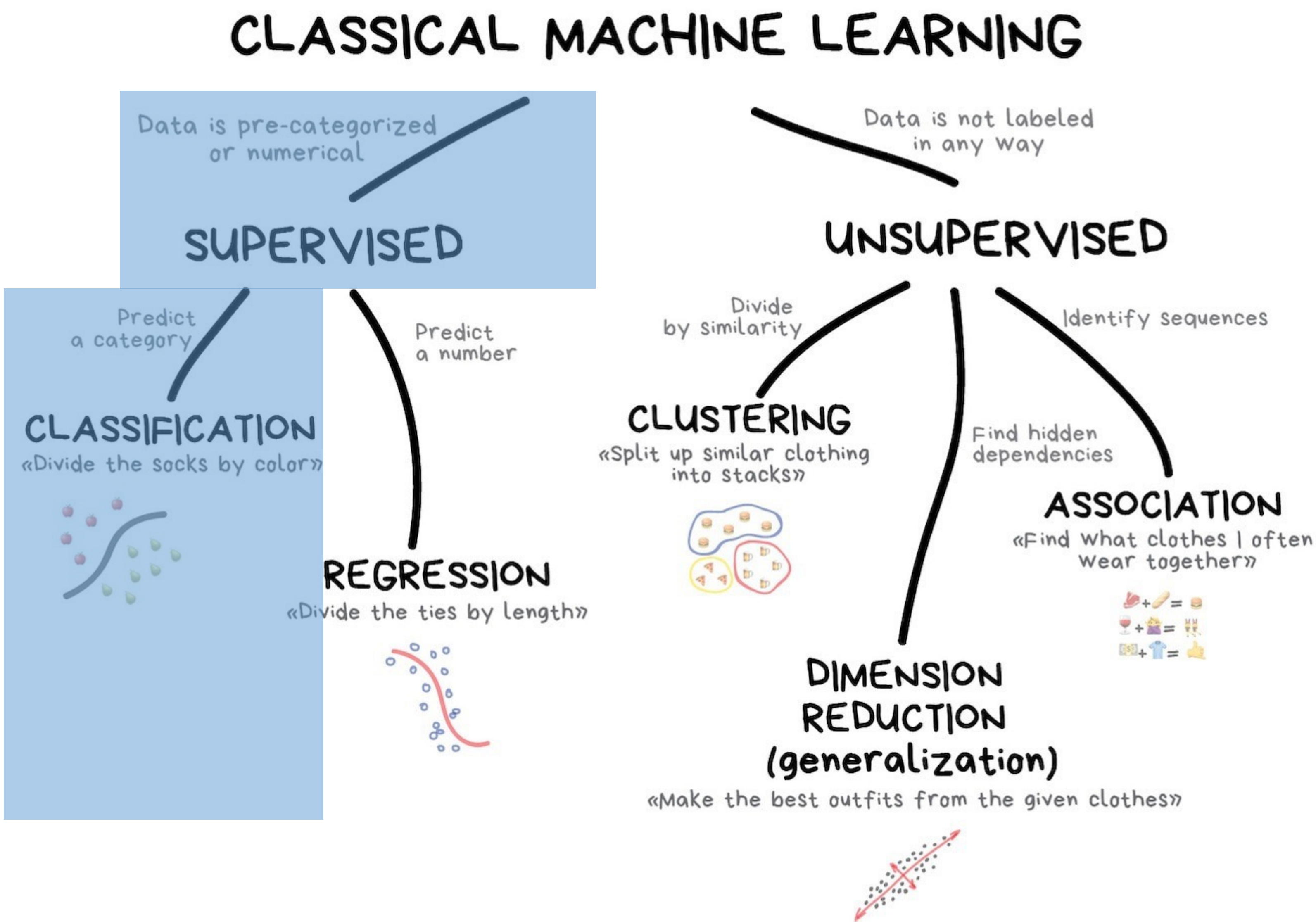
Logistic Regression

Topics Covered in This Class



Supervised Classification

- **Classification problem**
  - Output is categorical variable
    - ✓ Spam / Non-spam
    - ✓ Male / Female
    - ✓ Long / Medium / Short
  - Binary classification problem
    - ✓ The number of categories is 2.
    - ✓ Generally, these two categories are denoted as 0 and 1.
      - – 0 and 1 are not integer in this case

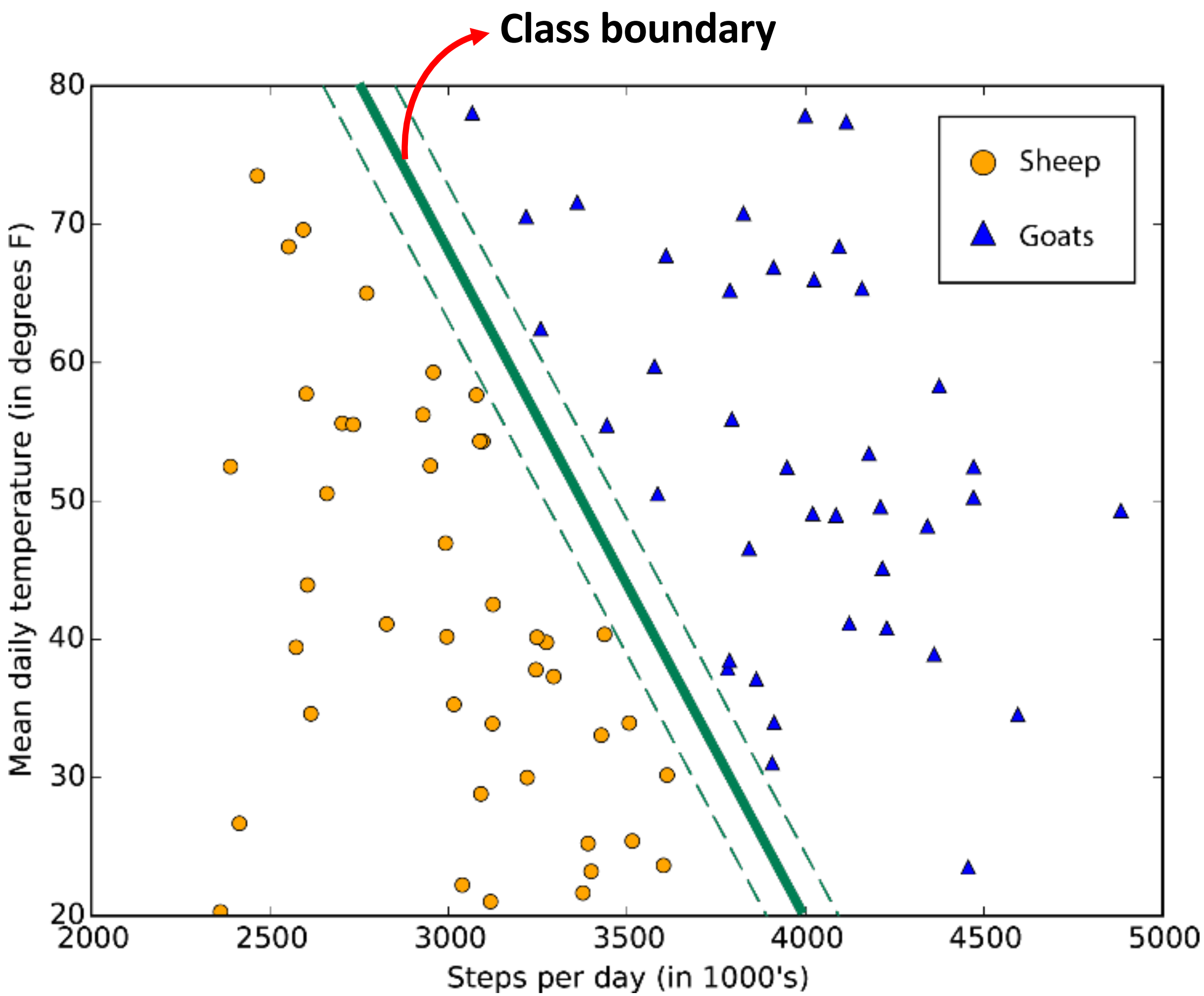
$$y \in \{0, 1\}$$

  - Multi-class classification problem
    - ✓ More than two classes

$$y \in \{1,2, \dots , C\}, \qquad C > 2$$

Supervised: Classification

- **Which one is a sheep?**



Type of Classifiers

- **A classifier is a function that assigns a class label $\hat{y}$ to a sample $\boldsymbol{x}$.**

$$\hat{y} = f(\boldsymbol{x})$$

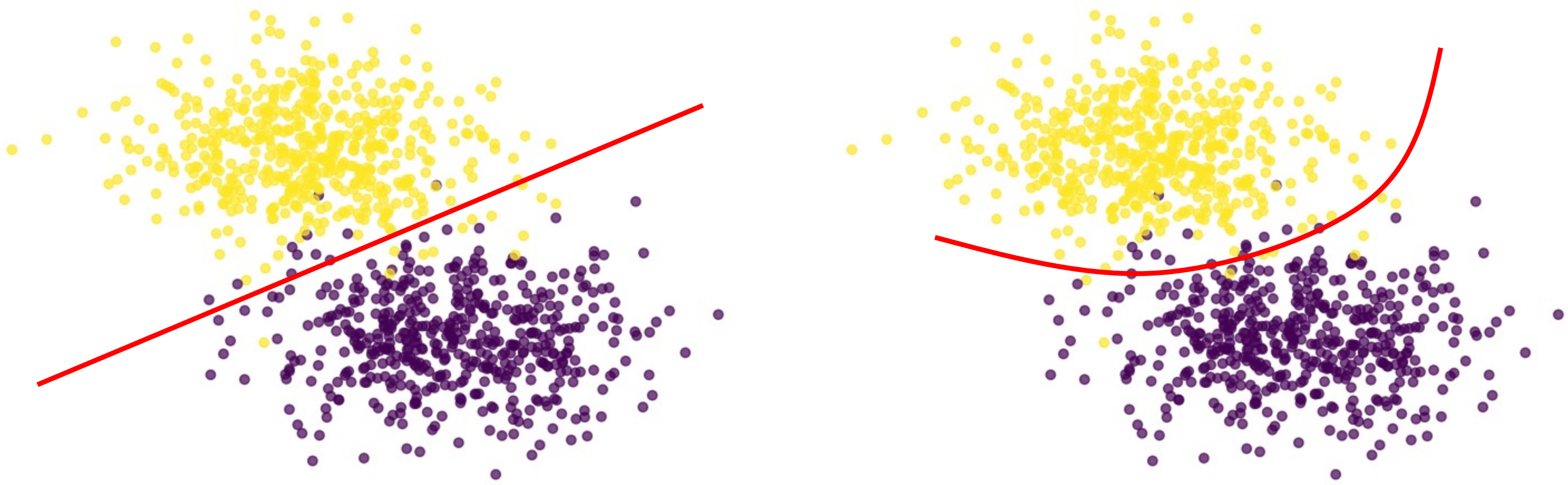
- **A probabilistic classifier obtains conditional distributions** $P(y|\boldsymbol{x})$, meaning that for a given $\boldsymbol{x}$, they assign probabilities to all $y \in \{1, \dots, C\}$
  - Hard classification

$$\hat{y} = \arg\max_{y} P(Y = y|\boldsymbol{x})$$

The Decision Boundary of Classifiers

- **Decision boundary**

$$y = f(x), \qquad y \in \{1, 2, \dots, C\}$$



Logistic Regression

- **Logistic regression**
  - Regression model where the dependent variable is categorical
  - The probabilities describing the possible outcomes is modeled as explanatory variables

$$f(x) = P(Y|X)$$

  - Logistic regression is a linear classification algorithm

$$f\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = f(\boldsymbol{\beta} \cdot \boldsymbol{x})$$

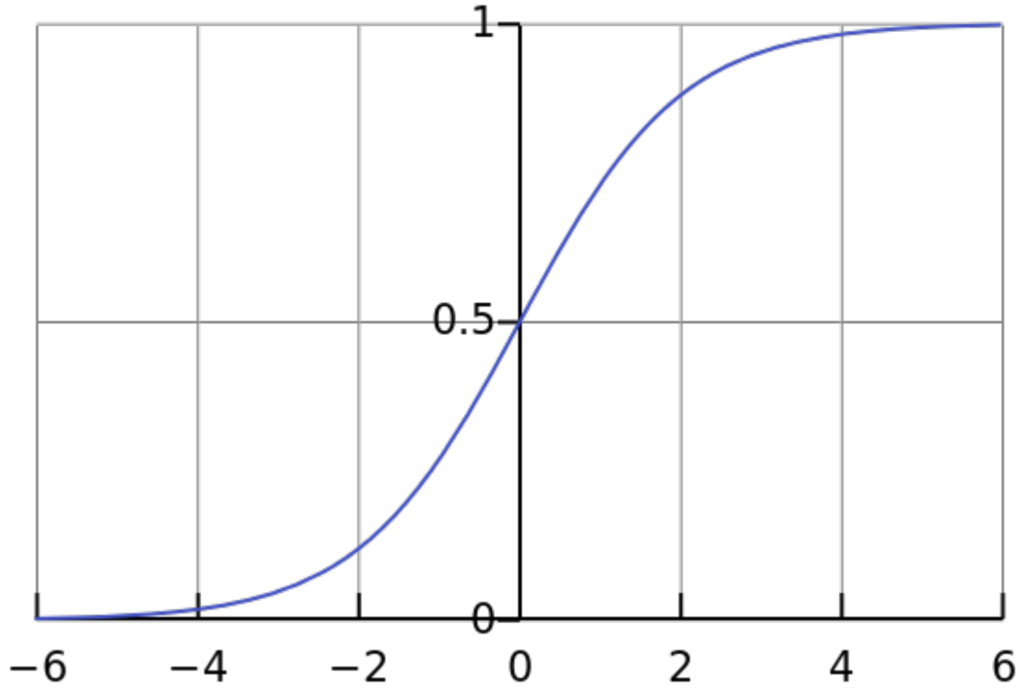
    ✓ $f(\boldsymbol{x})$ should be $0 \le f(\boldsymbol{x}) \le 1$.

How to confine outcome of $\boldsymbol{f}(\boldsymbol{x})$ within [0,1]?

Logistic Regression: Logistic Function

- **Logistic function is the function that can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one**

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



- **In logistic regression, t is determined by explanatory variables**

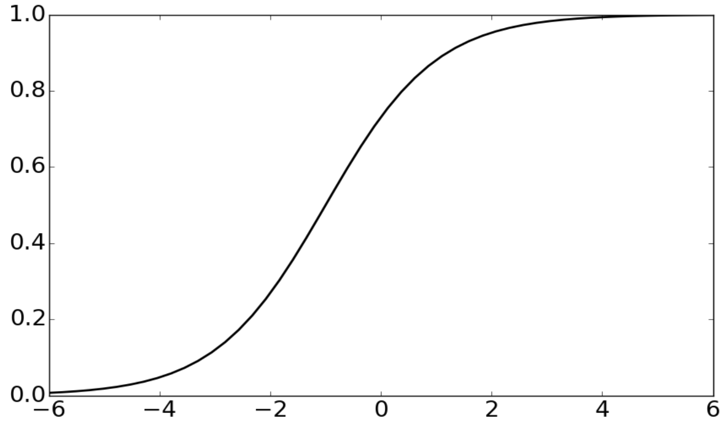
Logistic Regression

- t **is determined by linear combination of explanatory variables**

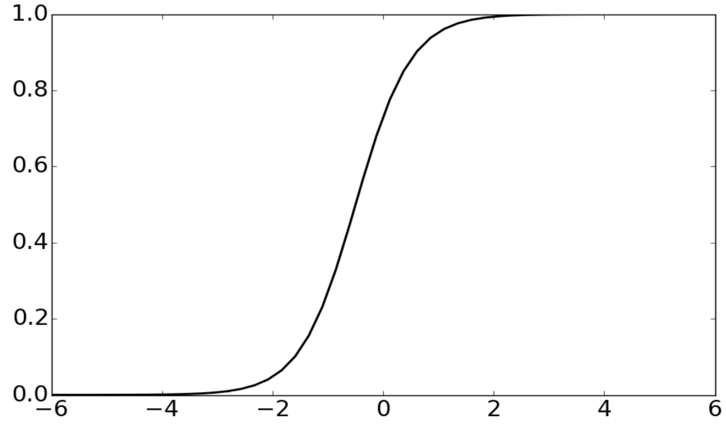
$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

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

$t = 1 + x$

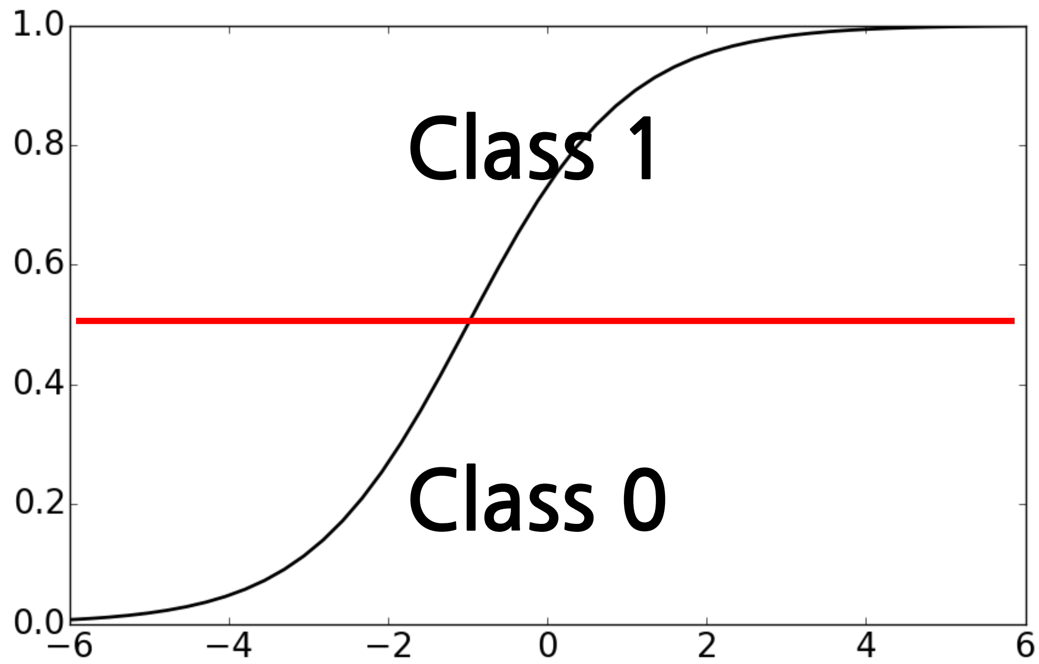


$t = 1 + 2x$



Logistic Regression

- **Determine class**
  - Set class boundary
    - ✓ Without any prior knowledge about class, set the boundary to 0.5



    - ✓ If we have some prior knowledge about the class distribution, then the classification boundary can be determined based on the knowledge.

Logistic Regression

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

- **Unknown parameters**
  - $\beta_0, \beta_1, \ldots \beta_p$

- **Logistic regression should estimate $\beta_0, \beta_1, \ldots, \beta_p$ 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 dataset $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ consists of n independent and identically distributed (iid) samples coming from a distribution with an unknown probability density function $f(\boldsymbol{x})$.
  - Assume $f(\boldsymbol{x})$ belongs to a certain type of distributions with parameters $\boldsymbol{\theta}$.
  - Joint probability density function for all observations

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n | \boldsymbol{\theta}) = f(\boldsymbol{x}_1|\boldsymbol{\theta}) \times f(\boldsymbol{x}_2|\boldsymbol{\theta}) \times \cdots \times f(\boldsymbol{x}_n|\boldsymbol{\theta})$$

because x_i is iid sample

  - Likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n | \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta})$$

An Example of Likelihood Function

- **Imagine the situation that a ball is drawn with replacement from the bag consisting of three blue balls and five white balls.**
  - 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?

- Probability of case 1

$$P(\text{Case1}) = p(\text{blue})\times p(\text{white})\times p(\text{blue})\times p(\text{white})\times p(\text{white})$$

- Probability of case 2

$$P(\text{Case2}) = p(\text{blue})\times p(\text{blue})\times p(\text{blue})\times p(\text{blue})\times p(\text{blue})$$

An Example of Likelihood Function

- **Imagine the situation that a ball is drawn with replacement from the bag consisting of three blue balls and five white balls.**
  - 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

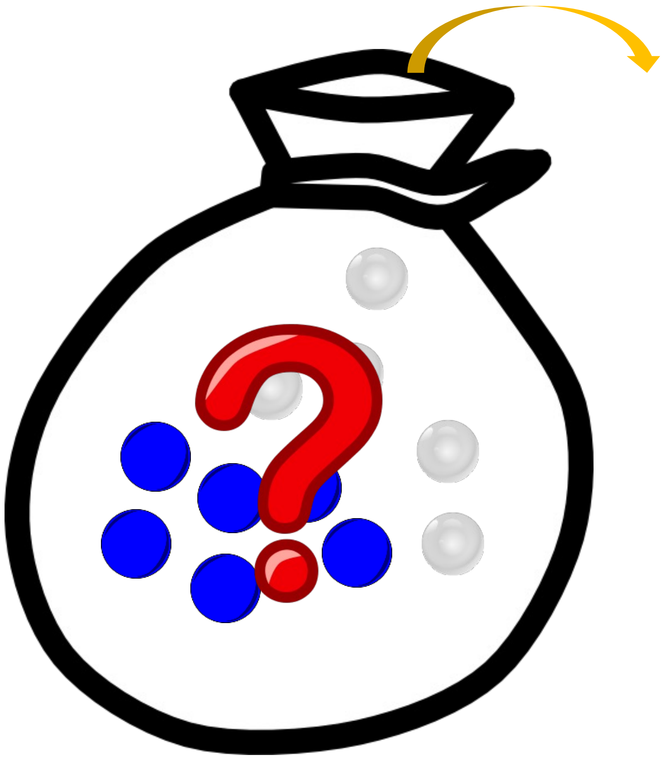
Which case is more probable?

⬇

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

An Example of Likelihood Function

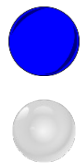
- **Want to estimate p_{blue} and p_{white} based on the sampling result.**



Sampling with replacement

⬇

3×
7×



An 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 p and the value 0 with failure probability $1 - p$
    - ✓ For random variable following Bernoulli distribution,

$$P(X = 1) = 1 - p(X = 0) = p$$

    - ✓ 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

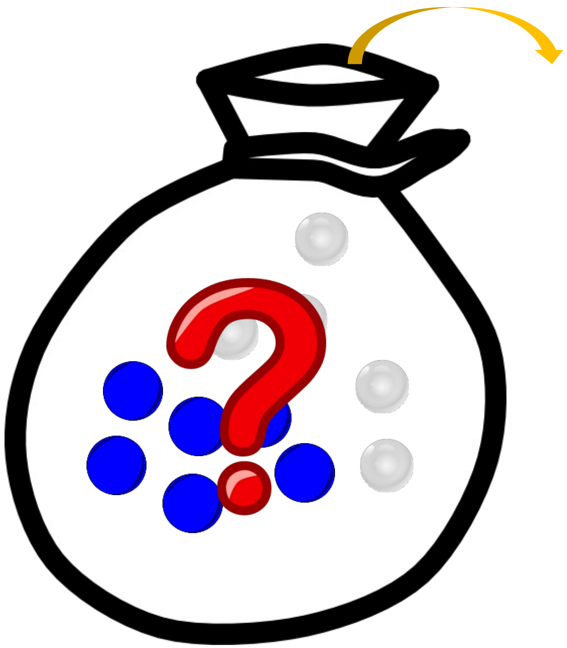
$$f(y;p) = p^y(1-p)^{1-y}, \qquad \text{for } y \in \{0, 1\}$$

  - For Bernoulli distribution, p is θ
    - ✓ In this example, assume that blue ball is 1

$$p = p_{blue}$$
$$1 - p = p_{white}$$

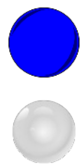
An Example of Likelihood Function

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



Sampling with replacement

3×
7×



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

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

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

An 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 a function of parameter θ**

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

  - 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 a method to find parameters that maximize likelihood function with given data samples

An Example of Likelihood Function

- **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

An Example of Likelihood Function

- **Likelihood function**

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

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

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

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

- **When $\sum_{i=1}^{n}(\theta^2-2x_i\theta+x_i^2)$ is minimum, $\mathcal{L}(\theta; \boldsymbol{x})$ 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 $\rightarrow$ There is a solution to minimize equation

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

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

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

How to Find Parameters for Logistic Regression?

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

- **Each sample has different p depending on the input**

$$y_i \sim Bernoulli(P_i)$$

  - P_i is the probability that output value is 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**

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

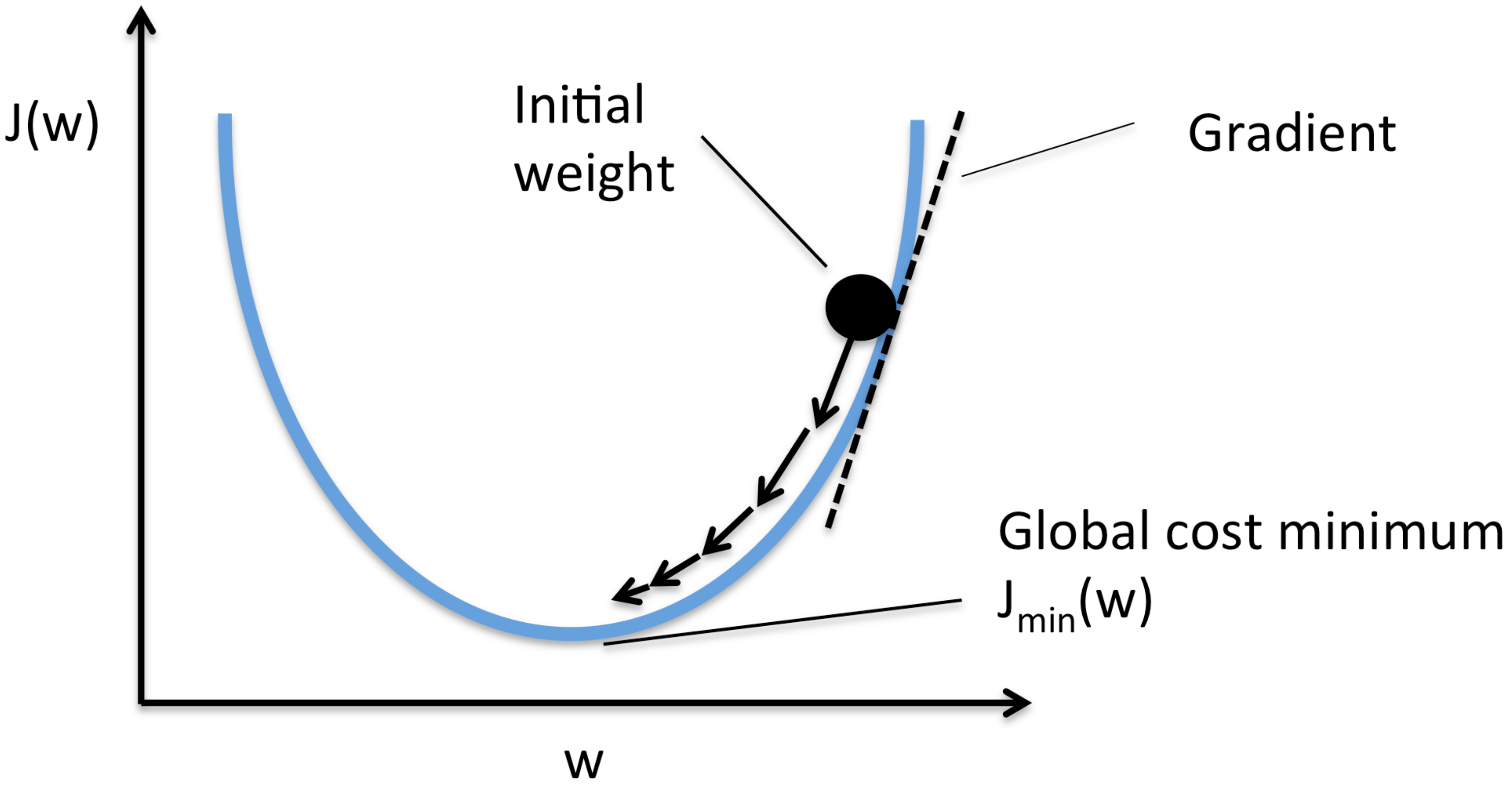
  - $P_i = P(y = 1|\boldsymbol{x}) = \frac{1}{1+e^{-\beta_0-\beta_{!}x_1-\cdots-\beta_p x_p}}$

- **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 $\beta_0, \beta_1, \dots, \beta_p$ that maximize $\log \mathcal{L}$

Gradient Descent



Odds and Odds Ratio

- **Odds reflect the likelihood that the event will occur**
  - 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 $P(y=0)$

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

- **Odds ratio is the ratio between odds when unit increment of a variable**
  - For the first variable x_1

$$Odds\ Ratio = \frac{odds \text{ when input is } x_1 = x+1}{odds \text{ when input is } x_1 = x} = \frac{\exp(\beta_0 + \beta_1(x+1) + \cdots + \beta_p x_p)}{\exp\left(\beta_0 + \beta_1 x + \cdots + \beta_p x_p\right)} = e^{\beta_1}$$

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

Logistic Regression: Odds

- **A logistic regression 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 regression model**

Logit function

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

- 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

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

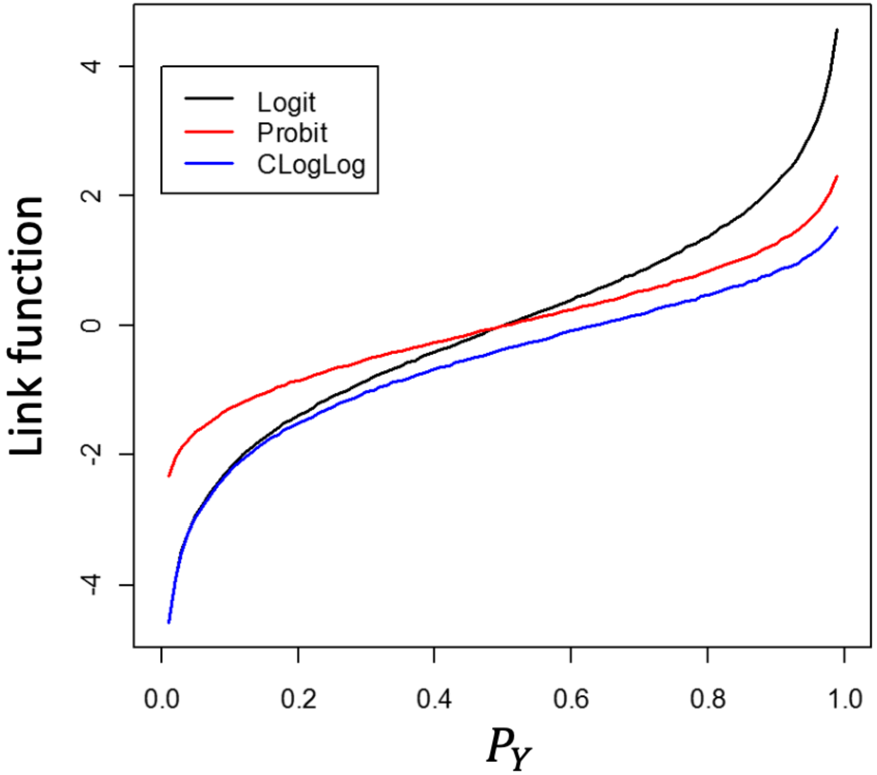
$$g(P) = \Phi^{-1}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

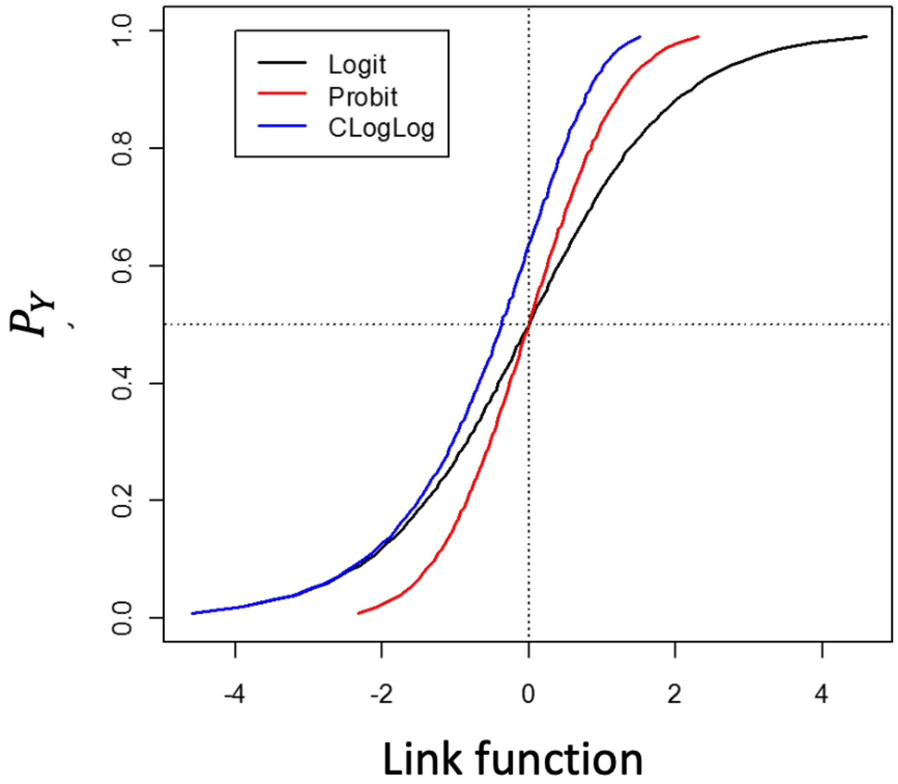
    - ✓ $\Phi^{-1}(x)$ is inverse cumulative density function of normal distribution

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

Properties of Link Functions

- **They can take any value on the real line for $0 \le P(y=1) \le 1$**
  - Consider logit function
    - ✓ If $P(Y=1)=0$, $\text{logit}(P(Y)=1) = \log 0 = -\infty$
    - ✓ If $P(Y=1)=1$, $\text{logit}(P(Y)=1) = \log \infty = \infty$





Logistic Regression for Multi-class

- **For K classes, $P(y_i = k)$ is the probability that i-th data point belong to class 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

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

✓ $\mathbf{x}_i = \left(1, x_{i1}, x_{i2,} \ldots, x_{ip}\right)$

✓ $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})$

✓ $\boldsymbol{\beta}_k \cdot \mathbf{x}_i = \beta_{k0} + \beta_{k1}x_{i1} + \cdots + \beta_{kp}x_{ip}$

⬇

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

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

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
    - ✓ $n_{\mathbf{x}} = \sum_{i=1}^{d} x_i$

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

Likelihood Function

- **Likelihood function**

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

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

- **Log-likelihood function**

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

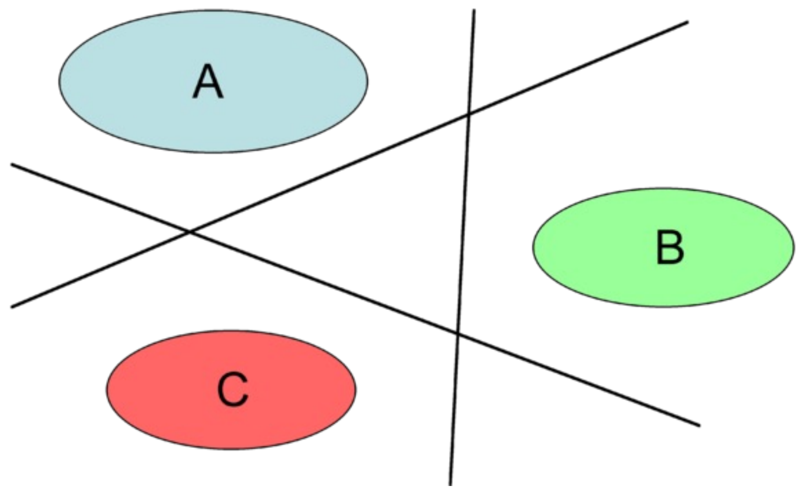
  - Through maximum likelihood estimation, determine $\boldsymbol{\beta}_k$ as the same as in binary logistic regression

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

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

→ One-versus-the rest approach

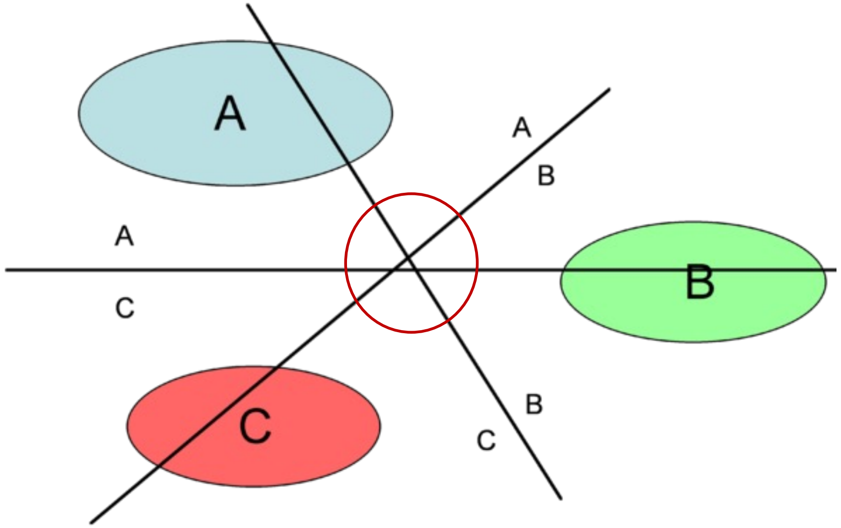


- **Problems of one-versus-the rest approach**
  - Because each classifier was trained on different task, there is no guarantee that the real-values quantities $y_k(\mathbf{x})$ will have appropriate scales
  - Imbalance of data on training

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

Thank you!