L7. Generalized Linear Models

- In a general linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in} + \epsilon_i$$

  - ✓ The response $y_i, i = 1, \ldots, m$, is modeled by a linear function of explanatory variables $x_{ip}, p = 1, \ldots, n$ plus an error term
  - ✓ The model is linear in the parameters
  - ✓ We assume that the errors ϵ_i are independent and identically distributed such that

$$\mathrm{E}[\epsilon_i] = 0, \quad \text{and} \quad \mathrm{var}[\epsilon_i] = \sigma^2$$

    - ➢ Typically we assume $\epsilon_i \sim N(0, \sigma^2)$

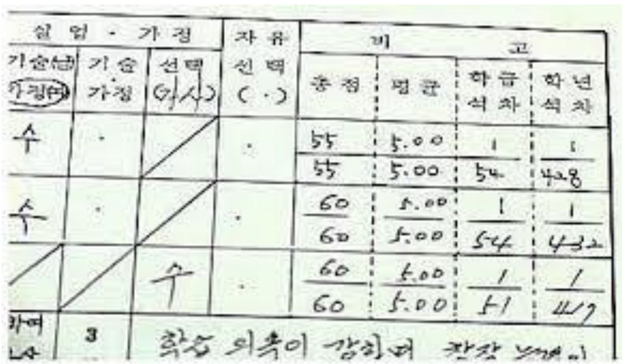
Restrictions of Linear Models

- Although a very useful framework, there are some situations where linear models are not appropriate
  - ✓ The range of Y is restricted (e.g., binary, count)
  - ✓ The variance of Y depends on the mean
- Generalized linear models extend the liner model framework to address both of these issues

Introduction to Logistic regression

University admission committee

High school grades



National Exam score

〈2016학년도 대학수학능력시험 성적통지표(예시)〉

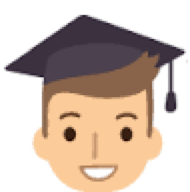
수험번호	성 명	생년월일	성별	출신고교 (반 또는 졸업년도)		
12345678	홍 길 동	97.09.05.	남	한국고등학교 (9)		
구 분	국어 영역	수학 영역	영어 영역	사회탐구 영역		제2외국어 /한문 영역
	B형	A형		생활과 윤리	사회 · 문화	일본어 I
표준점수	131	137	141	53	64	69
백 분 위	93	95	97	75	93	95
등 급	2	2	1	4	2	2

2015. 12. 2.

한국교육과정평가원장

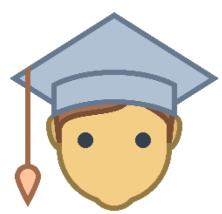
Rejected

Student 1
- Exam: 3/10
- Grades: 4/10



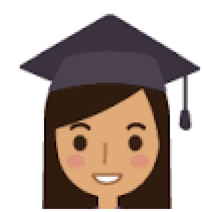
?

Student 2
- Exam: 7/10
- Grades: 6/10

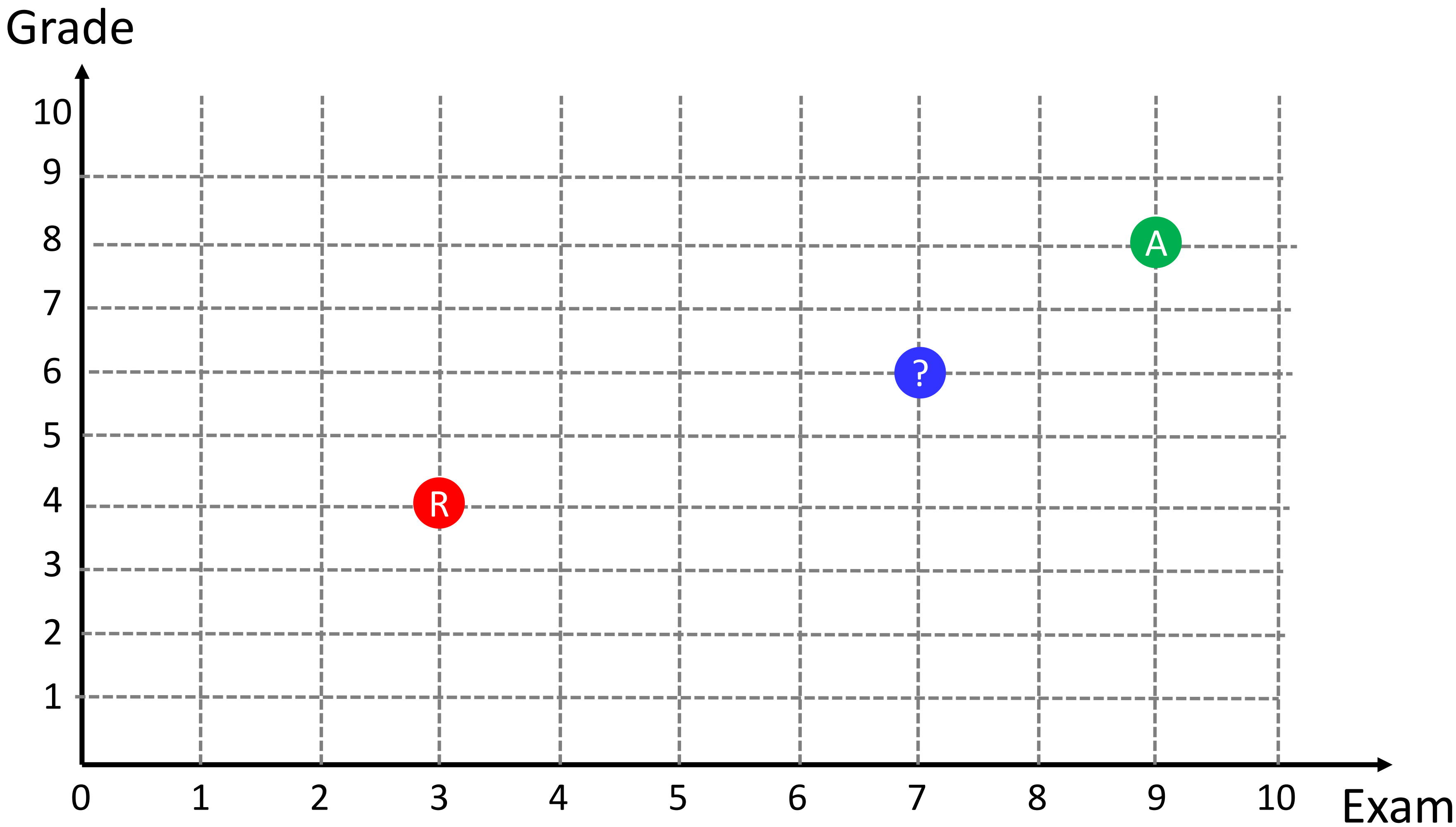


Accepted

Student 3
- Exam: 9/10
- Grades: 8/10

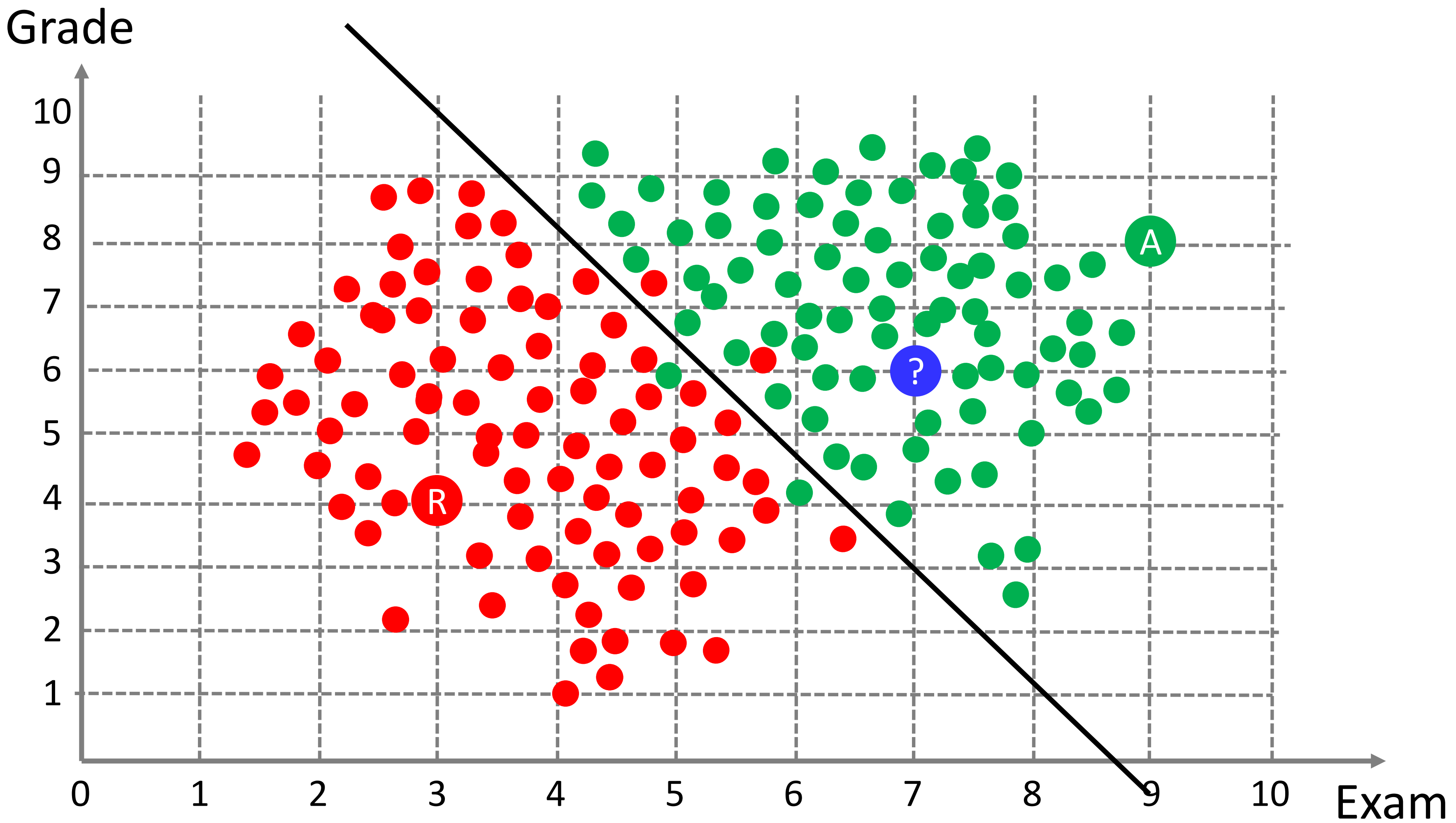


University admission committee



University admission committee

Look at the **historical data** on the admission results



Logistic regression

- Logistic regression is *discriminative* probabilistic linear classification :$p(y|x) = g(w^T x)$

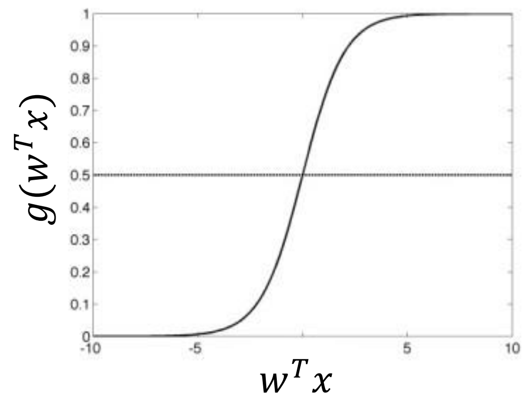
Let's denote p a probability of having $y = 1$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = w^T x$$

$$\frac{p}{1-p} = \exp(w^T x\,)$$

$$p = \frac{\exp(w^T x\,)}{1+\exp(w^T x\,)} = \frac{1}{1+\exp(-w^T x\,)} = g(w^T x)$$

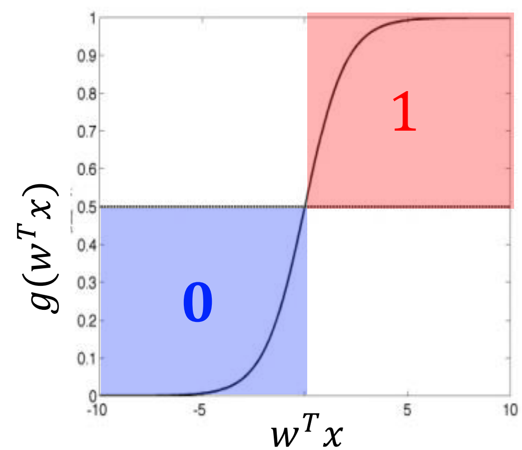
- Larger $w^T x$ →lareger→ $g(w^T x)$ → higher p for $y = 1$
- Smaller $w^T x$ →smaller→ $g(w^T x)$ → lower p for $y = 1$



$$g(z) = \frac{1}{(1+\exp(-w^T x))}$$

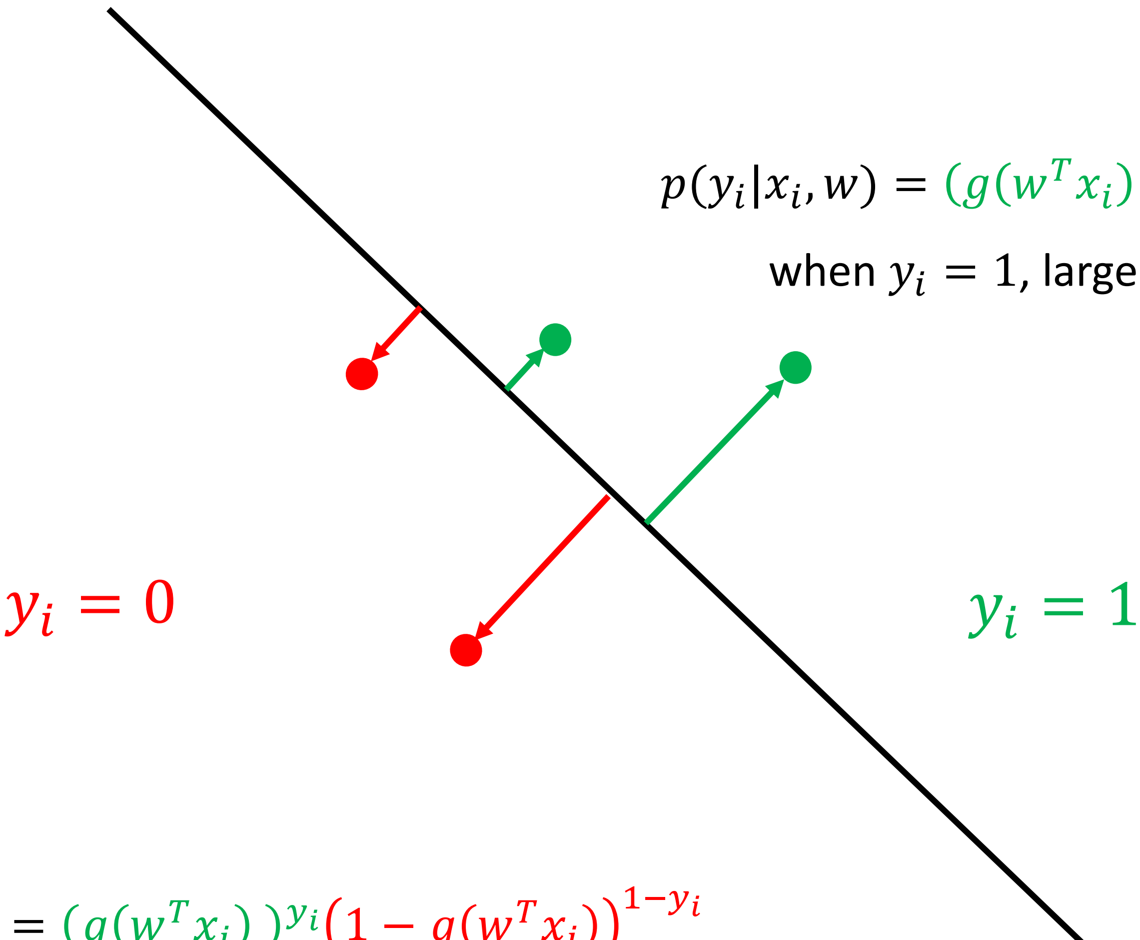
- Classification rule:

$$y = \begin{cases} 0, & \text{if } p(Y=1|x) = g(w^T x) < 0.5 \Leftrightarrow w^T x < 0 \\ 1, & \text{if } p(Y=1|x) = g(w^T x) \geq 0.5 \Leftrightarrow w^T x \geq 0 \end{cases}$$



University admission committee

How to draw **a separating line** ?



$$p(y_i|x_i,w) = (g(w^T x_i)\,)^{y_i}\left(1 - g(w^T x_i)\right)^{1-y_i}$$

when $y_i = 1$, larger $g(w^T x_i)$ is better

$$p(y_i|x_i,w) = (g(w^T x_i)\,)^{y_i}\left(1 - g(w^T x_i)\right)^{1-y_i}$$

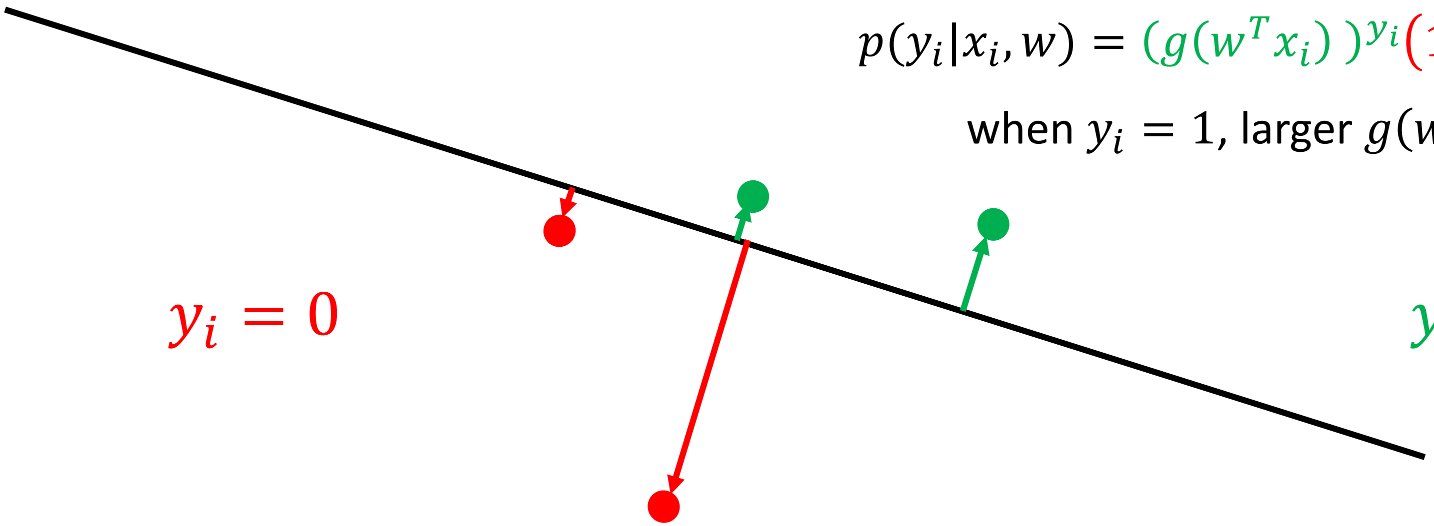
When $y_i = 0$, smaller $g(w^T x_i)$ is better

University admission committee

How to draw **a separating line** ?

$$p(y_i|x_i, w) = (g(w^T x_i))^{y_i} (1 - g(w^T x_i))^{1-y_i}$$

when $y_i = 1$, larger $g(w^T x_i)$ is better



$$p(y_i|x_i, w) = (g(w^T x_i))^{y_i} (1 - g(w^T x_i))^{1-y_i}$$

When $y_i = 0$, smaller $g(w^T x_i)$ is better

Logistic regression – objective function

- Likelihood for **a single point** (x_i, y_i) can be specified as

$$p(y_i|x_i, w) = (g(w^T x_i)\,)^{y_i} \left(1 - g(w^T x_i)\right)^{1-y_i}$$

- Likelihood for **whole training data** (X, y) can be specified as

$$p(y|X, w) = \prod_{i}^{m} p(y_i|x_i, w) = \prod_{i=1}^{m} (g(w^T x_i)\,)^{y_i} \left(1 - g(w^T x_i)\right)^{1-y_i}$$

Note that this is similar to the likelihood of Binomial dist.

- **Log**-likelihood

$$L(w) = \log \prod_{i}^{m} p(y_i|x_i, w) = \sum_{i=1}^{m} y_i \log g(w^T x_i) + (1 - y_i) \log\left(1 - g(w^T x_i)\right)$$

Logistic regression – learning (optimization)

- **Log**-likelihood

$$L(w) = \log \prod_{i}^{m} p(y_i | x_i, w) = \sum_{i=1}^{m} y_i \log g(w^T x_i) + (1 - y_i) \log\left(1 - g(w^T x_i)\right)$$

- We can find the parameters that maximizes the log-likelihood function

$$w^* = \mathrm{argmax}_w \ L(w)$$

- **Gradient ascent** algorithm

Repeat until convergence{

$w_j := w_j + \alpha \frac{\partial}{\partial w_j} L(w)$ (for every j)

}

α : learning rate

$$\frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{m} \left(y_i - g(w^T x_i)\right) x_{ij}$$

Logistic regression – learning (optimization)

- **Log**-likelihood

$$L(w) = \log \prod_{i}^{m} p(y_i|x_i, w) = \sum_{i=1}^{m} y_i \log g(w^T x_i) + (1 - y_i) \log\big(1 - g(w^T x_i)\big)$$

- We can find the parameters that maximizes the log-likelihood function

$$w^* = \text{argmax}_w \ L(w)$$

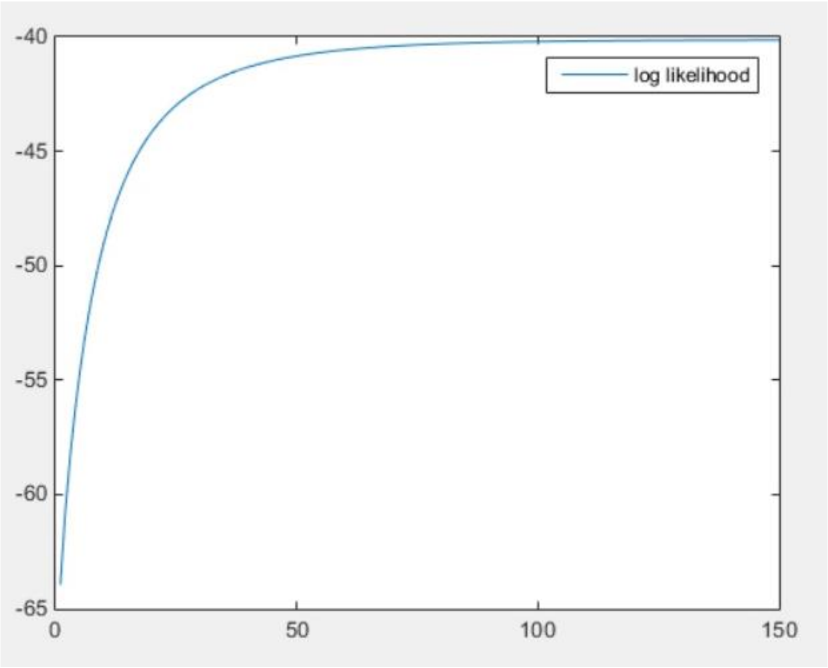
- **Stochastic gradient ascent** algorithm

Repeat until convergence{
for $i = 1, \dots, m$ {
$w_j := w_j + \alpha\big(y_i - g(w^T x_i)\big)x_{ij}$ (for every j)
}
}

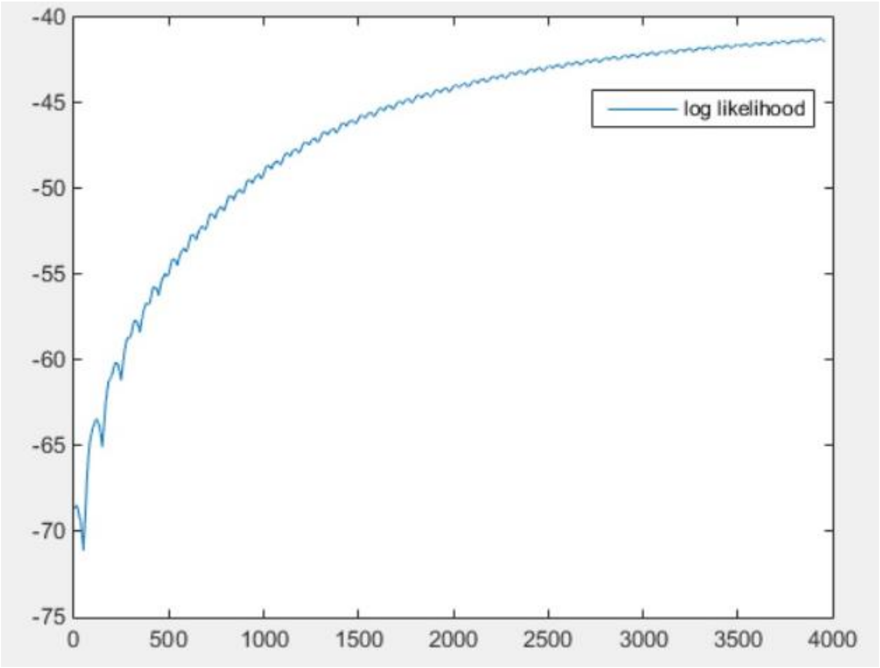
α : learning rate

$$\frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{m} \big(y_i - g(w^T x_i)\big)x_{ij} \sim \big(y_i - g(w^T x_i)\big)x_{ij}$$

Logistic regression – learning (optimization)

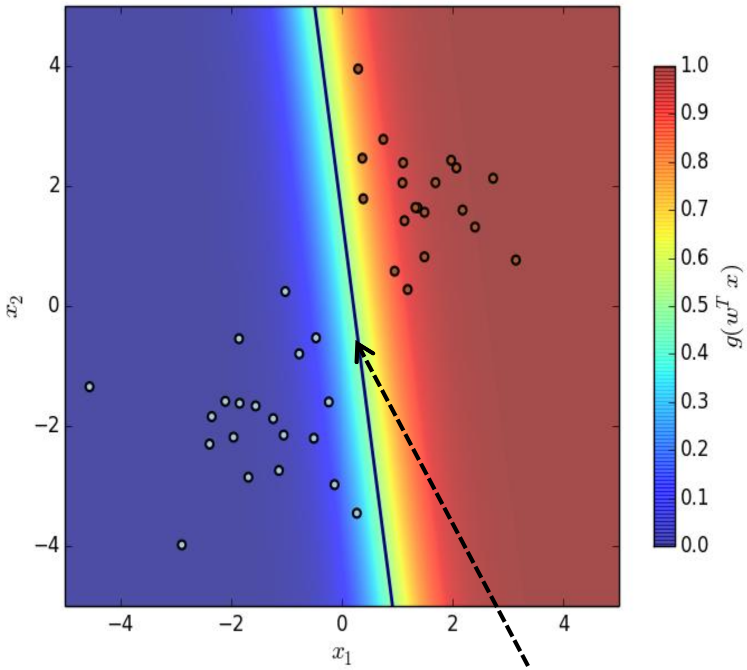


Gradient ascent의 log-likelihood 수렴

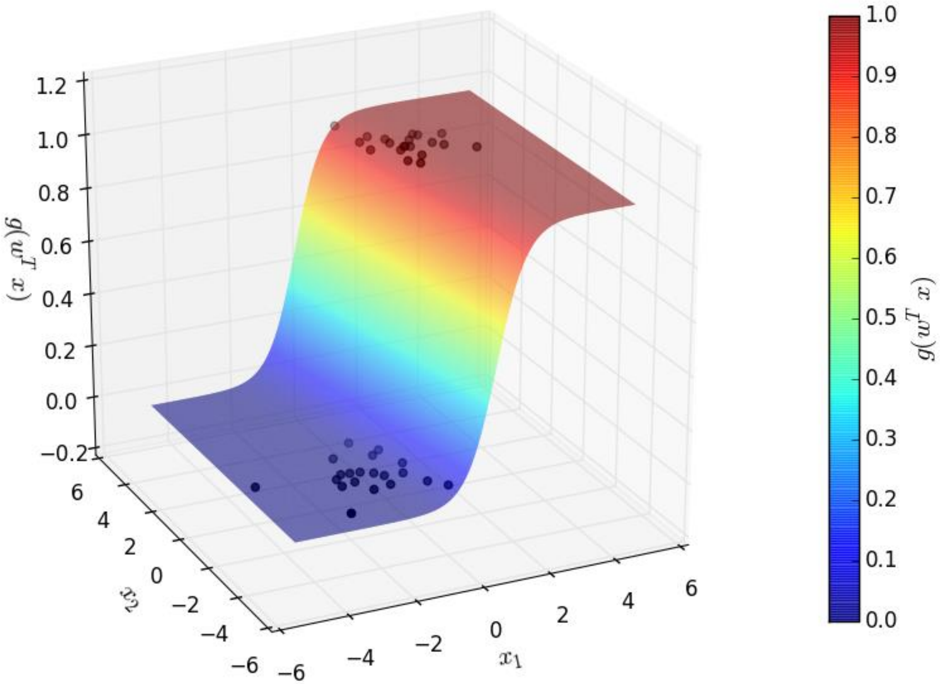


Stochastic gradient ascent의 log-likelihood 수렴

Logistic regression

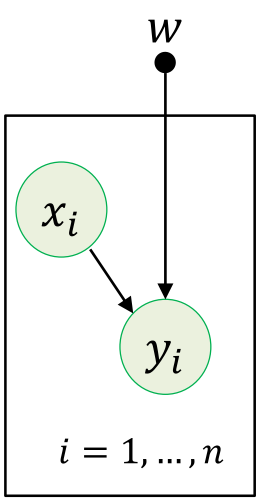


Classification line $\boldsymbol{w}^T \boldsymbol{x} = \mathbf{0}$



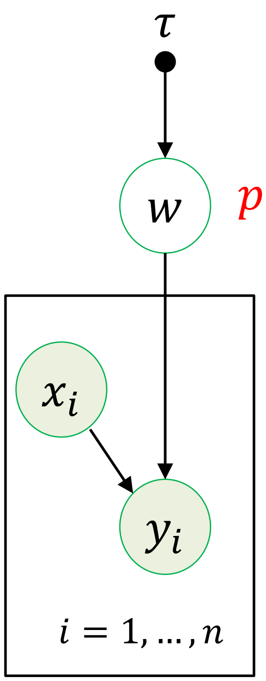
Logistic Regression

Fixed parameter
(to be determined)



Bayesian Logistic Regression

Fixed hyper-parameter



$p(w) = N(w|\mathbf{0}, \tau^2\mathbf{I})$

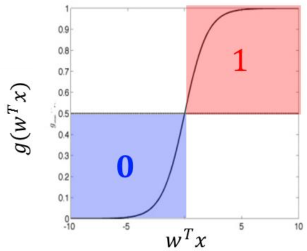
$$y_i = \begin{cases} 0, \text{if } g(w^T x_i) < 0.5 \Leftrightarrow w^T x_i < 0 \\ 1 \ \text{ if } g(w^T x_i) \geq 0.5 \Leftrightarrow w^T x_i \geq 0 \end{cases}$$

Bayesian Logistic Regression with Gaussian Prior (Ridge Logistic Regression)

- We have a logistic regression model :

$$p(Y=1|x) = g(w^T x) = \frac{1}{(1+\exp(-w^T x))}$$

$$p(Y=0|x) = 1 - g(w^T x)$$



- **Likelihood** can be specified as

$$p(y_i|x_i, w) = (g(w^T x_i))^{y_i}\big(1 - g(w^T x_i)\big)^{1-y_i}$$

for $y = (y_1, \dots, y_m)$

$$p(y|X, w) = \prod_{i}^{m} p(y_i|x_i, w) = \prod_{i=1}^{m} (g(w^T x_i))^{y_i}\big(1 - g(w^T x_i)\big)^{1-y_i}$$

- **Prior** on parameter w can be specified as

$$p(w_j) = N(w_j|0, \tau_i^2) = \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{w_j^2}{2\tau_j^2}\right)$$

for $w = (w_1, \dots, w_n)$

$$p(w) = \prod_{i=1}^{n} N(w_j|0, \tau_i^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{w_j^2}{2\tau_j^2}\right)$$

✓ τ_j^2 quantifies our belief that w_j is close to 0.

✓ For simple case, $\tau_j^2 = \tau^2$ for $j = 1, \dots, n$

Bayesian Logistic Regression with Gaussian Prior (Ridge Logistic Regression)

- We need to compute **the posterior**: (For simple case, $\tau_j^2 = \tau^2$ for $j = 1, \dots, n$)

$$p(w|X,y) \propto p(y|X,w)p(w)$$
$$= \prod_{i=1}^{m} (g(w^T x_i))^{y_i} \left(1 - g(w^T x_i)\right)^{1-y_i} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$$

$$\log p(w|X,y) \propto \sum_{i=1}^{m} y_i \log g(w^T x_i) + (1 - y_i) \log\left(1 - g(w^T x_i)\right) + n\log\left(\frac{1}{\sqrt{2\pi\tau^2}}\right) - \sum_{j=1}^{n} \frac{w_j^2}{2\tau^2}$$

- The **MAP** estimate of w is then simply

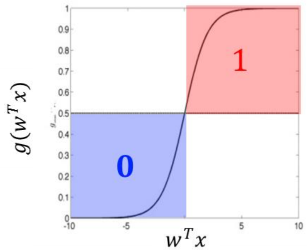
$$\hat{w} = \underset{w}{\operatorname{argmax}}\, p(w|X,y)$$
$$= \underset{w}{\operatorname{argmax}} \log p(w|X,y)$$
$$= \underset{w}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{m} y_i \log g(w^T x_i) + (1 - y_i) \log\left(1 - g(w^T x_i)\right)}_{\text{Data fitness}} \underbrace{- \lambda \|w\|_2^2}_{\text{complexity}}$$

Bayesian Logistic Regression with Laplace Prior (Lasso Logistic Regression)

- We have a logistic regression model :

$$p(Y=1|x) = g(w^T x) = \frac{1}{(1+\exp(-w^T x))}$$

$$p(Y=0|x) = 1 - g(w^T x)$$



- **Likelihood** can be specified as

$$p(y_i|x_i, w) = (g(w^T x_i))^{y_i} \left(1 - g(w^T x_i)\right)^{1-y_i}$$

for $y = (y_1, \ldots, y_m)$

$$p(y|X, w) = \prod_{i}^{m} p(y_i|x_i, w) = \prod_{i=1}^{m} (g(w^T x_i))^{y_i} \left(1 - g(w^T x_i)\right)^{1-y_i}$$

- **Prior** on parameter w can be specified using Laplacian as

$$p(w_j) = \frac{\lambda_j}{2} \exp(-\lambda_j |w_j|)$$

for $w = (w_1, \ldots, w_n)$

$$p(w) = \prod_{j=1}^{n} \frac{\lambda_j}{2} \exp(-\lambda_j |w_j|)$$

✓ τ_j^2 quantifies our belief that w_j is close to 0.
✓ For simple case, $\tau_j^2 = \tau^2$ for $j = 1, \ldots, n$

Bayesian Logistic Regression with Laplace Prior (Lasso Logistic Regression)

- We need to compute **the posterior**: (For simple case, $\tau_j^2 = \tau^2$ for $j = 1, \dots, n$)

$$p(w|X,y) = p(y|X,w)p(w)$$
$$= \prod_{i=1}^{m} (g(w^T x_i))^{y_i} (1 - g(w^T x_i))^{1-y_i} \prod_{j=1}^{n} \frac{\lambda}{2} \exp(-\lambda |w_j|)$$

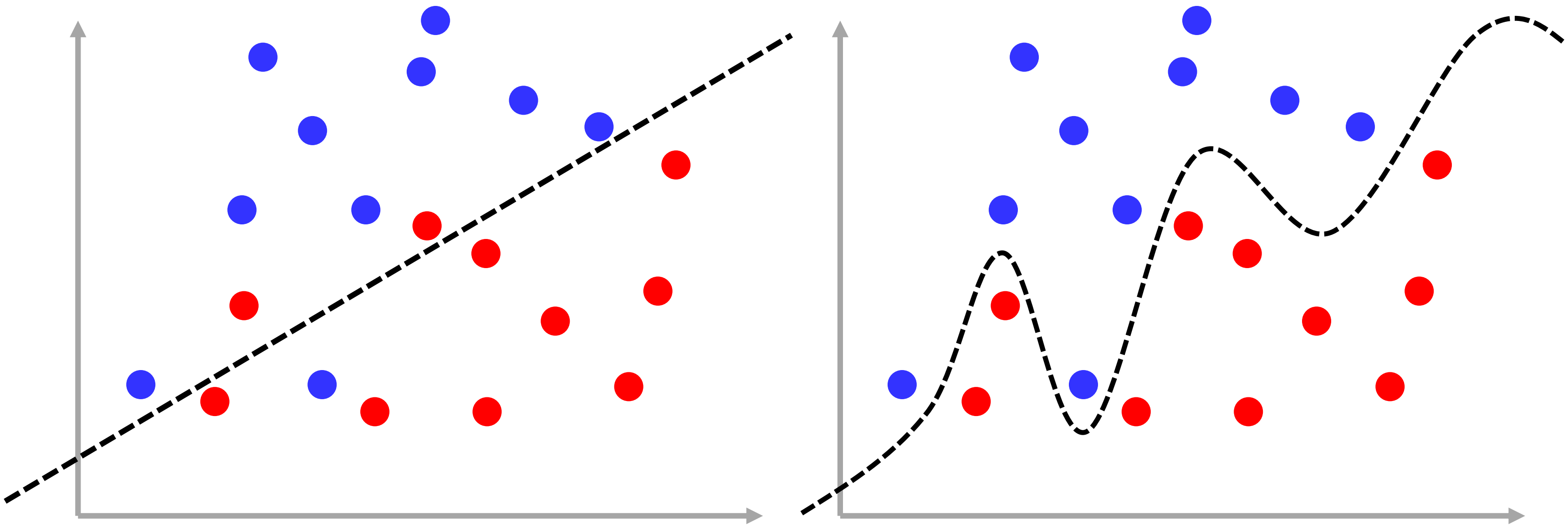
$$\log p(w|X,y) = \sum_{i=1}^{m} y_i \log g(w^T x_i) + (1-y_i) \log(1 - g(w^T x_i)) + n \log\left(\frac{\lambda}{2}\right) - \lambda \sum_{j=1}^{n} |w_j|$$

- The **MAP** estimate of w is then simply

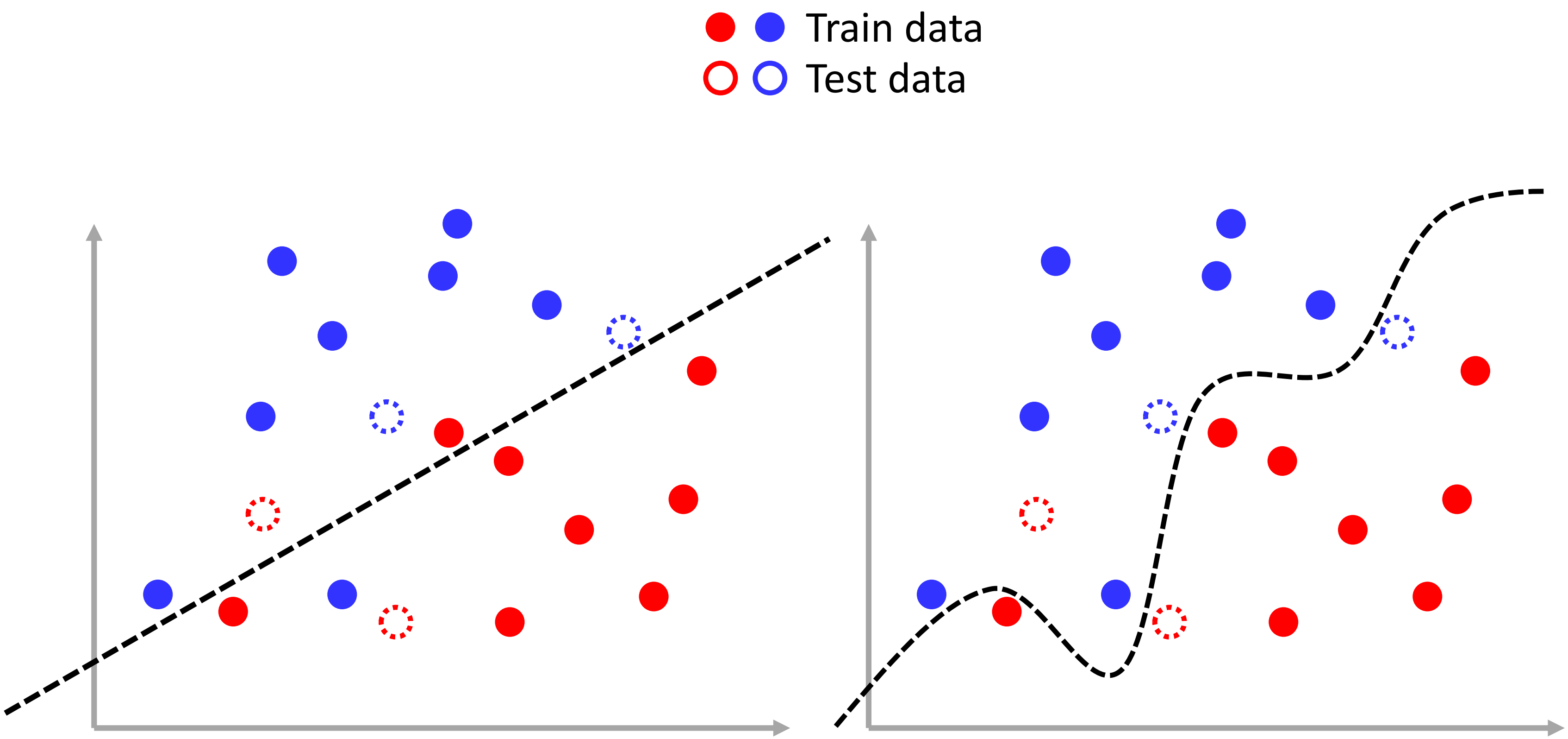
$$\hat{w} = \underset{w}{\operatorname{argmax}}\, p(w|X,y)$$
$$= \underset{w}{\operatorname{argmax}} \log p(w|X,y)$$
$$= \underset{w}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{m} y_i \log g(w^T x_i) + (1-y_i) \log(1 - g(w^T x_i))}_{\text{Data fitness}} \underbrace{- \lambda \sum_{j=1}^{n} |w_j|}_{\text{Complexity (sparsity)}}$$

Model Selection and Evaluation

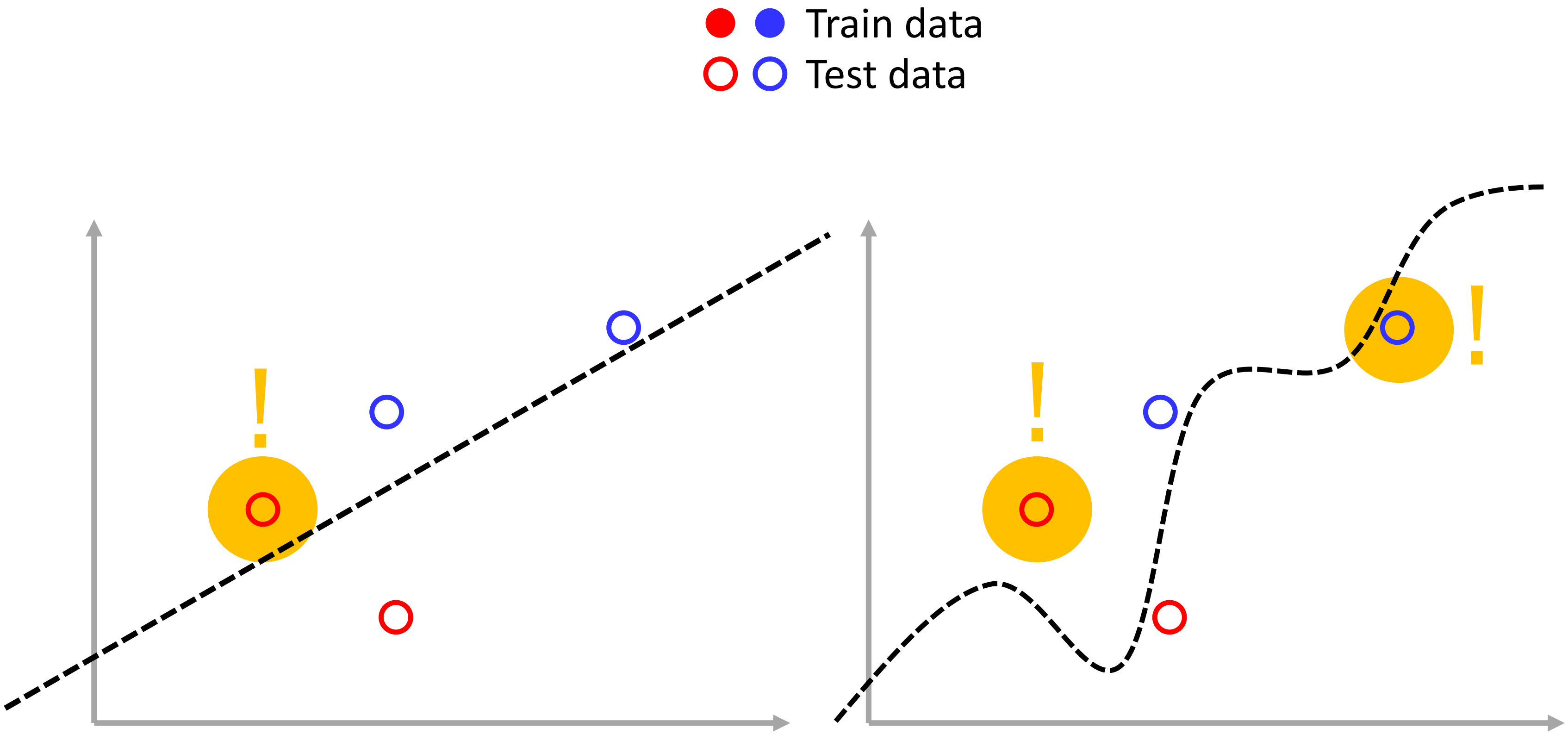
Which model is better



Which model is better



Which model is better

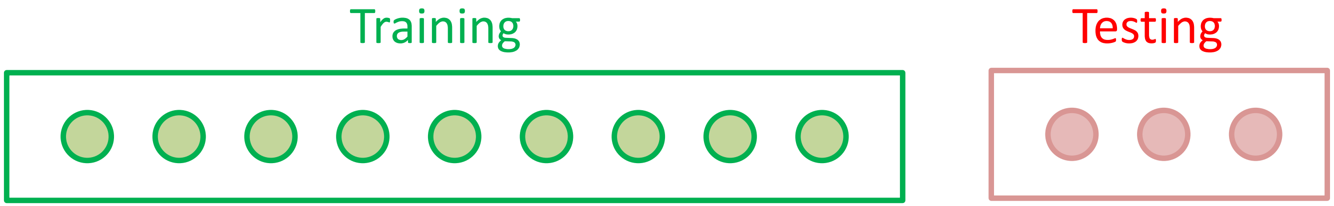


$$\text{Accuracy} = \frac{3}{4} \qquad\qquad \text{Accuracy} = \frac{2}{4}$$

Golden rule for machine learning:

Never use test data to train your model!

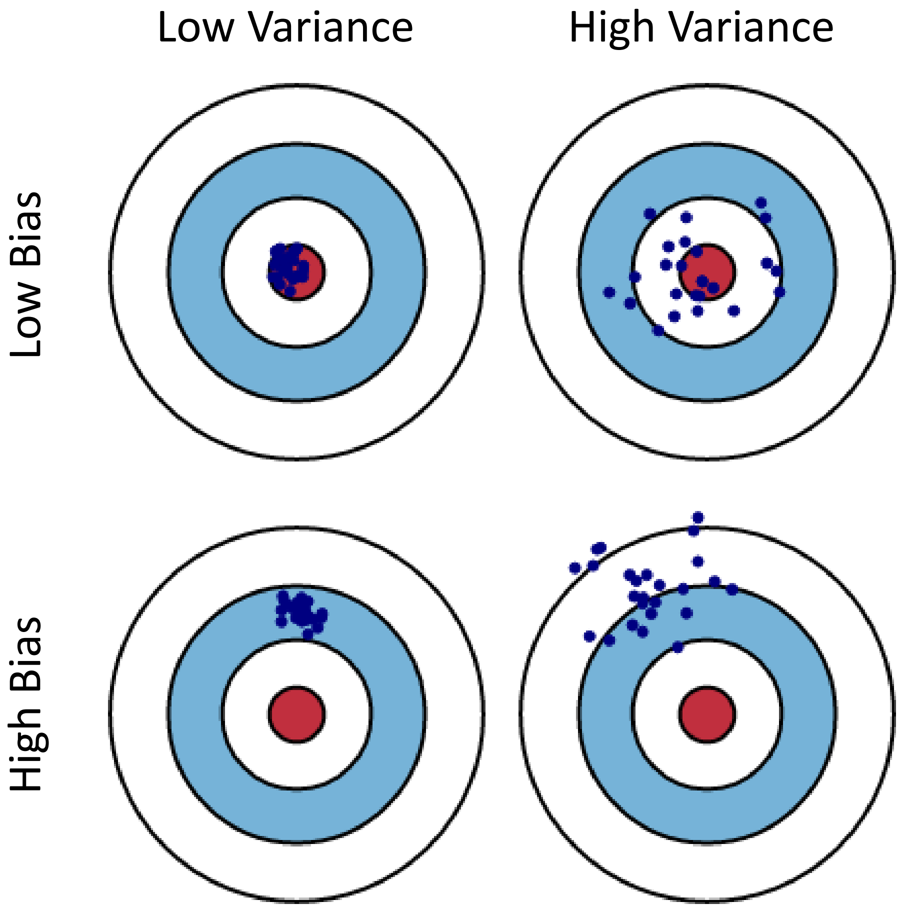
How do we not 'lose' the training data?



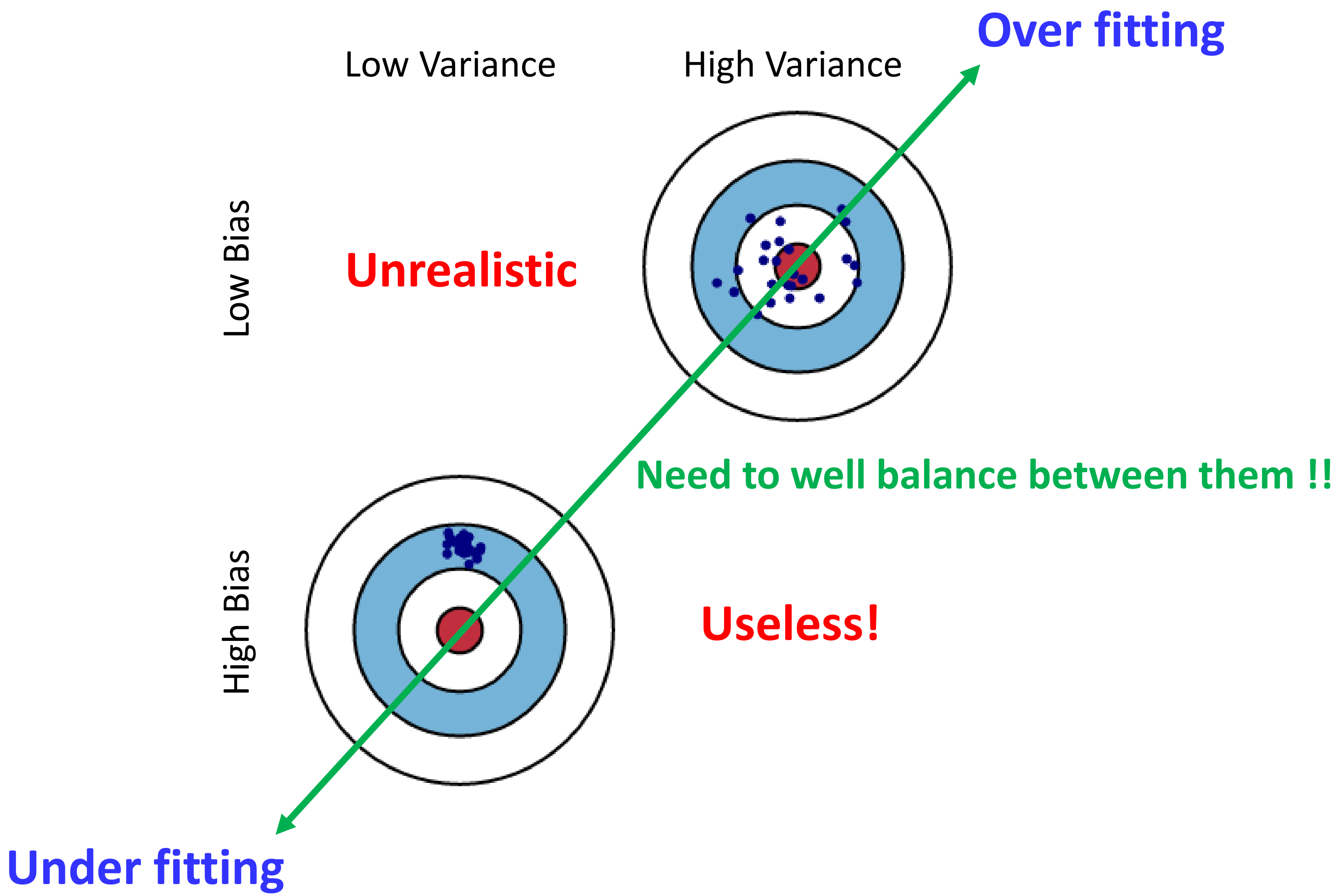
How do we not 'lose' the training data?

K-Fold Cross Validation

Under fitting and Over fitting : The Bias and Variance Trade off

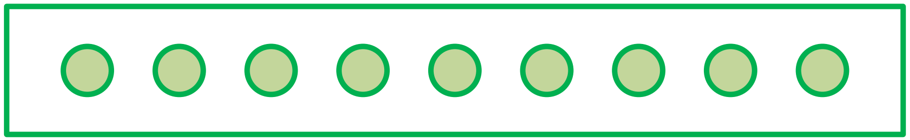


Under fitting and Over fitting : The Bias and Variance Trade off



Model selection and training

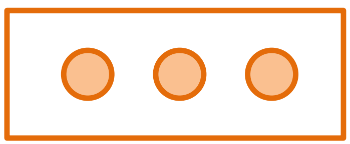
Training set



Training the model

- Fit the model parameters

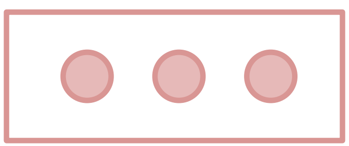
Validation set



Make decision about the model

- Select hyper parameters
  - Degree
  - Features,
  - Structures…

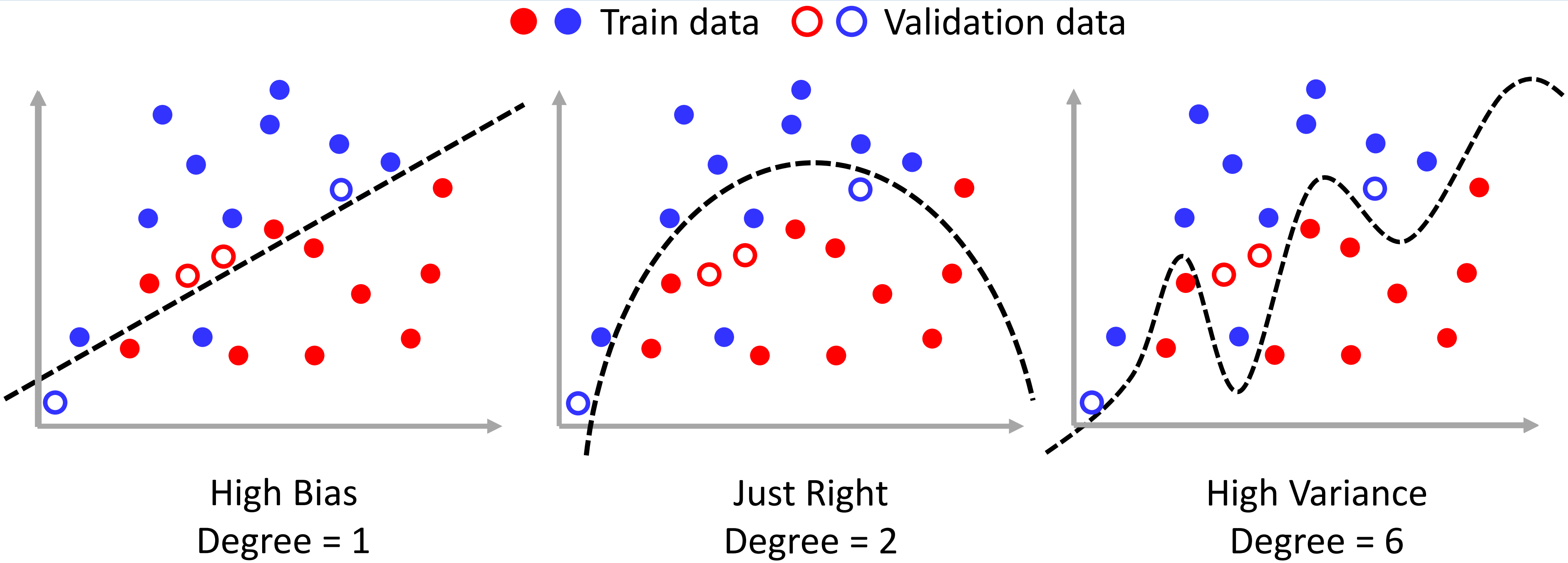
Test set



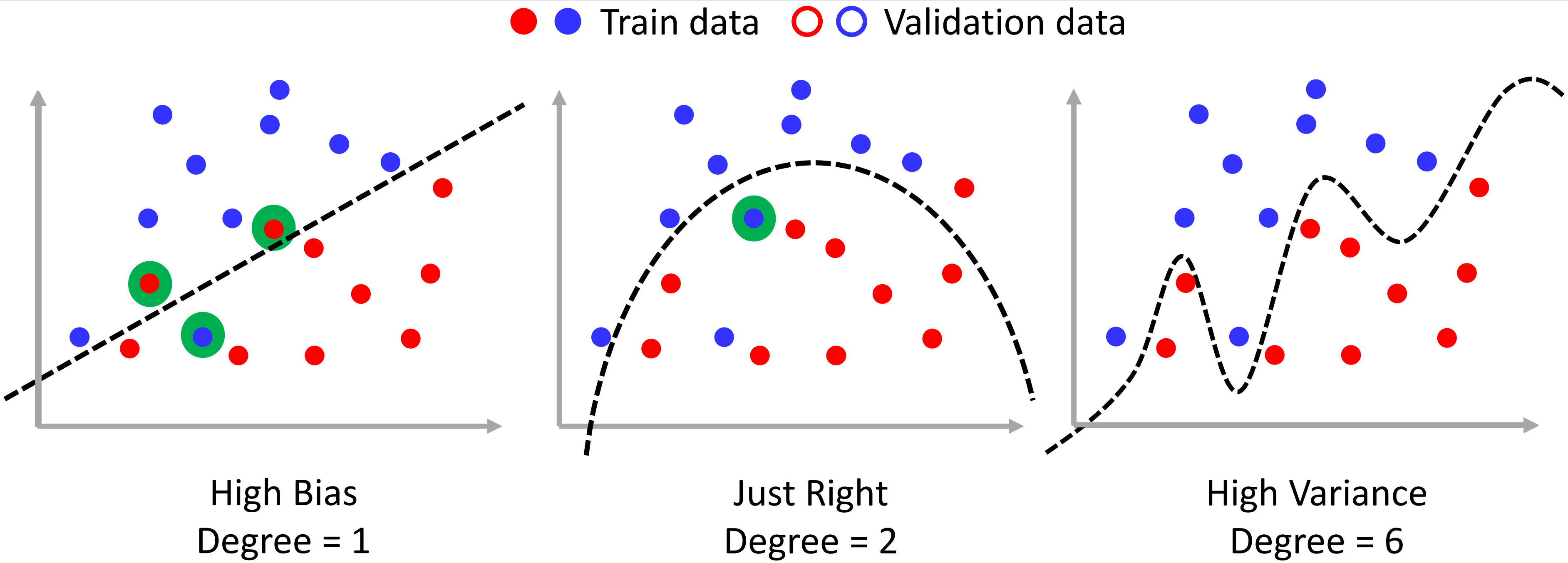
Final testing

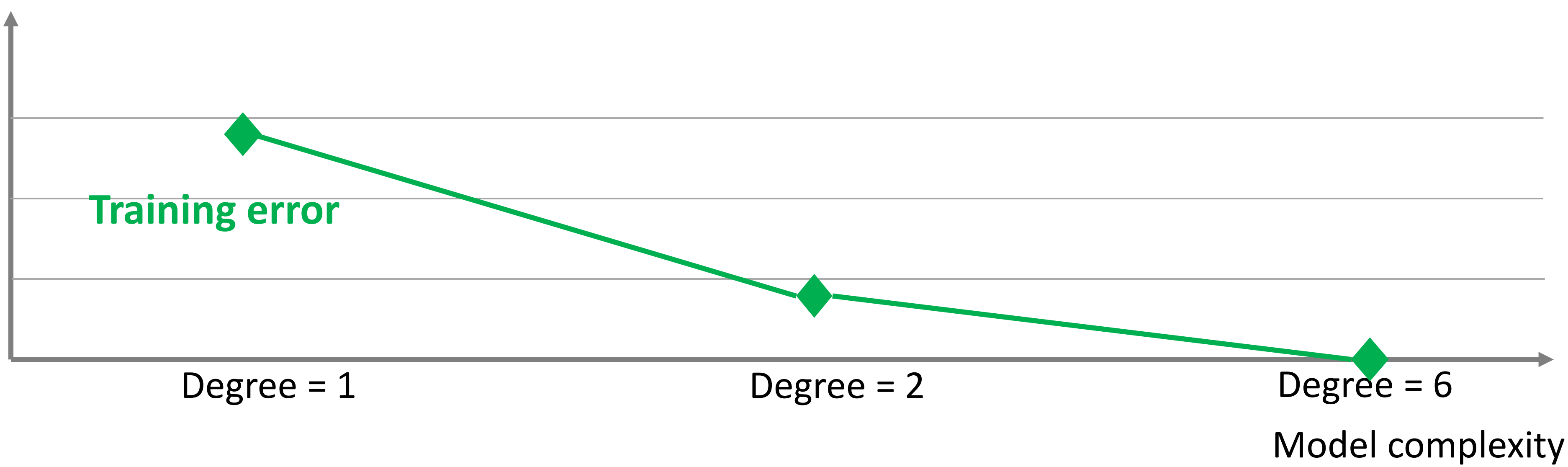
- Never make decision based on test set
- its just for evaluation!

Model Complexity Graph

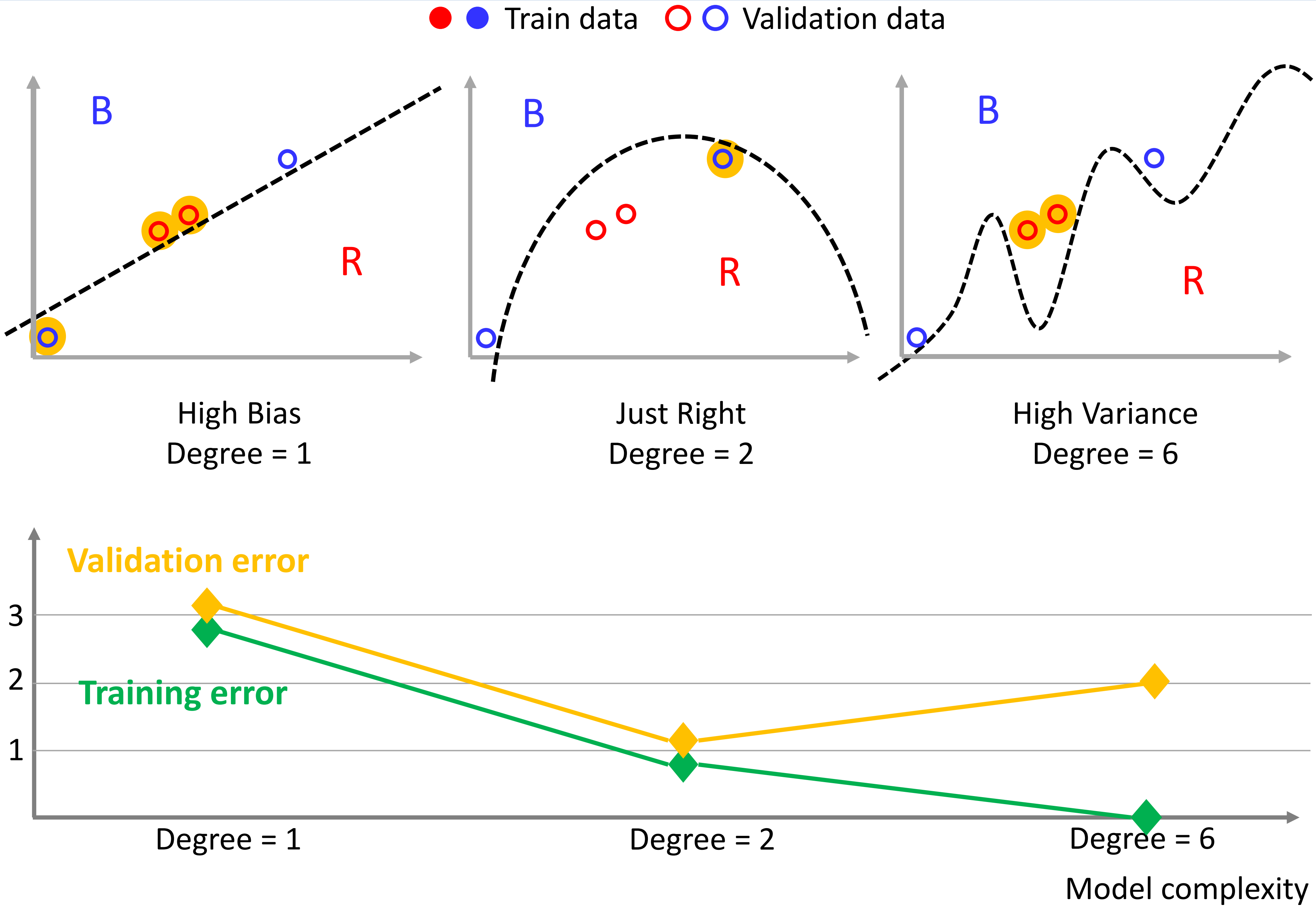


Model Complexity Graph

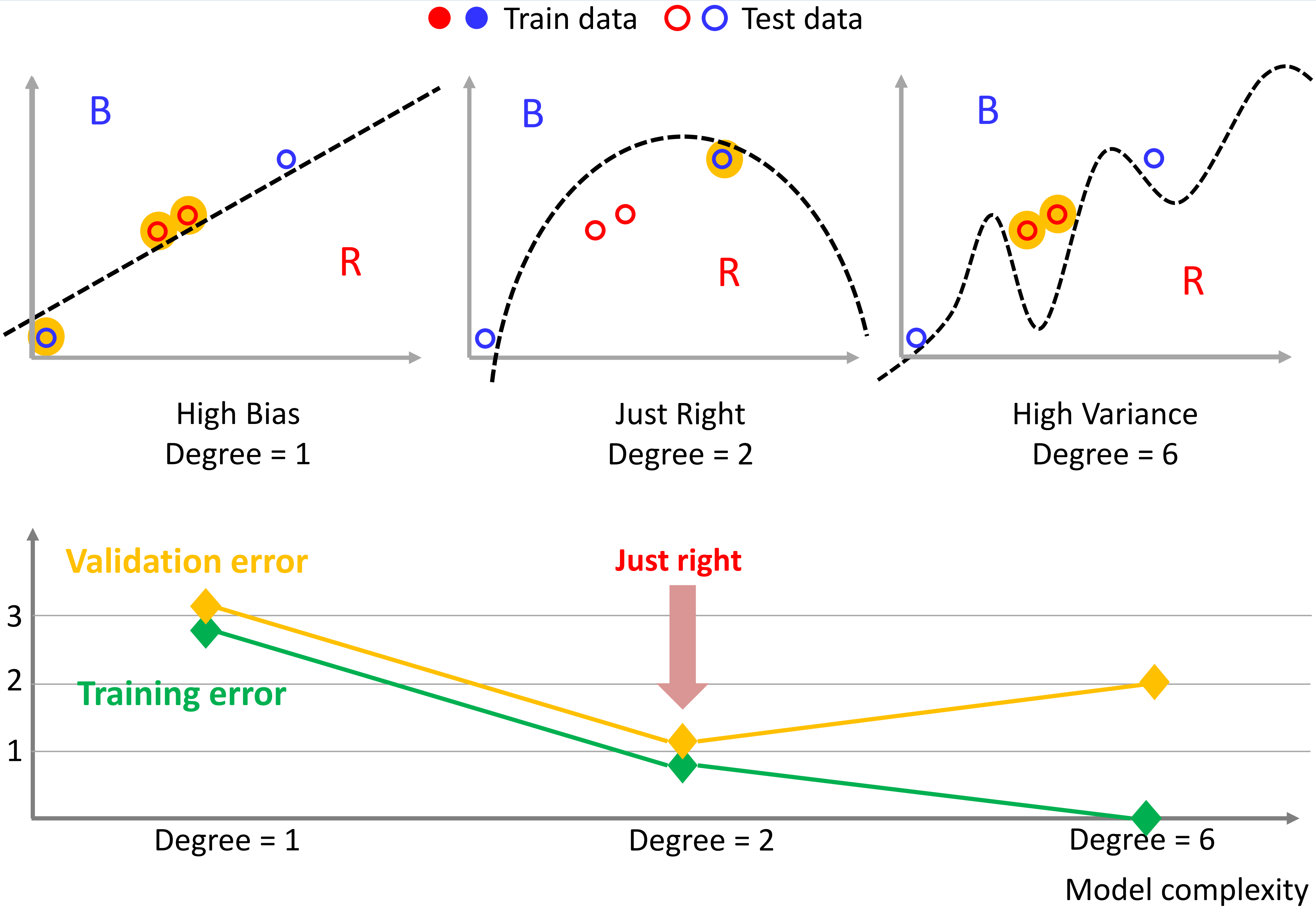




Model Complexity Graph

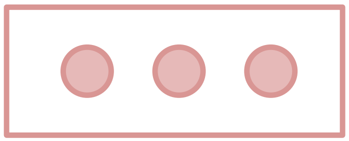


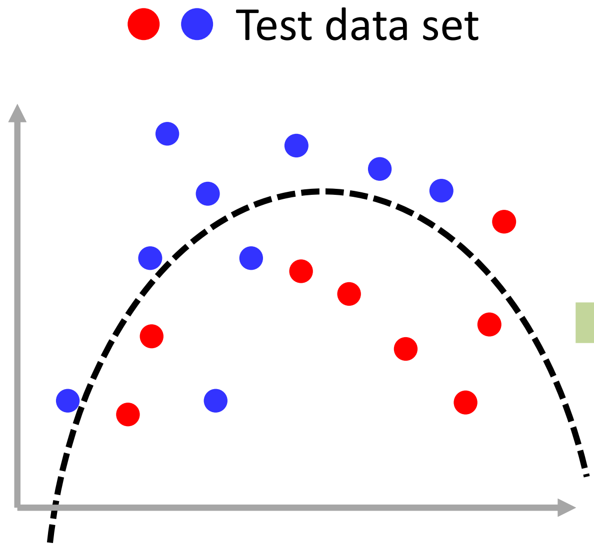
Model Complexity Graph



Test error

Test set





Just Right
Degree = 2

Evaluate performance:

- accuracy,
- Precision
- Recall
- etc.

Generalized Linear Model

- A generalized linear model is made up of
  - a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in} + \epsilon_i$$

  - ***A link function*** that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) \text{ or}$$
$$g\big(E(Y_i)\big) = g(\mu_i) = \eta_i$$

  - ***A variance function*** that describes how the variance, $\mathrm{var}(Y_i)$ depends on the mean

$$\mathrm{var}(Y_i) = \phi V\big(E(Y_i)\big) = \phi V(\mu_i)$$

Linear Regression as Generalized Linear Models (GLMs)

- For the general linear model with $Y_i \sim N(\mu_i, \sigma^2)$
  - a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

  - ***the link function***

$$g\big(E(Y_i)\big) = g(\mu_i) = \eta_i$$
$$g(\mu_i) = \mu_i$$
$$\Rightarrow \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

  - ***A variance function***

$$\operatorname{var}(Y_i) = \phi V\big(E(Y_i)\big) = \phi V(\mu_i)$$
$$V(\mu_i) = 1$$
$$\Rightarrow \operatorname{var}(Y_i) = \phi \times 1 = \sigma^2$$

Revisit to Logistic Regression: Motivation

- In many situations, we would like to forecast the ***outcome of a binary event***, given some relevant information:
  - Given the pattern of word usage in an e-mail, is it likely to be spam?
  - Given the temperature and cloud cover, is it likely to snow on Christmas?
  - Given a person's credit history, is he or she likely to default on a mortgage?

- One naïve way of forecasting y is simply to plunge ahead with the basic regression equation

$$E(Y_i|X_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

- Since Y_i can only take the values 0 or 1, the expected value of $Y^{(i)}$ is simply a weighted average of these two cases:

$$E(Y_i|X_i) = 1 \times P(Y_i = 1|X_i) + 0 \times P(Y_i = 0|X_i) = P(Y_i = 1|X_i)$$

- Therefore, the regression equation is just a linear model for the conditional probability that $Y_i = 1$, given the predictor X_i:

$$P(Y_i = 1|X_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

$$(0, 1) \qquad\qquad (-\infty, \infty)$$

- Suppose outcome Y_i is a (binary) random variable $Y_i \sim \text{Bernulli}(\pi_i)$
  - The mean is defined as
$$\mu_i = E(Y_i) = \pi_i$$
  - Then, the variance is
$$\text{var}(Y_i) = \pi_i(1 - \pi_i) = \mu_i(1 - \mu_i)$$

- **Generalized Linear Model for Binary Data is then modeled as**
  - ***The link function***
$$g\big(E(Y_i)\big) = g(\pi_i) = \eta_i$$
$$g(\pi_i) = \text{logit}(\pi_i) \qquad g\colon (0,1) \to (-\infty, \infty)$$
$$\Rightarrow \text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$
  - ***The variance function***
$$\text{var}(Y_i) = \phi V\big(E(Y_i)\big) = \phi V(\pi_i)$$
$$V(\pi_i) = \pi_i(1 - \pi_i)$$
$$\Rightarrow \text{var}(Y_i) = \phi \times \pi_i(1 - \pi_i)$$

Assumption of Logistic Regression

- Assumptions of the Logistic Regression Model

  ✓ The ith observation has the $\mathrm{Bernulli}(\pi_i)$ distribution. Each observation has its own probability of success

  ✓ The logit is linked to the linear predictor

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$$

$$\pi_i = \frac{(\exp \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}$$

  ✓ The observations are all independent of each other

Likelihood of Logistic Regression

- The likelihood of a single observation y_i is the probability of a $\mathrm{Bernulli}(\pi_i)$ where π_i is a function of the $n+1$ parameters $\beta_0, \ldots, \beta_n$

$$
\begin{aligned}
f(y_i|\beta_0, \ldots, \beta_n) &= (\pi_i)^{y_i}(1-\pi_i)^{1-y_i} \\
&= \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i) \\
&= \frac{(\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}))^{y_i}}{1+\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}
\end{aligned}
$$

- The joint likelihood all the sample is the product of the individual likelihood

$$
\begin{aligned}
f(y_1, \ldots, y_m|\beta_0, \ldots, \beta_n) &= \prod_{i=1}^{m} f(y_i|\beta_0, \ldots, \beta_n) \\
&= \prod_{i=1}^{m}\left(\frac{(\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}))^{y_i}}{1+\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}\right) \\
&= \exp\left(\beta_0 \Sigma y_i + \Sigma \beta_j \Sigma x_{ij} y_i\right)\prod_{i=1}^{m}\left(\frac{1}{1+\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}\right)
\end{aligned}
$$

Parameter Estimation of Logistic Regression using MLE

- The frequentist approach to estimation in the logistic regression model would be to find the maximum likelihood estimators.

  - MLE estimator finds the simultaneous solutions of

$$\frac{\partial \log f(y_1, \dots, y_m | \beta_0, \dots, \beta_n)}{\partial \beta_j} = 0 \;\; for\, j = 0, \dots, n$$

  - In general, it may be messy to find the simultaneous solution of these equations algebraically

  - MLE estimators can be iteratively reweighted least squares

Bayesian Approach to Logistic Regression

- In the Bayesian approach, we want to find the posterior distribution of the parameters given the data

$$p(\beta_0, \dots, \beta_n | y_1, \dots, y_m) = \frac{p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) p(\beta_0, \dots, \beta_n)}{p(y_1, \dots, y_m)}$$

$$= \frac{p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) p(\beta_0, \dots, \beta_n)}{\int_{\beta_0, \dots, \beta_n} p(y_1, \dots, y_m, \beta_0, \dots, \beta_n)}$$

✓ Likelihood is given as

$$p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) = \prod_{i=1}^{n} \left(\frac{(\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in}))^{y_i}}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in})} \right)$$

✓ Prior capturing the belief on the parameters can be represented as

$$p(\beta_0, \dots, \beta_n) = \mathrm{N}(\mathbf{b}_o, \mathbf{V}_o)$$

$$\mathbf{b}_o = \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix}, \mathbf{V}_o = \begin{pmatrix} s_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_n^2 \end{pmatrix}$$

(We will employ sampling strategies to infer posterior distribution in the next lecture)

Motivation for Poisson Regression

- In many situations, we would like to forecast ***the number of a event***, given some relevant information:
  - Given time and whether in a city, what is the number of cars passing by?
  - Given a certain disease, what is the number of survivals after 1-year ?
  - Given stock market records today, what will be the number transactions tomorrow?

- One naïve way of forecasting Y is simply to plunge ahead with the basic regression equation

$$E(Y_i|X_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

$(0, \infty)$ $(-\infty, \infty)$

Poisson Regression as GLMs

- Suppose outcome Y_i is a count variable following $Y_i \sim \text{Poisson}(\lambda_i)$
  - The mean is defined as

$$\mu_i = E(Y_i) = \lambda_i$$

  - Then, the variance is

$$\text{var}(Y_i) = \lambda_i$$

- **Generalized Linear Model for Count Data is then modeled as**
  - ***The link function***

$$g\big(E(Y_i)\big) = g(\lambda_i) = \eta_i$$
$$g(\lambda_i) = \log(\lambda_i) \qquad g\colon (0,\infty) \to (-\infty,\infty)$$
$$\Rightarrow \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots + \beta_n x_{in}$$

  - ***The variance function***

$$\text{var}(Y_i) = \phi V\big(E(Y_i)\big) = \phi V\big(\lambda^{(i)}\big)$$
$$V(\lambda_i) = \lambda_i$$
$$\Rightarrow \text{var}(Y_i) = \phi \times \lambda_i$$

Assumption of Poisson Regression

- Assumptions of the Poisson Regression Model

  ✓ The ith observation has the $\text{Poisson}(\lambda_i)$ distribution. Each observation has its own probability distribution

  ✓ The **log function** (link function) is linked to the linear predictor

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$$

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})$$

  ✓ The observations are all independent of each other

Likelihood of Poisson Regression

- The likelihood of a single observation y_i is the probability of a $\mathrm{Bernulli}(\pi_i)$ where π_i is a function of the $n+1$ parameters $\beta_0, \dots, \beta_n$

$$
\begin{aligned}
f(y_i|\beta_0,\dots,\beta_n) &\propto \lambda_i^{y_i}\exp(-\lambda_i) \\
&\propto (\exp(\eta_i))^{y_i}\exp(-\exp(\eta_i)) \\
&\propto (\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}))^{y_i}\exp(-\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}))
\end{aligned}
$$

- The joint likelihood all the sample is the product of the individual likelihood

$$
\begin{aligned}
f(y_1,\dots,y_m|\beta_0,\dots,\beta_n) &\propto \prod_{i=1}^{m} f(y_i|\beta_0,\dots,\beta_n) \\
&\propto \prod_{i=1}^{m} \lambda_i^{y_i}\exp(-\lambda_i) \\
&\propto \exp(-\Sigma\lambda_i)\prod_{i=1}^{m}\lambda_i^{y_i} \\
&\propto \exp(-\Sigma\exp(\eta_i))\prod_{i=1}^{m}(\exp(\eta_i))^{y_i} \\
&\propto \exp(-\Sigma\exp\left(\Sigma x_{ij}\beta_j\right))\ \exp\left(\Sigma y_i \Sigma x_{ij}\beta_j\right)
\end{aligned}
$$

Parameter Estimation of Poisson Regression using MLE

- The frequentist approach to estimation in the logistic regression model would be to find the maximum likelihood estimators.

  - MLE estimator finds the simultaneous solutions of

$$\frac{\partial \log f(y_1, \dots, y_m | \beta_0, \dots, \beta_n)}{\partial \beta_j} = 0 \;\; for\, j = 0, \dots, n$$

  - In general, it may be messy to find the simultaneous solution of these equations algebraically

  - MLE estimators can be iteratively reweighted least squares

Bayesian Approach to Poisson Regression

- In the Bayesian approach, we want to find the posterior distribution of the parameters given the data

$$p(\beta_0, \dots, \beta_n | y_1, \dots, y_m) = \frac{p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) p(\beta_0, \dots, \beta_n)}{p(y_1, \dots, y_m)}$$

$$= \frac{p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) p(\beta_0, \dots, \beta_n)}{\int_{\beta_0, \dots, \beta_n} p(y_1, \dots, y_m, \beta_0, \dots, \beta_n)}$$

✓ Likelihood is given as

$$p(y_1, \dots, y_m | \beta_0, \dots, \beta_n) = \exp\left(-\Sigma \exp\left(\Sigma x_{ij}\beta_j\right)\right) \ \exp\left(\Sigma y_i \Sigma x_{ij}\beta_j\right)$$

✓ Prior capturing the belief on the parameters can be represented as

$$p(\beta_0, \dots, \beta_n) = \mathrm{N}(\mathbf{b}_o, \mathbf{V}_o)$$

$$\mathbf{b}_o = \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix}, \mathbf{V}_o = \begin{pmatrix} s_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_n^2 \end{pmatrix}$$

(We will employ sampling strategies to infer posterior distribution in the next lecture)

Motivation to Survival Analysis

- Sometimes we have observed times until some event occurs for a sample of individuals or items.
  - The survival times of individuals in the study?
  - The time until failure of an object operating in a controlled high stress test setting?

- Data of this type is called survival time data, and the event is referred to as "death"

- A Poisson process often is used to model the waiting time until an event
  - ✓ when arrivals occur according to a Poisson process, the waiting time distribution follows the exponential distribution.

The Proportional Hazards Model

- Let T be the random variable the time until "death" of something. Suppose its density is given by the exponential distribution:

$$f(t) = \lambda e^{-\lambda t} \quad for \quad t > 0$$

- The probability of death by time t is given by the cumulative distribution function (CDF) of the random variable and is

$$F(t) = \int_0^t f(t)dt = \int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t}$$

- The survival function is the probability of surviving to time t and is given by

$$S(t) = P(T > t) = 1 - F(t) = e^{-\lambda t}$$

- The **hazard function** gives the instantaneous probability of death at time t given survival up until time t. It is given by

$$h(t) = \frac{f(t)}{S(t)} = \lambda$$

  - Thus, when time until death follows the exponential distribution, the hazard function will be constant.

Assumption of the Proportional Hazard Model

- Each individual has their own constant hazard function, Individual i has hazard function

$$h_i(t) = \lambda e^{\eta_i}$$

  - ✓ We will express the parameter η_i as a linear function of the predictor variables

$$\eta_i = \sum_{j=1}^{n} x_{ij}\beta_j$$

- For each individual we have
  - ✓ t_i which is either time of death, or time at end of study
  - ✓ $w_i = \begin{cases} 0 & \text{observation is censored} \\ 1 & \text{observation is not censored} \end{cases}$

If $w_i = 0$, we don't know T_i, the time of death of ith individual, we only know that $T_i > t_i$
If $w_i = 1$, we know $T_i = t_i$, we know the time of death exactly

Likelihood for Censored Survival Data

- The contribution to the likelihood of an individual that died is given by $f_i(t)$, and the contribution of an individual that is alive at the end of the study is $S_i(t)$.
- The likelihood of individual i is

$$\lambda \rightarrow \lambda e^{\eta_i}$$
$$f(t) = \lambda e^{-\lambda t} \rightarrow \lambda e^{\eta_i} e^{-\lambda t_i e^{\eta_i}}$$
$$S(t) = e^{-\lambda t} \rightarrow e^{-\lambda t_i e^{\eta_i}}$$

$$\begin{aligned}
L_i\big((t_i, w_i)|\eta_i\big) &= (f_i(t))^{w_i}(S_i(t))^{1-w_i} \\
&= \left(\lambda e^{-\lambda t_i}\right)^{w_i}\left(e^{-\lambda t_i}\right)^{1-w_i} \\
&= \left(\lambda e^{\eta_i} e^{-\lambda e^{\eta_i} t_i}\right)^{w_i}\left(e^{-\lambda e^{\eta_i} t_i}\right)^{1-w_i} \\
&= (\lambda e^{\eta_i})^{w_i} \times e^{-\lambda e^{\eta_i} t_i} \\
&= e^{-\lambda e^{\eta_i} t_i}[\lambda e^{\eta_i}]^{w_i} \\
&= e^{-\lambda e^{\eta_i} t_i}[\lambda t_i e^{\eta_i}]^{w_i} \times \left(\frac{1}{t_i}\right)^{w_i}
\end{aligned}$$

- The likelihood of the whole sample equals the product of the individual likelihoods

$$\begin{aligned}
L\big((t_1, w_1), \dots, (t_n, w_n)|\eta_1, \dots, \eta_n\big) &= \prod_{i=1}^{n} L_i\big((t_i, w_i)|\eta_i\big) \\
&= \prod_{i=1}^{n} e^{-\lambda e^{\eta_i} t_i}[\lambda t_i e^{\eta_i}]^{w_i} \times \left(\frac{1}{t_i}\right)^{w_i}
\end{aligned}$$

Likelihood for Censored Survival Data

- The likelihood of the whole sample equals the product of the individual likelihoods

$$L\big((t_1, w_1), \dots, (t_n, w_n) | \eta_1, \dots, \eta_n\big) = \prod_{i=1}^{n} L_i\big((t_i, w_i) | \eta_i\big)$$

$$= e^{-\Sigma \lambda e^{\eta_i} t_i} \prod_{i=1}^{n} [\lambda e^{\eta_i} t_i]^{w_i} \times \prod_{i=1}^{n} (t_i)^{-w_i}$$

- Let us parameterize to the form $\mu_i = \lambda e^{\eta_i} t_i$

$$L(w_1, \dots, w_n | \mu_1, \dots, \mu_n) \propto e^{-\Sigma \mu_i} \prod_{i=1}^{n} \mu_i^{w_i}$$

  - ✓ This is similar to the likelihood for a random sample of n independent Poisson random variables with parameters $\mu_1, \dots, \mu_n$

$$L(y_1, \dots, y_n | \lambda_1, \dots, \lambda_n) \propto e^{-\Sigma \lambda_i} \prod_{i=1}^{n} \lambda_i^{y_i}$$

  - ✓ This means that given λ, we can treat the censoring variables w_i as a independent random sample of Poisson random variables with respective parameters μ_i

- In terms of the parameters $\beta_0, \dots, \beta_n$ the likelihood becomes

$$L(w_1, \dots, w_n | \beta_0, \dots, \beta_n) \propto e^{-t_i \Sigma e^{x_{ij} \beta_j}} \prod_{i=1}^{n} \left(t_i \Sigma e^{x_{ij} \beta_j}\right)^{w_i}$$