

Module 4: Classification Part 1

TMA4268 Statistical Learning V2023

Daesoo Lee,

Department of Mathematical Sciences, NTNU



NTNU

Norwegian University of
Science and Technology

30/01/2023



Overview

Overview

- Classification and discrimination
- Logistic regression
- Bayes classifier
- KNN



What is classification?

What is classification?

$$\begin{array}{|c|} \hline y_1 \\ \hline y_2 \\ \hline \vdots \\ \hline y_n \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \hline 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 1 & x_{n1} & x_{n2} & \cdots & x_{np} \\ \hline \end{array} \cdot \begin{array}{|c|} \hline \beta_0 \\ \hline \beta_1 \\ \hline \vdots \\ \hline \beta_p \\ \hline \end{array} + \begin{array}{|c|} \hline \epsilon_1 \\ \hline \epsilon_2 \\ \hline \vdots \\ \hline \epsilon_n \\ \hline \end{array}$$

$$Y = X\beta + \epsilon$$

- Y is a qualitative variable for regression.
- Y can be a quantitative variable \rightarrow *classification*!
 - Spam filters `email` $\in \{\text{spam}, \text{ham}\}$,
 - Eye color $\in \{\text{blue}, \text{brown}, \text{green}\}$.
 - Medical condition $\in \{\text{disease1}, \text{disease2}, \text{disease3}\}_5$

What is classification?

- We often build models that **predict probabilities of categories**, *given* X .

$$Y = X\beta + \epsilon$$

- *e.g.*, $Y \in \{\text{spam, not. spam}\}$
- *e.g.*, $\hat{Y} = \{0.2, 0.8\}$

What is classification?

What are the methods?

Three methods for classification are discussed here:

- Logistic regression
- K -nearest neighbors
- Linear and quadratic discriminant analysis (in the next lecture)



Logistic Linear Regression

Logistic Linear Regression

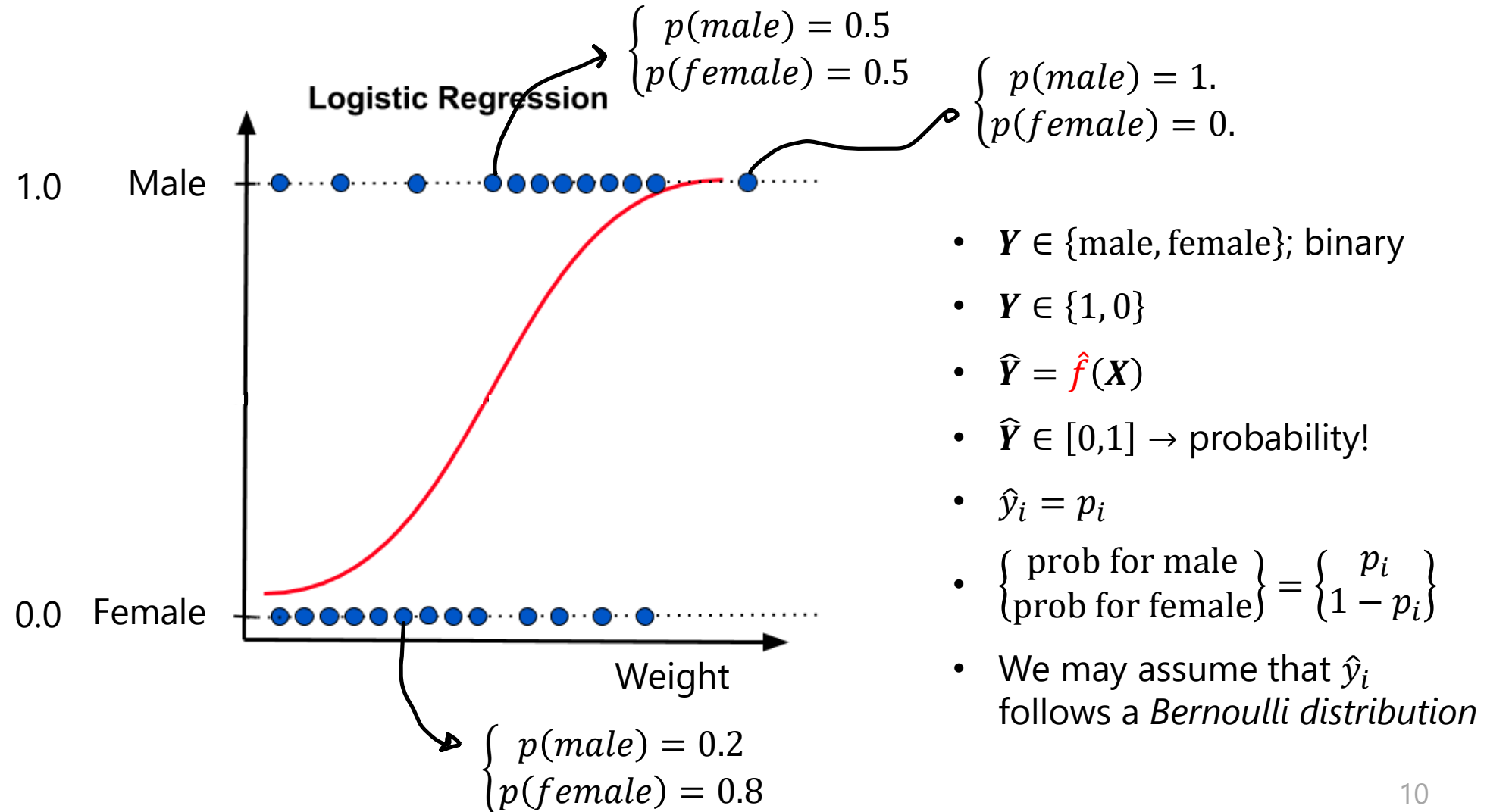
Logistic Regression



- $Y \in \{\text{male, female}\}$; binary
- $Y \in \{1, 0\}$

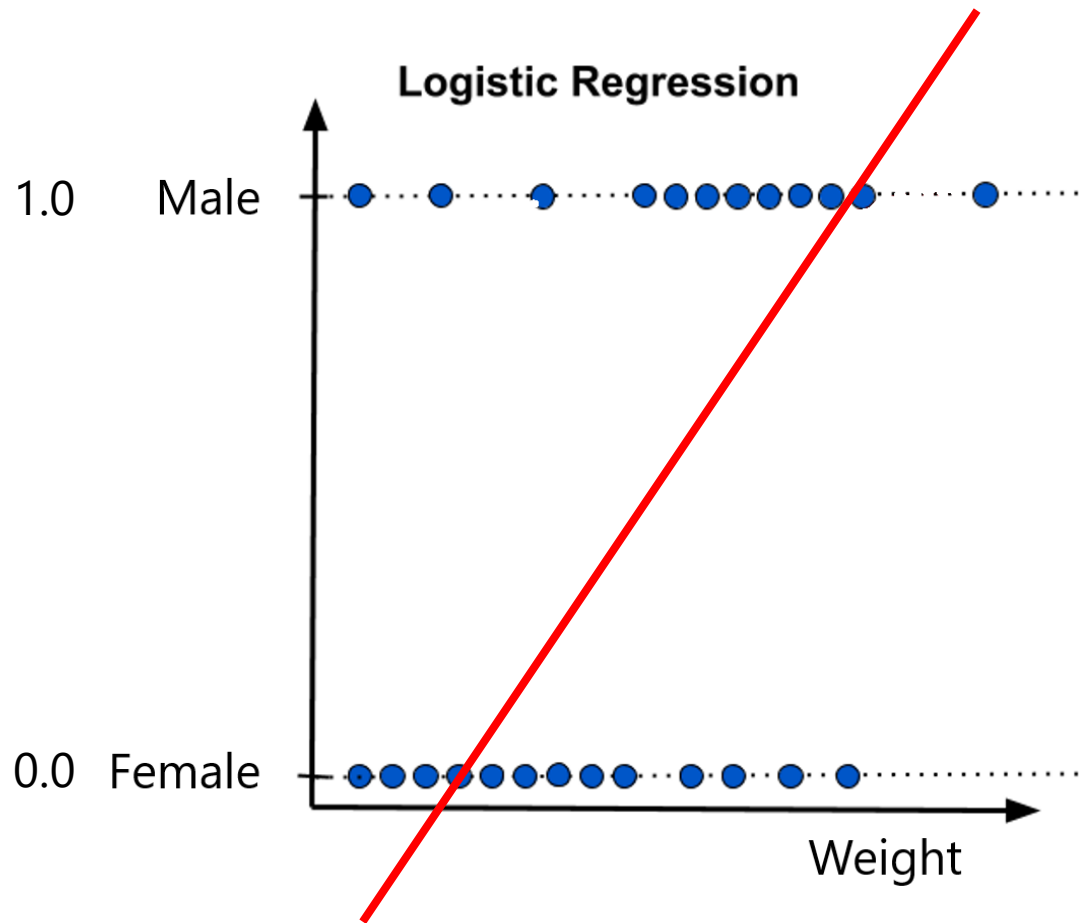
Logistic Linear Regression

Logistic Regression



Logistic Linear Regression

Why don't we just use a linear regression?

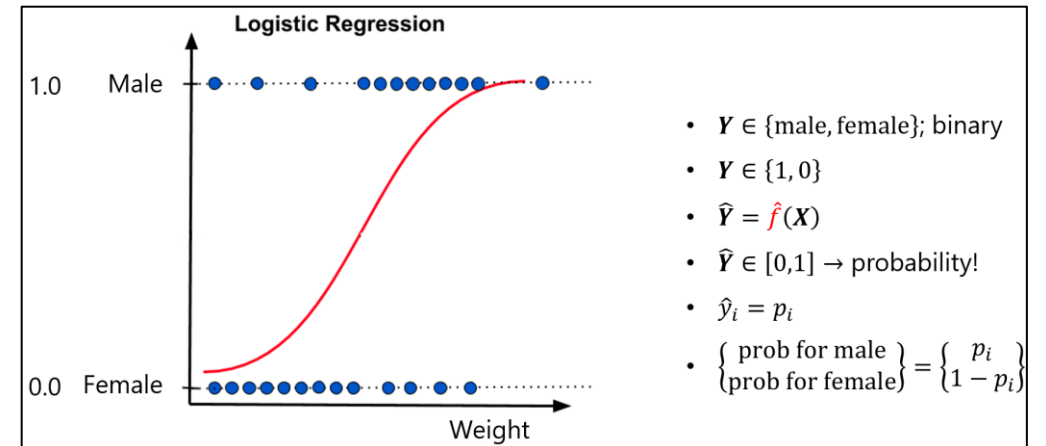
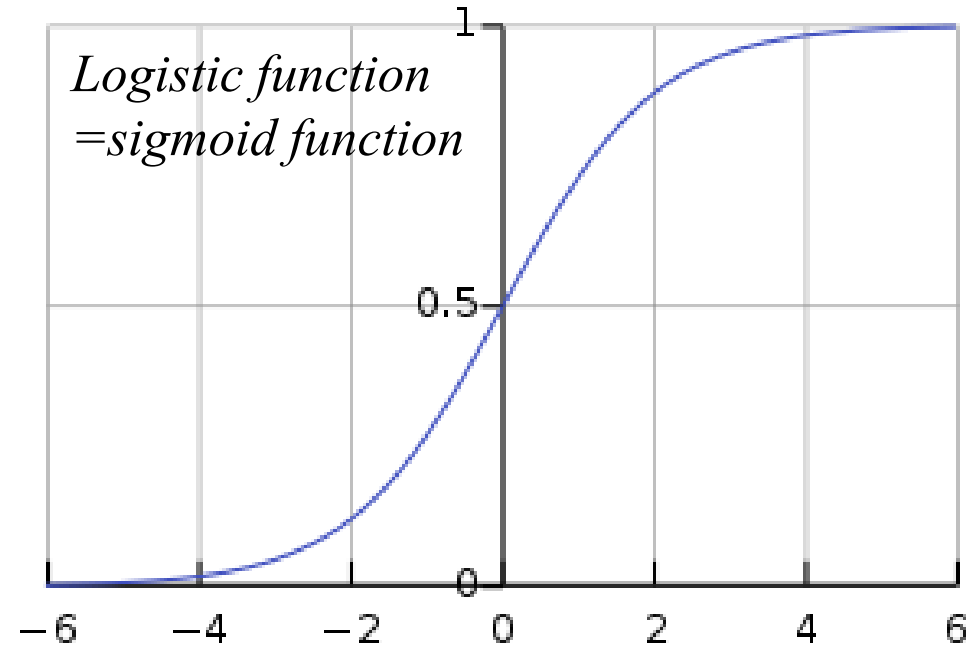


- The output of \hat{f} can vary between $-\infty$ and $+\infty$. For binary classification, the output should be within 0 and 1.
- Linear regression assumes $\epsilon \sim N(0, \sigma^2)$

Logistic Linear Regression

How to model \hat{f} ?

- $\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$
- Replace
 - 1) t with $\beta_0 + \beta_1 x$, and
 - 2) $\sigma(t)$ with $p(x)$.
- t is a linear transformation of x .
- $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$
- $\hat{f} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x)}}$



Logistic Linear Regression

Example

- We're now using the *Default* dataset from the *ISLR* library

```
{r}  
library(ISLR)  
head(Default)
```

	default <dbl>	student <fctr>	balance <dbl>	income <dbl>
1	0	No	729.5265	44361.625
2	0	Yes	817.1804	12106.135
3	0	No	1073.5492	31767.139
4	0	No	529.2506	35704.494
5	0	No	785.6559	38463.496
6	0	Yes	919.5885	7491.559

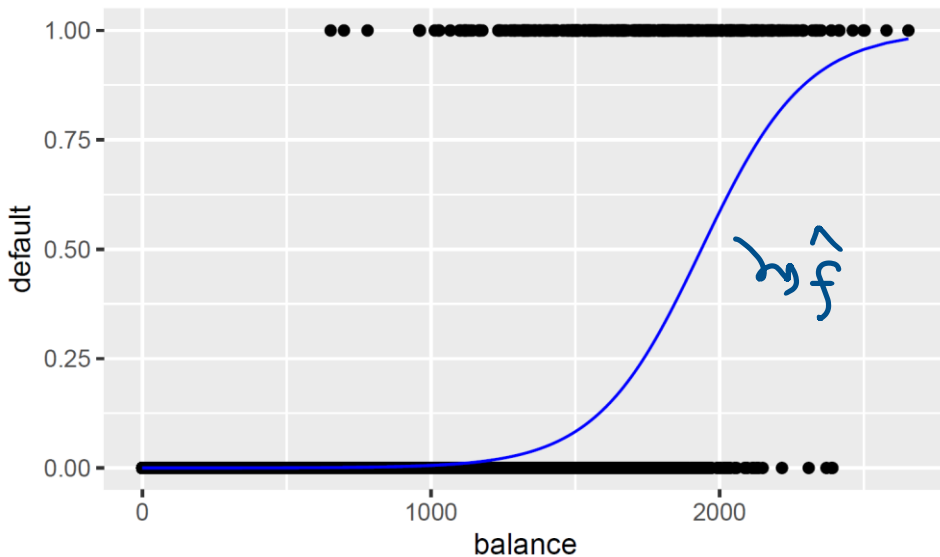
Logistic Linear Regression

Example

- $default = f(balance) + \epsilon$

default: "failure to fulfil an obligation, especially to repay a loan or appear in a law court." (0 or 1)

- $\widehat{default} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 balance)}}$



```
{r}
library(ISLR)
data(Default)
Default$default <- as.numeric(Default$default) - 1
glm_default = glm(default ~ balance, data = Default, family = "binomial")
summary(glm_default)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-10.651330614	0.3611573721	-29.49221	3.623124e-191
balance	0.005498917	0.0002203702	24.95309	1.976602e-137

Q. what's the probability of default given balance of 2,000 ?

$$\widehat{default} = \frac{1}{1 + e^{-(-10.65 + 0.0055 \cdot 2000)}} \approx 0.59$$

$$\therefore \begin{cases} \text{prob for default} \\ \text{prob for not. default} \end{cases} = \begin{cases} 0.59 \\ 1 - 0.59 \end{cases}$$

Logistic Linear Regression

Estimating the coefficients

- Remember

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$\begin{cases} \text{prob for male} \\ \text{prob for female} \end{cases} = \begin{cases} p_i \\ 1 - p_i \end{cases} \rightarrow (p_i)^{y_i} (1 - p_i)^{1 - y_i} \text{ where } y_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

- Given n independent observation pairs $\{x_i, y_i\}$, the *likelihood* function of a logistic regression model is written as:

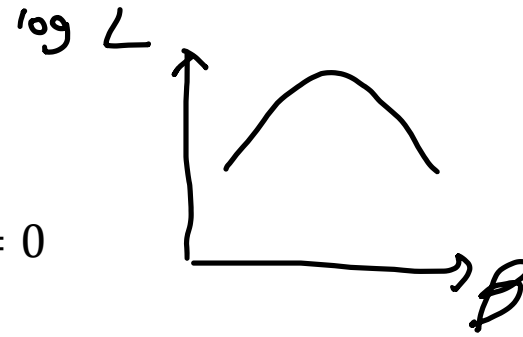
$$L(\boldsymbol{\beta}) = \prod_{i=1}^n (p_i)^{y_i} (1 - p_i)^{1 - y_i}$$

- The estimates are found by maximizing the above likelihood.
- Usually, *log-likelihood* is used instead of likelihood. (a value of likelihood becomes too small for large n)

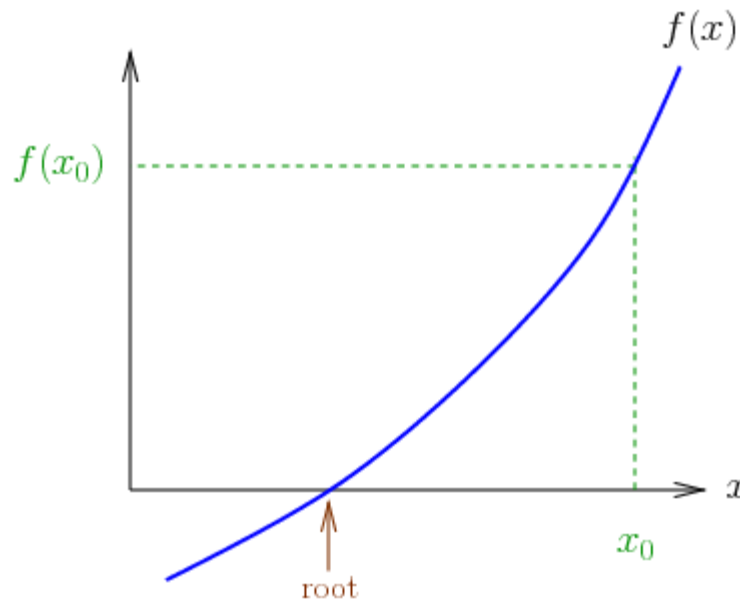
Logistic Linear Regression

Estimating the coefficients

$$\frac{\partial \log(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = 0$$



- This doesn't have an analytical solution.
- The equation is solved numerically using the *Newton-Raphson algorithm* (it's an optimization algorithm).



In our case,
 $\frac{\partial \log(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}}$ instead of $f(x)$

Logistic Linear Regression

Example (continued)

```
library(ISLR)
summary(Default)
Default$default <- as.numeric(Default$default) - 1
glm_default = glm(default ~ balance, data = Default, family = "binomial")
summary(glm_default)
```

```
call:
glm(formula = default ~ balance, family = "binomial", data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2697	-0.1465	-0.0589	-0.0221	3.7589

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.065e+01	3.612e-01	-29.49	<2e-16 ***
balance	5.499e-03	2.204e-04	24.95	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
Residual deviance: 1596.5 on 9998 degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8

- The z-statistic is equal to $\frac{\hat{\beta}}{SE(\hat{\beta})}$, and is approximately $N(0,1)$ distributed.
- Check the p -value for *Balance*. Conclusion?

Logistic Linear Regression

Odds

- $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$ can be generalized to

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}}$$

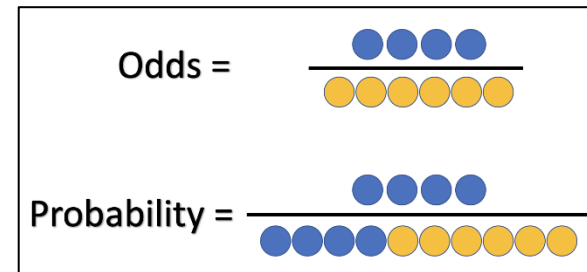
$$\Rightarrow \frac{1}{p_i} = 1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}$$

$$\Rightarrow \frac{1}{p_i} - 1 = \frac{1 - p_i}{p_i} = e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})} = \frac{1}{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

$$\Rightarrow \frac{1 - p_i}{p_i} = \frac{1}{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

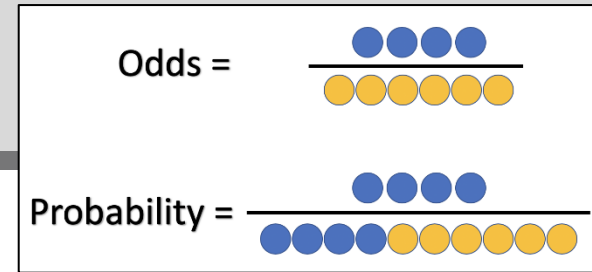
$$\Rightarrow \frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

- The quantity $p_i / (1 - p_i)$ is called *odds*. Odds represent *chances* (e.g., in betting)



$$\frac{p_i}{1 - p_i} = \frac{\frac{\text{blue}}{\text{blue} + \text{yellow}}}{1 - \frac{\text{blue}}{\text{blue} + \text{yellow}}} = \frac{\frac{\text{blue}}{\text{blue} + \text{yellow}}}{\frac{\text{yellow}}{\text{blue} + \text{yellow}}} = \frac{\text{blue}}{\text{yellow}}$$

Logistic Linear Regression



What's the deal with Odds here?

- The key point here is that we can *interpret the logistic regression in terms of odds (chances)*.
(not the formula or nitty-gritty details!)
- $\frac{p_i}{1-p_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}} = e^{\beta_0} e^{\beta_1 x_{i1}} \dots e^{\beta_p x_{ip}}$

The odds ratio

- To understand the effect of a regression coefficient β_j ,
let's see what happens if we increase x_{ij} to $x_{ij} + 1$, while all other covariates are kept fixed.
- $$\frac{\text{odds}(Y_i=1 \mid X_j=x_{ij}+1)}{\text{odds}(Y_i=1 \mid X_j=x_{ij})} = \frac{e^{\beta_0} e^{\beta_1 x_{i1}} \dots e^{\beta_j (x_{ij}+1)} \dots e^{\beta_p x_{ip}}}{e^{\beta_0} e^{\beta_1 x_{i1}} \dots e^{\beta_j x_{ij}} \dots e^{\beta_p x_{ip}}} = \frac{e^{\beta_j (x_{ij}+1)}}{e^{\beta_j x_{ij}}} = \frac{e^{\beta_j x_{ij}} e^{\beta_j}}{e^{\beta_j x_{ij}}} = e^{\beta_j}$$
- Interpretation: By increasing covariate x_{ij} by one, we change the odds for $Y_i = 1$ by a factor of e^{β_j} .
NB! The odds ratio of 1 represents no change.

Logistic Linear Regression

$$\frac{\text{odds}(Y_i = 1 \mid X_j = x_{ij} + 1)}{\text{odds}(Y_i = 1 \mid X_j = x_{ij})} = \frac{e^{\beta_j(x_{ij}+1)}}{e^{\beta_j x_{ij}}} = e^{\beta_j}$$

NB! The odds ratio of 1 represents no change.

Example

```
```\nlibrary(ISLR)\ndata(Default)\nDefault$default <- as.numeric(Default$default) - 1\nglm_default = glm(default ~ balance + income + student, data = Default, family = "binomial")\nsummary(glm_default)$coefficients\n```\n
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.086905e+01	4.922555e-01	-22.080088	4.911280e-108
balance	5.736505e-03	2.318945e-04	24.737563	4.219578e-135
income	3.033450e-06	8.202615e-06	0.369815	7.115203e-01
studentYes	-6.467758e-01	2.362525e-01	-2.737646	6.188063e-03

## Questions:

- What happens with the odds to default when *income* increases by 10,000 dollars?

$$\text{odds ratio} = \frac{e^{\beta_j(x_{ij}+10000)}}{e^{\beta_j x_{ij}}} = e^{\beta_j 10000} = e^{(3.03 \cdot 10^{-6}) \cdot 10000} = 1.03$$

- What happens with the odds to default when *balance* increases by 100 dollars?

$$\text{odds ratio} = \frac{e^{\beta_j(x_{ij}+100)}}{e^{\beta_j x_{ij}}} = e^{\beta_j 100} = e^{(5.74 \cdot 10^{-3}) \cdot 100} = 1.78$$



# The Bayes classifier

# The Bayes classifier

## Bayes Classifier

- Assume that we can estimate the probability that a new observation  $x_0$  belongs to class  $k$ , for  $K$ .  
The probability that  $Y = k$  given  $x_0$  is:

$$Pr(Y = k | X = x_0)$$

- Then, the Bayes classifier performs classification by

$$\operatorname{argmax}_{k \in \{1, 2, \dots, K\}} Pr(Y = k | X = x_0)$$

- For instance,

$$Pr(Y|x_0) = \begin{cases} Pr(Y = male | X = 50kg) = 0.1 \\ Pr(Y = female | X = 50kg) = 0.9 \end{cases}$$
$$\operatorname{argmax}(Pr(Y|x_0)) = female$$

- What if you have  $\mathbf{x} = (x_0, x_1, x_2, \dots, x_p)$  instead of  $x_0$ ? ( $p$  denotes a number of features)
- $Pr(Y = k | X = \mathbf{x})$  becomes much more complex! To tackle this issue, *Naïve Bayes classifier* is introduced.

# The Bayes classifier

## Naïve Bayes Classifier

- $Pr(Y = k | X = \mathbf{x}) = p(Y_k|\mathbf{x})$  where  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  and  $p$  denotes a number of features.
- Our problem was that  $p(Y_k|\mathbf{x})$  becomes much more complex than  $p(Y_k|x_j)$ .

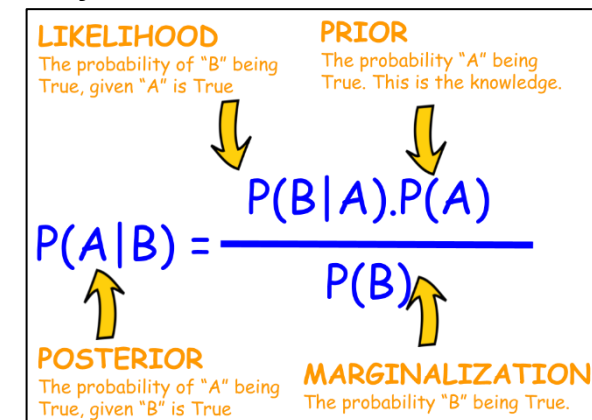
- Using the Bayes theorem,  $p(Y_k|\mathbf{x})$  can be expressed as

$$p(Y_k|\mathbf{x}) \propto p(\mathbf{x}|Y_k)p(Y_k) = p(x_1, x_2, \dots, x_p|Y_k)p(Y_k)$$

- With the independent assumption -- all features in  $\mathbf{x}$  are mutually independent (the term "naïve" comes from here):

$$p(Y_k|\mathbf{x}) \propto p(x_1|Y_k)p(x_2|Y_k) \cdots p(x_p|Y_k)p(Y_k)$$

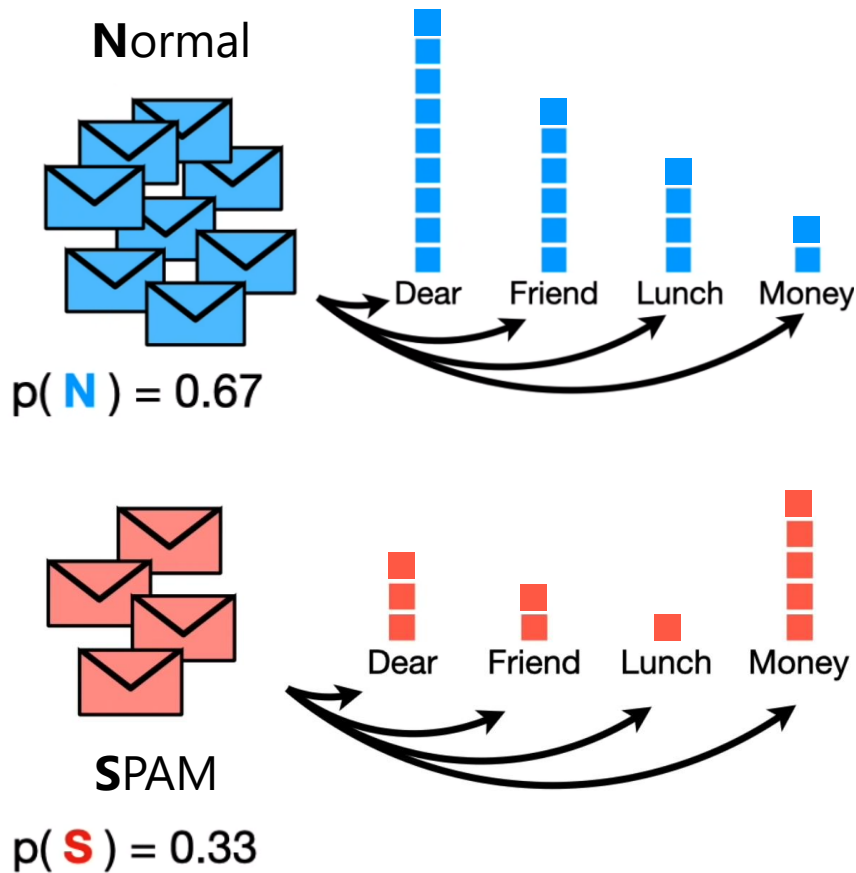
## Bayes Theorem



# The Bayes classifier

$$p(Y_k | \mathbf{x}) \propto p(x_1 | Y_k) p(x_2 | Y_k) \cdots p(x_p | Y_k) p(Y_k)$$

## Example: Naïve Bayes Classifier



$$\begin{aligned}p(\mathbf{Dear} | \mathbf{N}) &= 0.43 \\p(\mathbf{Friend} | \mathbf{N}) &= 0.29 \\p(\mathbf{Lunch} | \mathbf{N}) &= 0.19 \\p(\mathbf{Money} | \mathbf{N}) &= 0.10\end{aligned}$$

$$\begin{aligned}p(\mathbf{Dear} | \mathbf{S}) &= 0.27 \\p(\mathbf{Friend} | \mathbf{S}) &= 0.18 \\p(\mathbf{Lunch} | \mathbf{S}) &= 0.09 \\p(\mathbf{Money} | \mathbf{S}) &= 0.45\end{aligned}$$

### Email: "Dear Friend"

$$\begin{aligned}p(\mathbf{N} | \text{"Dear Friend"}) &= p(\mathbf{Dear} | \mathbf{N}) p(\mathbf{Friend} | \mathbf{N}) p(\mathbf{N}) \\&= 0.43 \cdot 0.29 \cdot 0.67 = 0.084\end{aligned}$$

$$\begin{aligned}p(\mathbf{S} | \text{"Dear Friend"}) &= p(\mathbf{Dear} | \mathbf{S}) p(\mathbf{Friend} | \mathbf{S}) p(\mathbf{S}) \\&= 0.27 \cdot 0.18 \cdot 0.33 = 0.016\end{aligned}$$

### Email: "Lunch Money Money Money"

$$\begin{aligned}p(\mathbf{N} | \text{"Lunch Money Money Money"}) &= p(\mathbf{Lunch} | \mathbf{N}) p(\mathbf{Money} | \mathbf{N})^3 p(\mathbf{N}) \\&= 0.19 \cdot 0.1^3 \cdot 0.67 = 0.00127\end{aligned}$$

$$\begin{aligned}p(\mathbf{S} | \text{"Lunch Money Money Money"}) &= p(\mathbf{Lunch} | \mathbf{S}) p(\mathbf{Money} | \mathbf{S})^3 p(\mathbf{S}) \\&= 0.09 \cdot 0.45^3 \cdot 0.33 = 0.0027\end{aligned}$$



# The Bayes classifier

## Properties of the Bayes classifier

- The overall Bayes error rate is given as

$$1 - E \left[ \max_j \Pr(Y = j | X) \right]$$

where the expectation is over  $X$ .

# The Bayes classifier

## Training error

- Training error rate

$$\frac{1}{n} \sum_{i=1}^n \mathbf{I}(y_i \neq \hat{y}_i)$$

$\mathbf{I}$  is an indicator function and is defined as:

$$\mathbf{I} = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{else} \end{cases}$$

- The training error rate is the fraction of misclassification made on our training set.

# The Bayes classifier

## Test error

- *Test error* is the same as the training error except that it is calculated on our *test set*.
- This gives a *better indication of the true performance* of the classifier than the training error.
- We assume that a *good classifier* is a classifier with a *low test error*.

### Assessing the Model Accuracy

#### Measuring the Quality of Fit



You study for an exam, and often with the previous years' exams.

$X_{train}$  (training dataset)



You take the exam.

$X_{test}$  (test dataset)

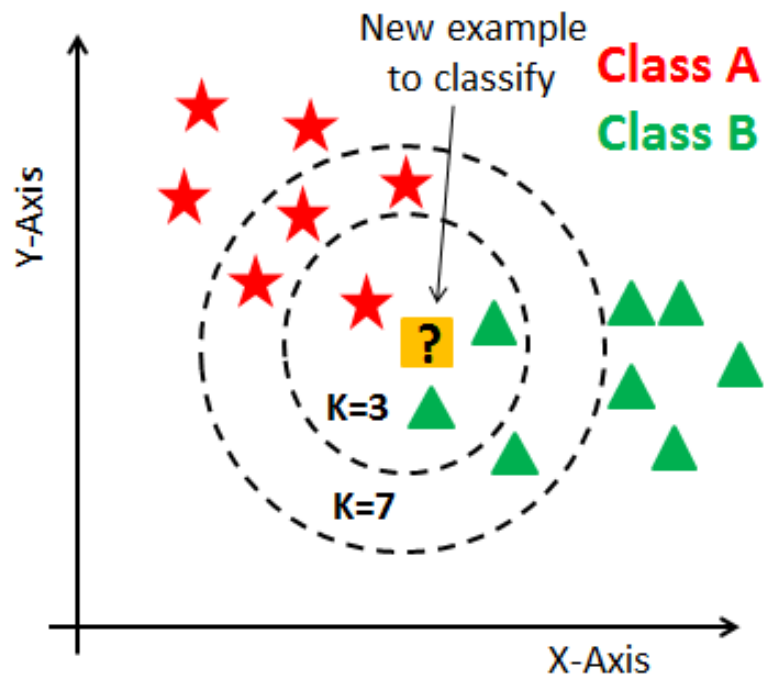
Remember this from Module 2!



# K-Nearest Neighbor (KNN) classifier

# K-Nearest Neighbor (KNN) classifier

- KNN classifier estimates  $p(Y_k|\mathbf{x})$  *non-parametrically*.
- Classification is done by a *majority vote*:



- $K = 3$

$$\begin{cases} p(\text{green}|\mathbf{x}) = 2/3 \\ p(\text{red}|\mathbf{x}) = 1/3 \end{cases}$$

- $K = 7$

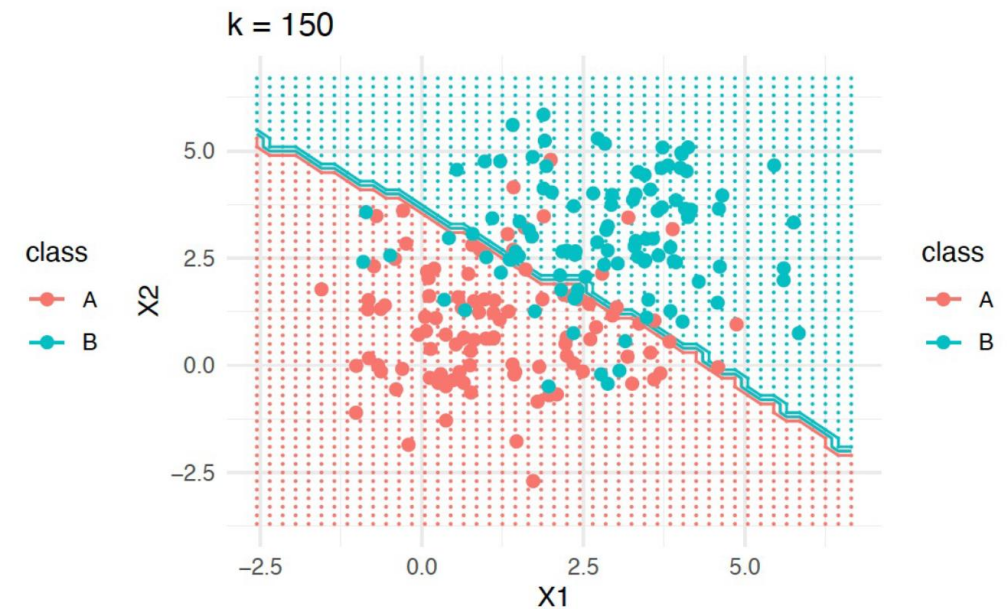
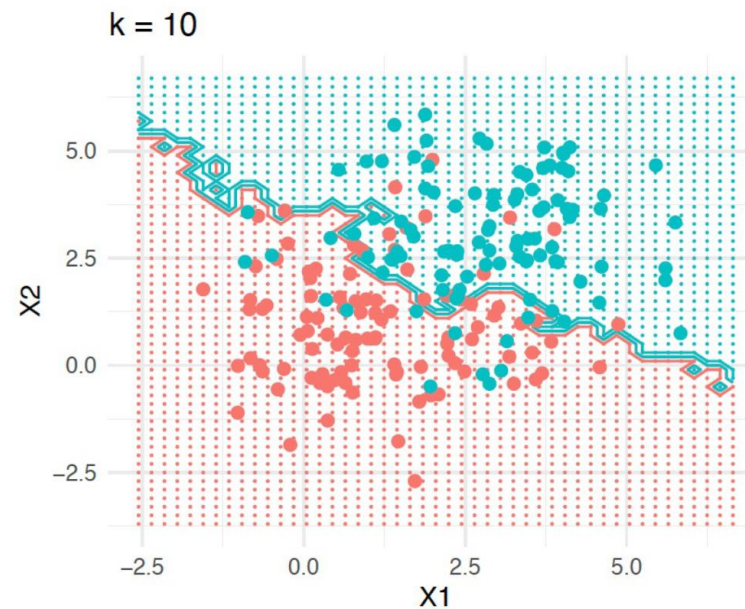
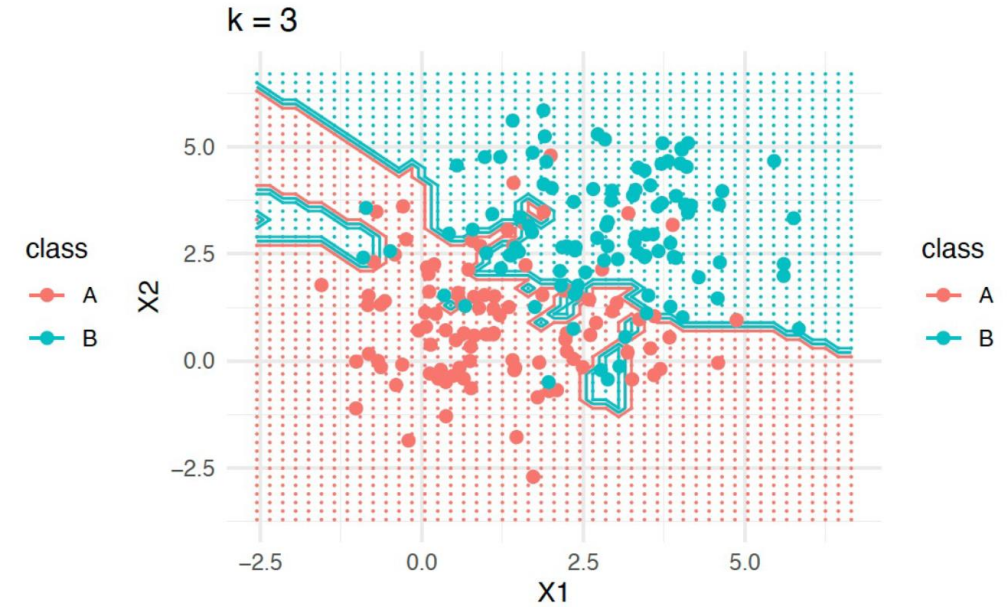
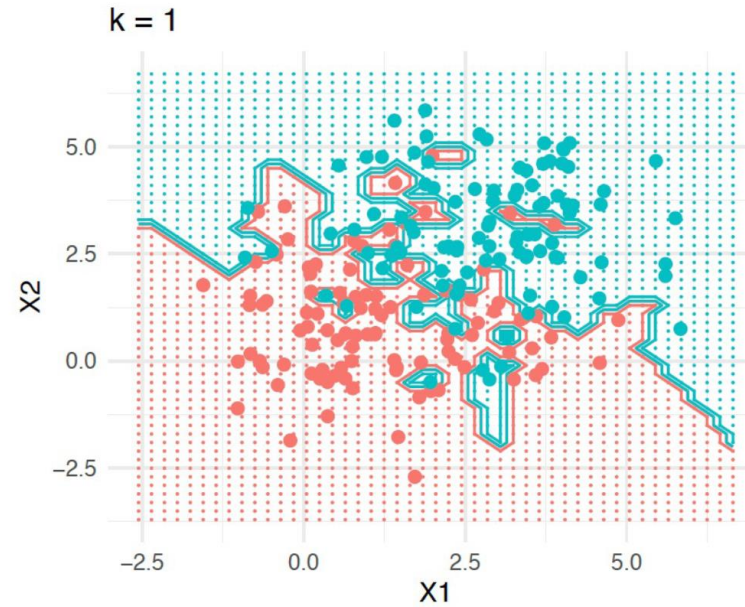
$$\begin{cases} p(\text{green}|\mathbf{x}) = 3/7 \\ p(\text{red}|\mathbf{x}) = 4/7 \end{cases}$$

- $K$ : number of neighbors

# K-Nearest Neighbor (KNN) classifier

## KNN

- Big dots: training data
- Little dots: test data



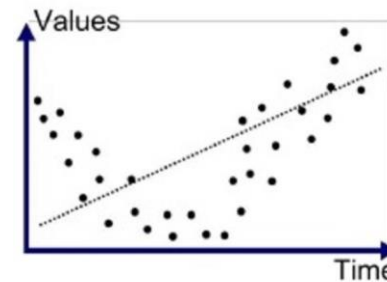
# K-Nearest Neighbor (KNN) classifier

## How to choose $K$ ?

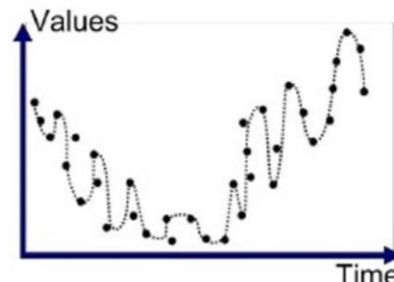
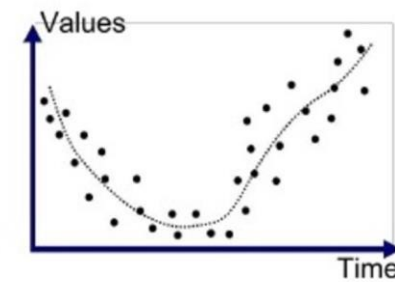
- $K = 1$ : the classification is made to the same class as the one nearest neighbor.
- $K$  large: the decision boundary tends towards a straight line.

## Discussion:

- When is the bias large? When is the variance large?
- How to find the optimal  $K$ ?



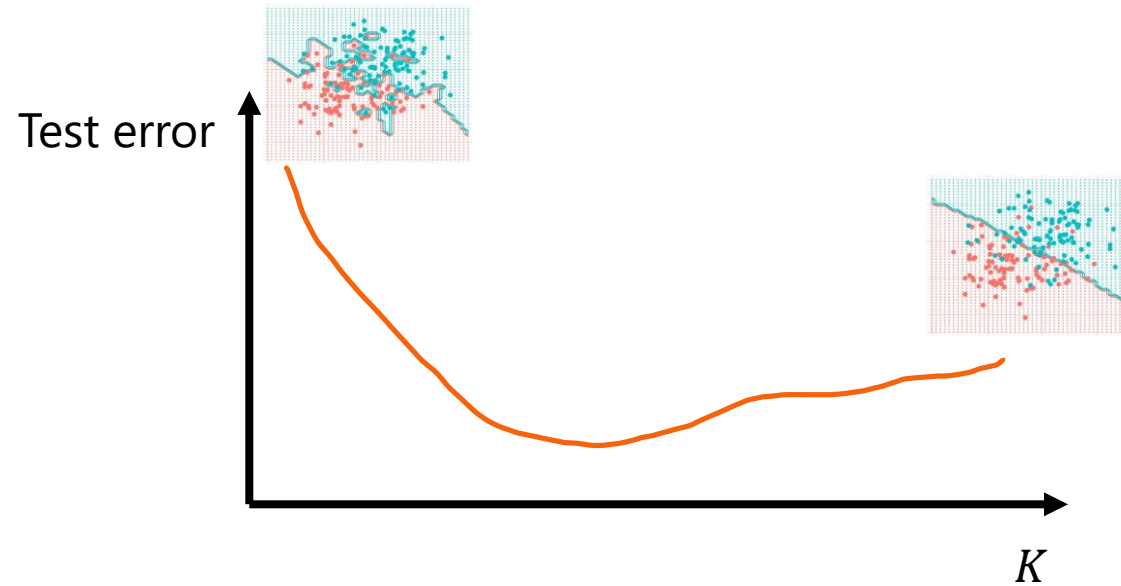
High Bias  
Low Variance



Low Bias  
High Variance

# K-Nearest Neighbor (KNN) classifier

**How to find the optimal  $K$ ?**

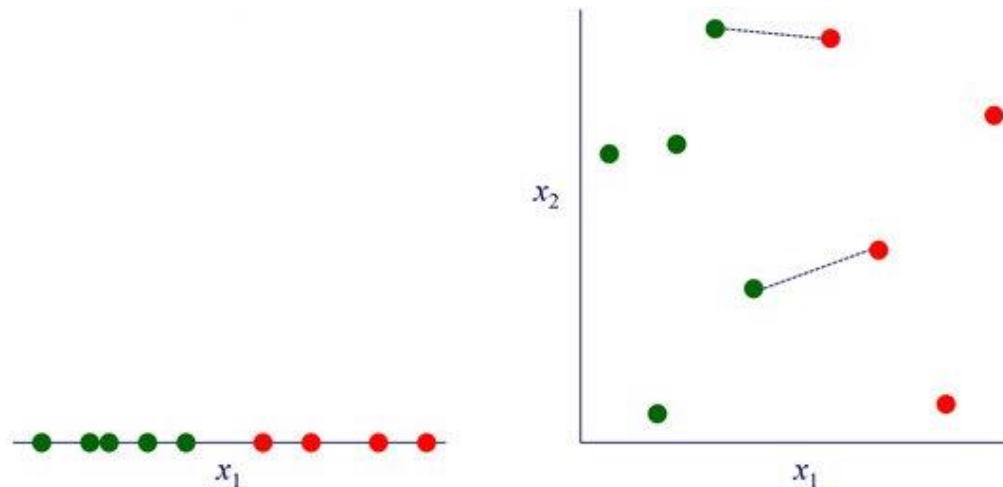




# K-Nearest Neighbor (KNN) classifier

## The curse of dimensionality

- KNN can be quite good if the number of predictors  $p$  is small and the number of observations  $n$  is large. We need enough close neighbors to make a good classification.
- The effectiveness of the KNN classifier falls quickly when the dimension of the predictor space is high.
- Why?  
Because the nearest neighbors tend to be far away in high dimensional space.  
KNN is based on “nearest neighbors” and neighbors are going far away.. Not good!





# Two Paradigms for Classification

# Two Paradigms for Classification

## Two approaches to estimate $\Pr(Y = k \mid X = x)$

### Diagnostic Paradigm

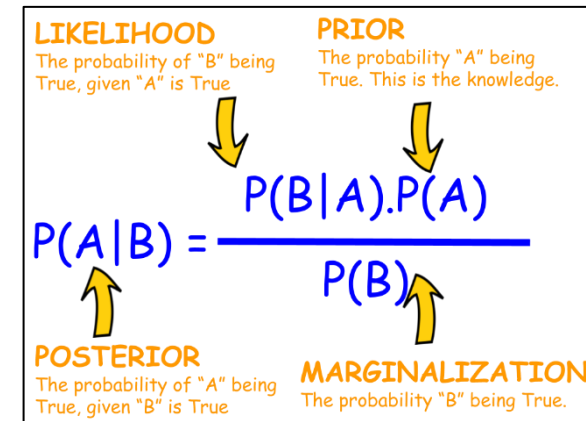
- *Directly* estimating  $\Pr(Y = k \mid X = x)$
- *e.g.*, Logistic regression, KNN classification

### Sampling Paradigm

- *Indirectly* estimating  $\Pr(Y = k \mid X = x)$   
by modeling the likelihood  $\Pr(X = x \mid Y = k)$  and the prior  $\Pr(Y = k)$ .

$$\Pr(Y = k \mid X = x) \propto \Pr(X = x \mid Y = k) \Pr(Y = k)$$

### Bayes Theorem



*Remember Naïve Bayes Classifier?*

$$p(Y_k | \mathbf{x}) \propto p(x_1 | Y_k) p(x_2 | Y_k) \cdots p(x_p | Y_k) p(Y_k)$$

A group of six students are running away from the camera down a school hallway. They are all wearing backpacks. The student on the far left is a girl with curly hair, wearing a blue backpack and plaid shorts. Next to her is a boy with a blue backpack and a yellow shirt. In the center is a girl with long dark hair, wearing a purple backpack and jeans. To her right is a boy with a black backpack and a blue shirt, who has his arms raised in the air. Next is a girl with a teal shirt and dark shorts, carrying a brown bag. On the far right is a boy with a brown backpack and a green shirt. The hallway has large windows on the left and a white wall on the right. The text "The End!" is overlaid in the center of the image.

The End!