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?

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

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{Y}$ is a qualitative variable for regression.
- $\boldsymbol{Y}$ can be a quantitative variable → *classification!*

- Spam filters `email` $\in$ {spam, ham},
- `Eye color` $\in$ {blue, brown, green}.
- `Medical condition` $\in$ {disease1, disease2, disease3}

What is classification?

- We often build models that **predict probabilities of categories**, *given* $\boldsymbol{X}$.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

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

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)

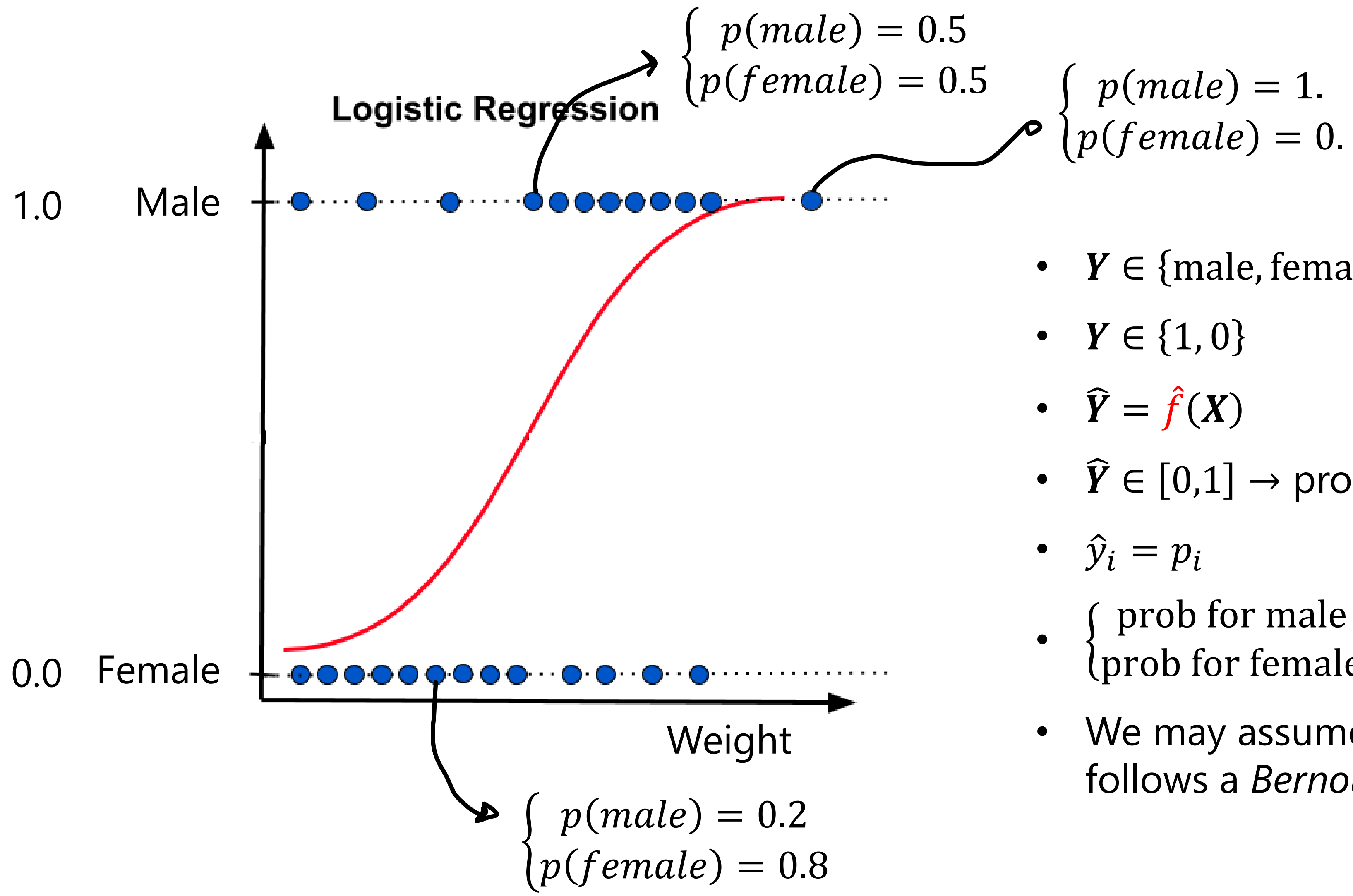
Logistic Linear Regression

Logistic Regression

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

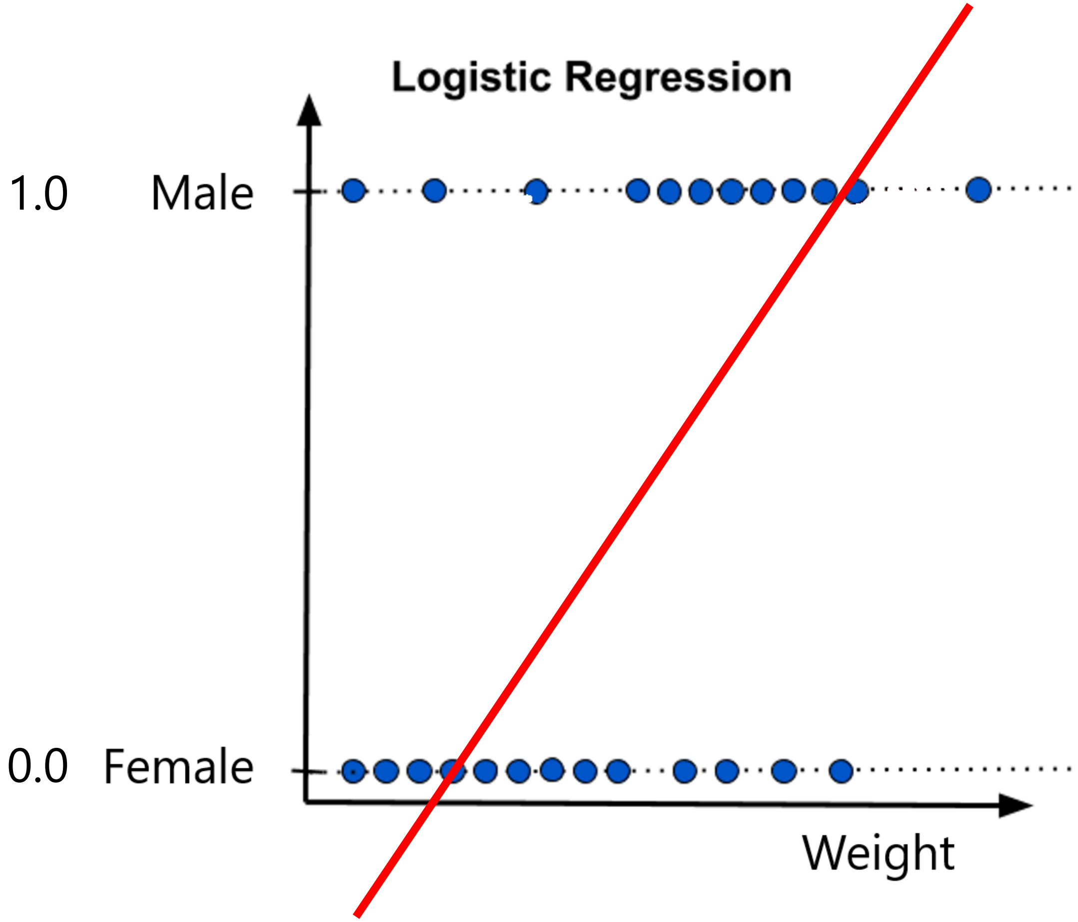
Logistic Linear Regression

Logistic Regression



- $\boldsymbol{Y} \in \{\text{male}, \text{female}\}$; binary
- $\boldsymbol{Y} \in \{1, 0\}$
- $\widehat{\boldsymbol{Y}} = \hat{f}(\boldsymbol{X})$
- $\widehat{\boldsymbol{Y}} \in [0,1] \rightarrow$ probability!
- $\hat{y}_i = p_i$
- $\begin{Bmatrix} \text{prob for male} \\ \text{prob for female} \end{Bmatrix} = \begin{Bmatrix} p_i \\ 1 - p_i \end{Bmatrix}$
- We may assume that $\hat{y}_i$ follows a *Bernoulli distribution*

Why don't we just use a linear regression?

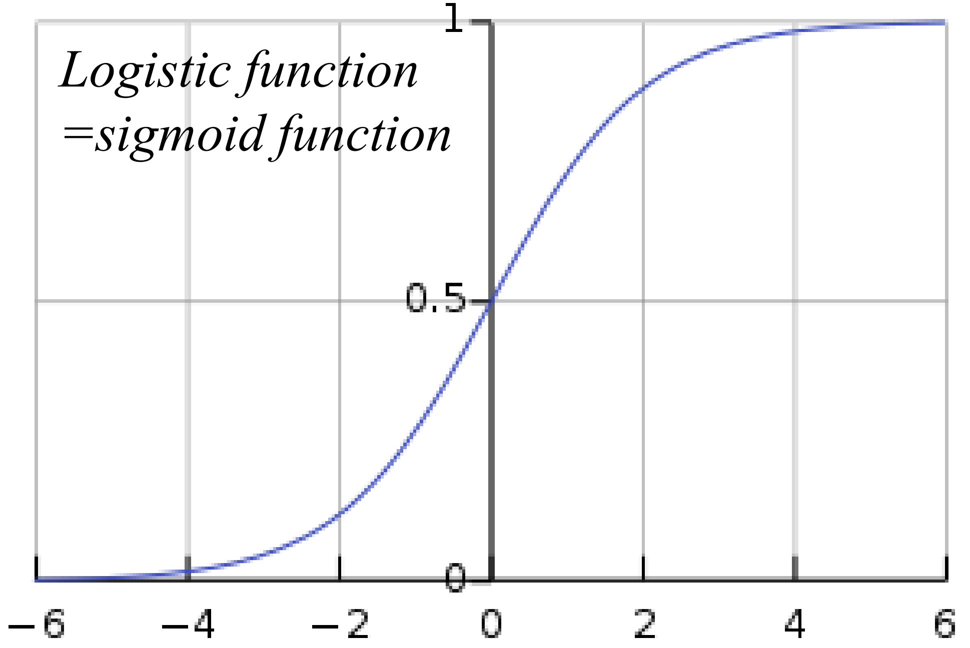


- The output of $\hat{f}$ can vary between $-\infty$ and $+\infty$.
  For binary classification, the output should be within 0 and 1.
- Linear regression assumes $\epsilon \sim N(0, \sigma^2)$

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.
- $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
- $\hat{f} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x}} = \frac{1}{1+e^{-(\widehat{\beta}_0+\widehat{\beta}_1 x)}}$



Example

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

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

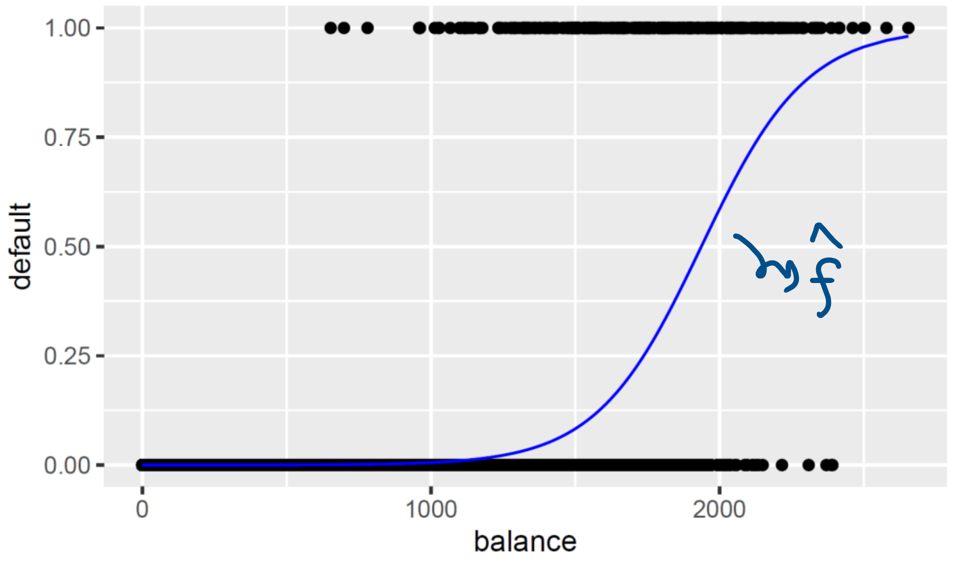
	default	student	balance	income
	<dbl>	<fctr>	<dbl>	<dbl>
1	0	No	729.5265	44361.625
2	0	Yes	817.1804	12106.135
3	0	No	1073.5492	31767.139
4	0	No	529.2506	35704.494
5	0	No	785.6559	38463.496
6	0	Yes	919.5885	7491.559

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 ?

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

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

Logistic Linear Regression

Estimating the coefficients

- Remember

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

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

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

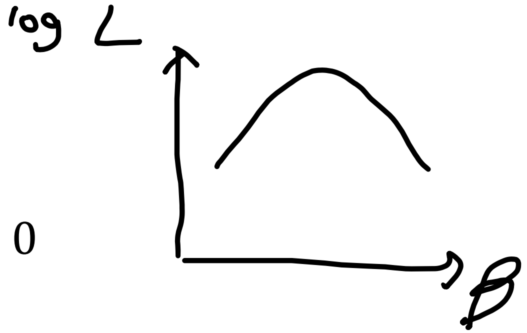
$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} (p_i)^{y_i}(1-p_i)^{1-y_i}$$

- 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)

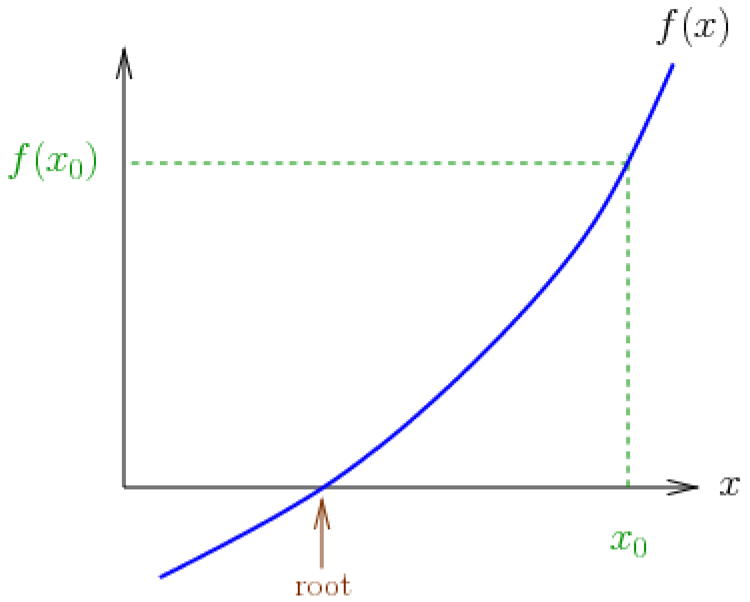
Logistic Linear Regression

Estimating the coefficients

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



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



Example (continued)

```{r}
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

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

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

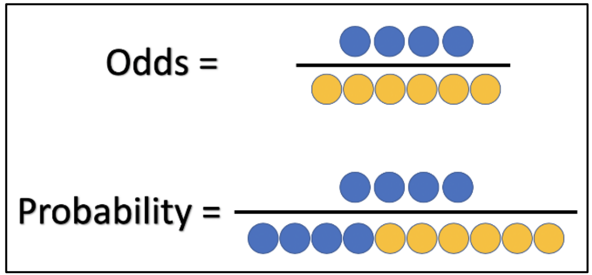
$$\Rightarrow \frac{1}{p_i} - 1 = \frac{1-p_i}{p_i} = e^{-\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)} = \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}$$

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

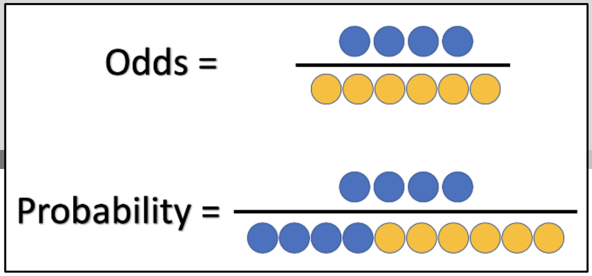
$$\Rightarrow \frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}$$

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

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



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!)

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

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.

- $\frac{\text{odds}(Y_i=1 \mid X_j=x_{ij}+1)}{\text{odds}(Y_i=1 \mid X_j=x_{ij})} = \frac{e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_j (x_{ij}+1)} \ldots e^{\beta_p x_{ip}}}{e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_j x_{ij}} \ldots e^{\beta_p x_{ip}}} = \frac{e^{\beta_j (x_{ij}+1)}}{e^{\beta_j x_{ij}}} = \frac{e^{\beta_j x_{ij}} e^{\beta_j}}{e^{\beta_j x_{ij}}} = e^{\beta_j}$

- Interpretation: By increasing covariate x_{ij} by one unit, we change the odds for $Y_i = 1$ by a factor of e^{β_j}.

Logistic Linear Regression

Example

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

```
                 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?

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

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

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

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:

$$Pr(Y = k \mid X = x_0)$$

- Then, the Bayes classifier is

$$C^{\text{Bayes}}(x_0) = \text{argmax}_{k\in\{1,2,\dots,K\}}\, Pr(Y = k \mid X = x_0)$$

- For instance,

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

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

The Bayes classifier

Naïve Bayes Classifier

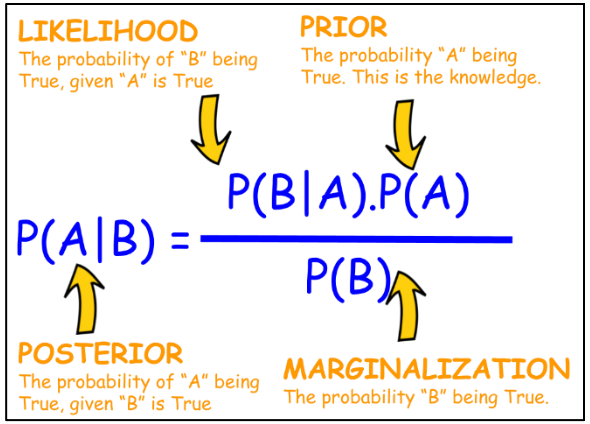
- $Pr(Y = k \mid X = \boldsymbol{x}) = p(Y_k|\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ and p denotes a number of features.
- Our problem was that $p(Y_k|\boldsymbol{x})$ becomes much more complex than $p(Y_k|x_j)$.
- Using the Bayes theorem, $p(Y_k|\boldsymbol{x})$ can be expressed as

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

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

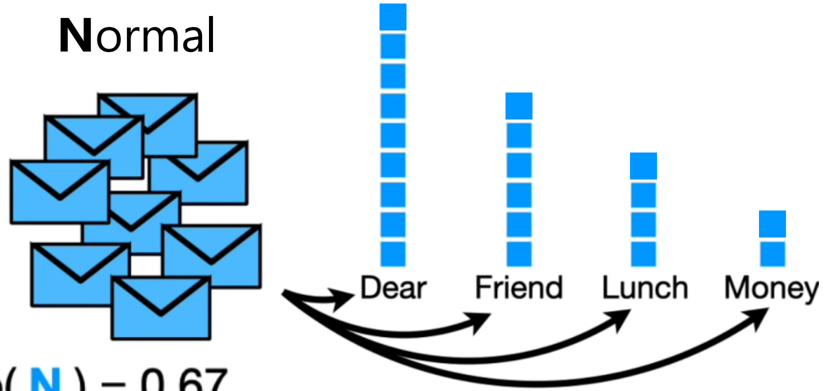
$$p(Y_k|\boldsymbol{x}) \propto p(x_1|Y_k)p(x_2|Y_k)\cdots p(x_p|Y_k)p(Y_k)$$

Bayes Theorem



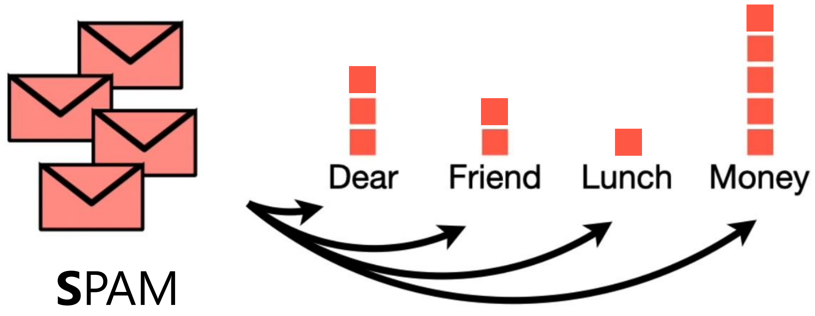
$$p(Y_k|\boldsymbol{x}) \propto p(x_1|Y_k)p(x_2|Y_k)\cdots p(x_p|Y_k)p(Y_k)$$

Example: Naïve Bayes Classifier



p(**N**) = 0.67

p(**Dear** | **N**) = 0.43
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.19
p(**Money** | **N**) = 0.10



p(**S**) = 0.33

p(**Dear** | **S**) = 0.27
p(**Friend** | **S**) = 0.18
p(**Lunch** | **S**) = 0.09
p(**Money** | **S**) = 0.45

Email: “Dear Friend”

$$p(\mathrm{N}\mid\text{"Dear Friend"}) = p(\mathrm{N})p(\mathrm{Dear}|\mathrm{N})p(\mathrm{Friend}|\mathrm{N}) = 0.67\cdot 0.43\cdot 0.29 = 0.084$$

$$p(\mathrm{S}\mid\text{"Dear Friend"}) = p(\mathrm{S})p(\mathrm{Dear}|\mathrm{S})p(\mathrm{Friend}|\mathrm{S}) = 0.33\cdot 0.27\cdot 0.18 = 0.016$$

Email: “Lunch Money Money Money”

$$p(\mathrm{N}\mid\text{"Lunch Money Money Money"}) = p(\mathrm{N})p(\mathrm{Lunch}|\mathrm{N})p(\mathrm{Money}|\mathrm{N})^3 = 0.67\cdot 0.19\cdot 0.1^3 = 0.00127$$

$$p(\mathrm{S}\mid\text{"Lunch Money Money Money"}) = p(\mathrm{S})p(\mathrm{Lunch}|\mathrm{S})p(\mathrm{Money}|\mathrm{S})^3 = 0.33\cdot 0.09\cdot 0.45^3 = 0.0027$$

https://youtu.be/O2L2Uv9pdDA

The Bayes classifier

Properties of the Bayes classifier

- The overall Bayes error rate is given as

$$1 - \mathrm{E}\left[\max_j \Pr(Y = j \mid X)\right]$$

  where the expectation is over X.

Training error

- Training error rate

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

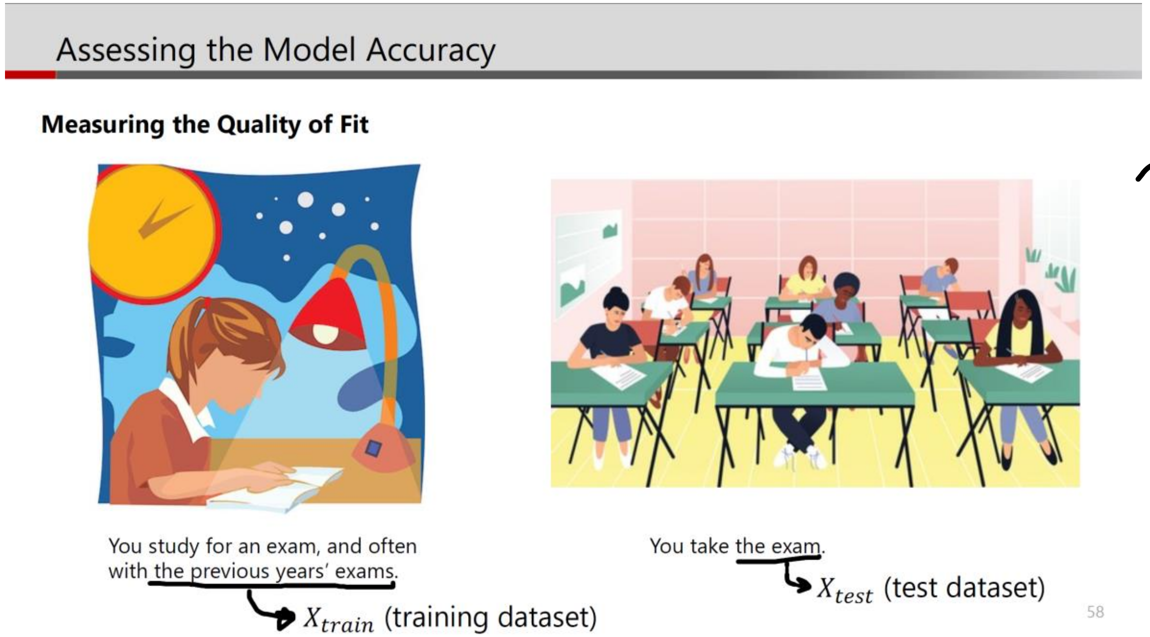
$\mathbf{I}$ is an indicator function and is defined as:

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

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

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*.

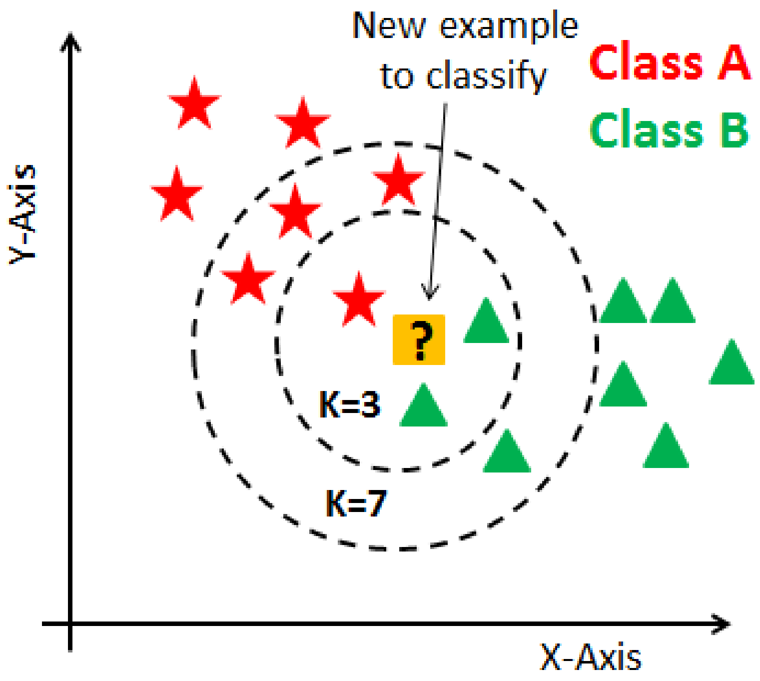


Remember this from Module 2!

K-Nearest Neighbor (KNN) classifier

K-Nearest Neighbor (KNN) classifier

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



- $K = 3$

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

- $K = 7$

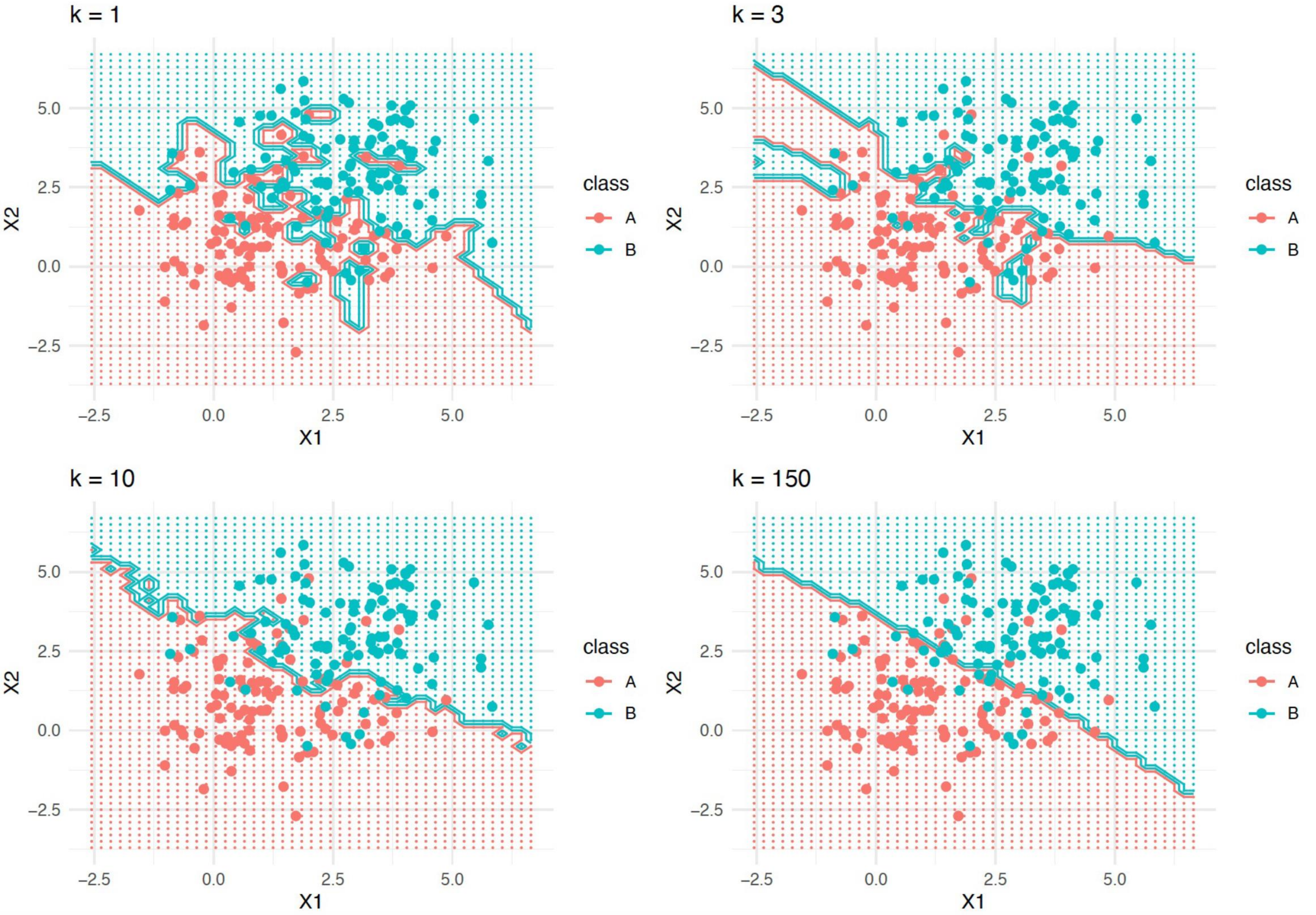
$$\begin{cases} p(green|\boldsymbol{x}) = 3/7 \\ p(red|\boldsymbol{x}) = 4/7 \end{cases}$$

- K: number of neighbors

K-Nearest Neighbor (KNN) classifier

KNN

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



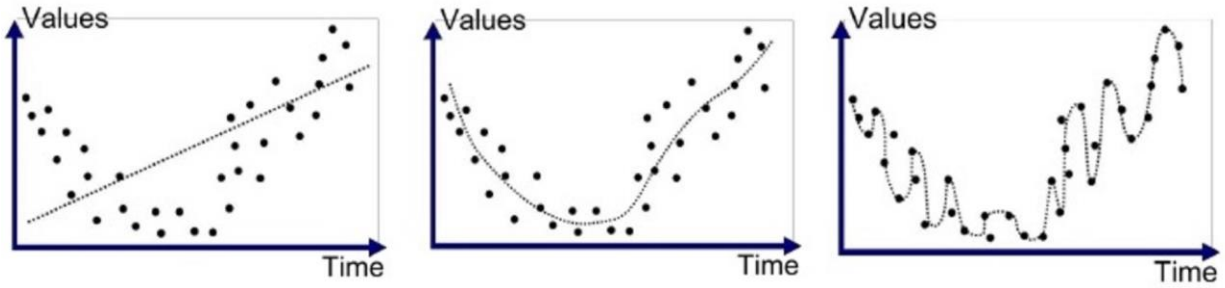
K-Nearest Neighbor (KNN) classifier

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?

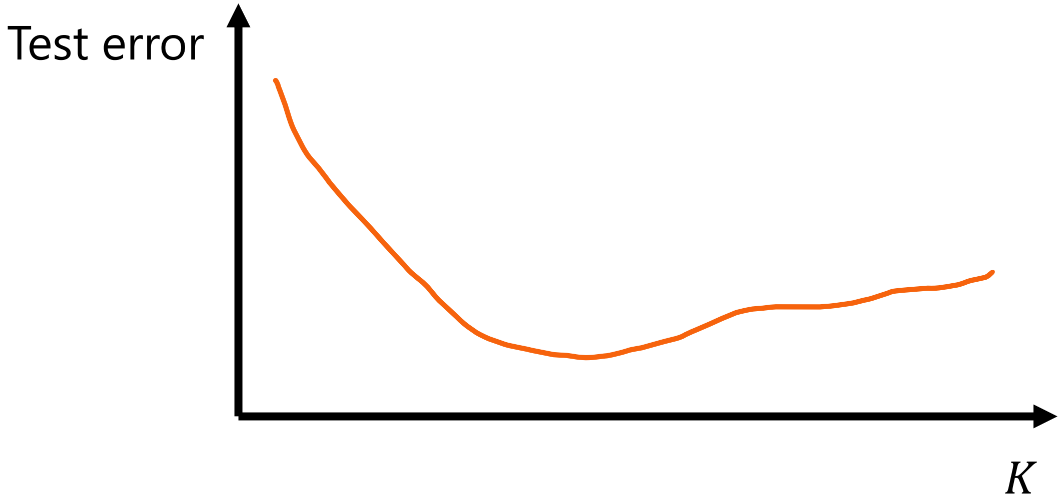


High Bias
Low Variance

Low Bias
High Variance

K-Nearest Neighbor (KNN) classifier

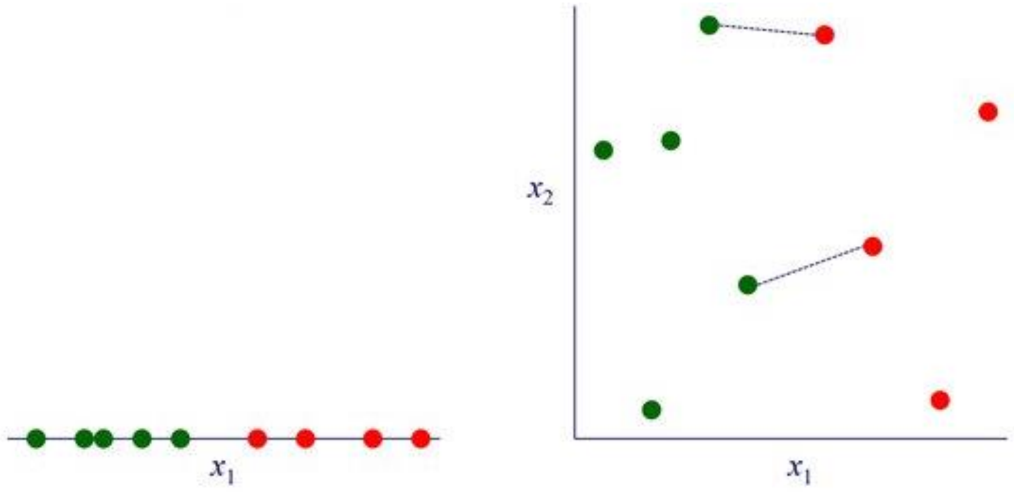
How to find the optimal K?



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

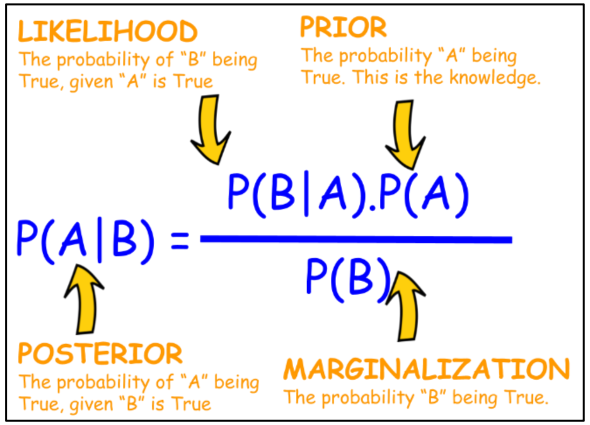
- *Directly* estimating $\Pr(Y = k \mid X = x)$
- *e.g.*, Logistic regression, KNN classification

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?

$$p(Y_k|\boldsymbol{x}) \propto p(x_1|Y_k)p(x_2|Y_k)\cdots p\left(x_p|Y_k\right)p(Y_k)$$


The End!