Ve572 Lecture 8

Manuel and Jing

UM-SJTU Joint Institute

June 12, 2018

- In linear regression, we explore the relation between the response

$$Y_i \qquad i = 1, 2, \ldots, n$$

and two or more predictors, k of them in general,

$$X_{i1}, \quad X_{i2}, \quad \ldots \quad X_{ij}, \quad \ldots \quad X_{ik} \qquad i = 1, 2, \ldots, n$$

where we assume that we observed on n cases, and the relation is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$$

the true parameters β_0, β_1, ..., β_k are unknown, thus e_i is also unknown.

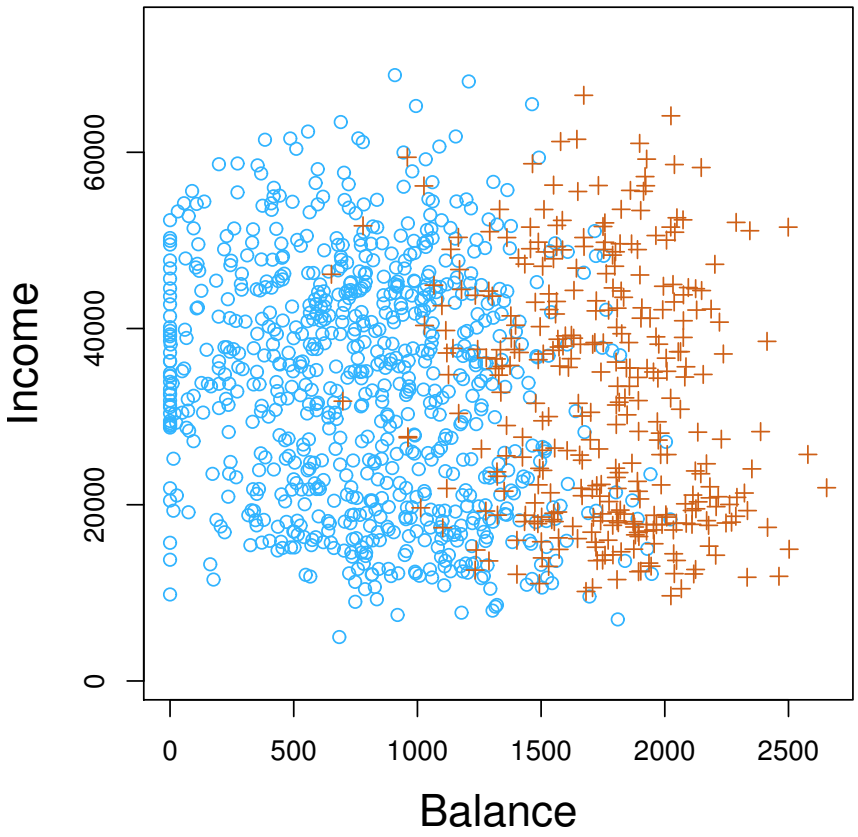
- Using the data $(y_1, x_{11}, x_{12}, \ldots, x_{1k}), \cdots, (y_n, x_{n1}, x_{n2}, \ldots, x_{nk})$, we have

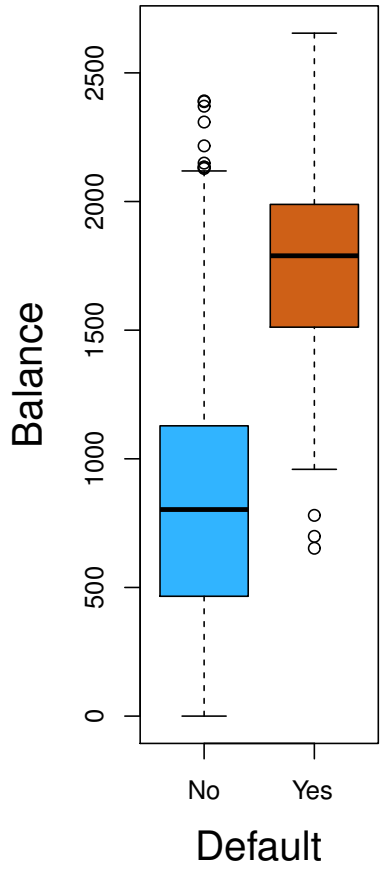
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{e}_i$$

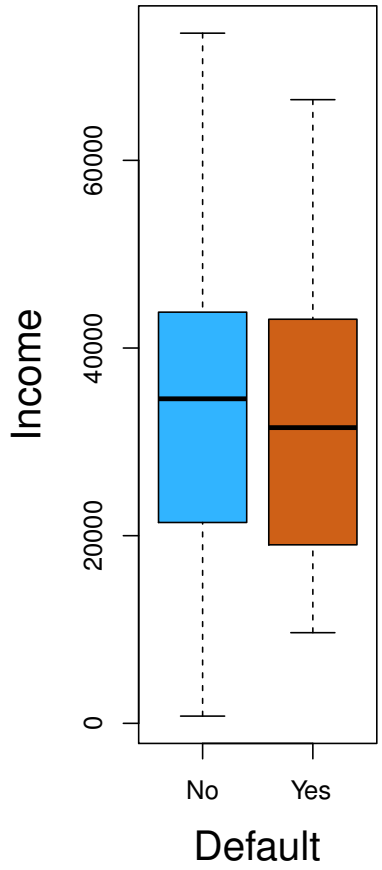
where $\hat{\beta}_i$, for $i = 0, 1, \ldots, k$, are chosen to minimise $\sum_{i=1}^{n} \hat{e}_i^2$.

- Consider the following dataset

`Income`	Average annual income
`Balance`	Average monthly credit card balance
`Default`	Whether the card has defaulted
`Student`	Whether the card belongs to a student







- The bank is interested in predicting `Default` using `Income` and `Balance`.

Q: Why linear regression is not going to be useful/meaningful here? e.g.

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where X corresponds to `Balance` and Y corresponds `Default`

$$Y = \begin{cases} 0 & \text{if } \texttt{Default} \text{ is no;} \\ 1 & \text{if } \texttt{Default} \text{ is yes.} \end{cases}$$

- Recall linear regression models the conditional mean

$$\mathbb{E}\left[Y \mid X = x\right] = \beta_0 + \beta_1 x$$

- By finding $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, we essentially find the estimate

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

to predict the conditional mean for every possible value of $X = x$.

- However, here `Default` is a categorical variable, the conditional mean

$$\mathbb{E}\left[Y \mid X = x\right] = \beta_0 + \beta_1 x$$

  is not meaningful to take any other value other than the original two values.
- Our definition of Y lead us to what is really needed for a binary variable

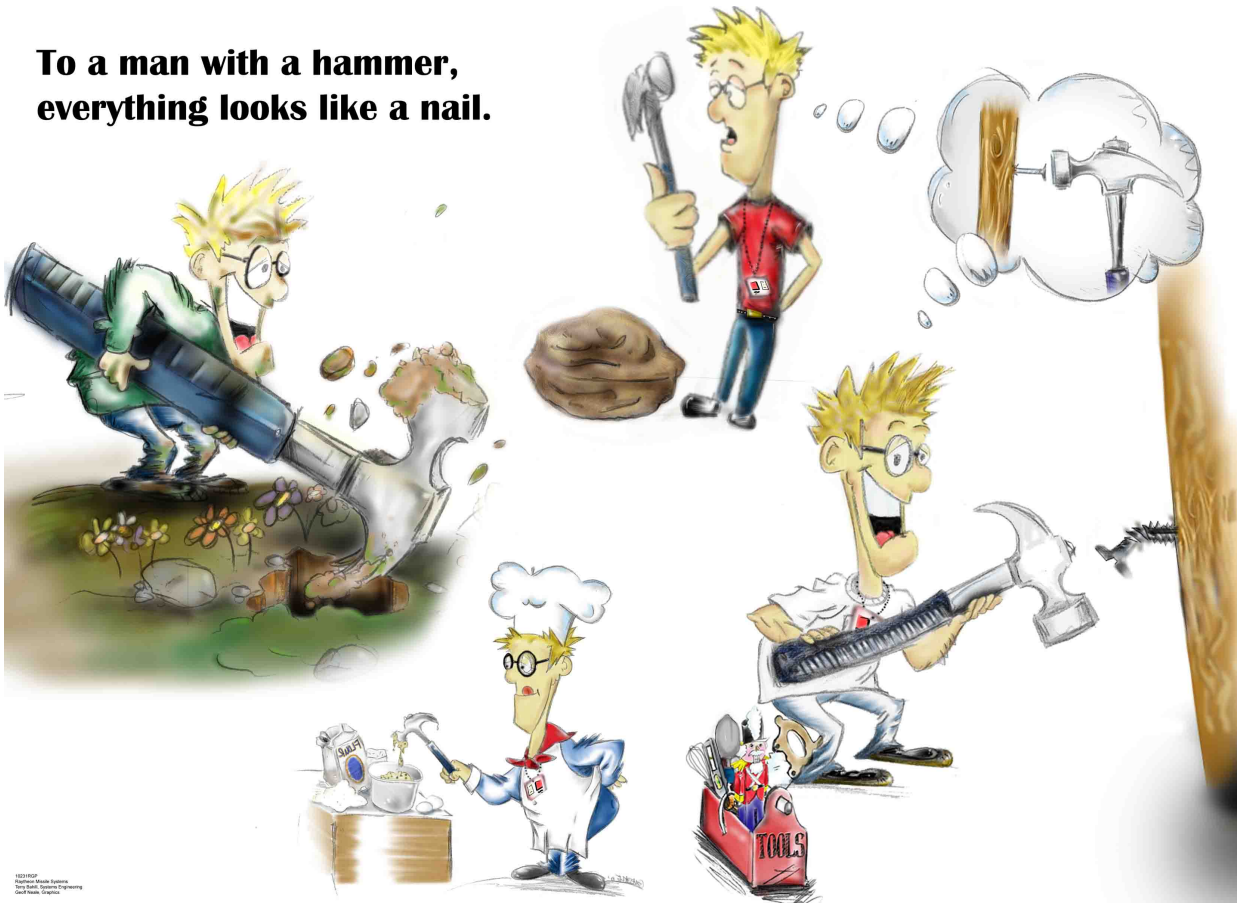
$$Y = \begin{cases} 0 & \text{if } \texttt{Default} \text{ is no;} \\ 1 & \text{if } \texttt{Default} \text{ is yes.} \end{cases}$$

  since the conditional mean is also the conditional probability

$$\Pr\left(Y = 1 \mid X = x\right) = \mathbb{E}\left[Y \mid X = x\right]$$

- However, linear regression fails to restrict $\hat{y}$ to be between $[0, 1]$ in general.
- We could indulge in a discussion on constrained optimisation problem.

- The real problem is not we don't have a more sophisticated hammer,



it is we don't have a nail to start with.

- For a categorical response variable Y, the process of predicting it is known as

  classification

- It often involves modelling the probability of each of the categories using

$$X_{i1}, \quad X_{i2}, \quad \ldots \quad X_{ij}, \quad \ldots \quad X_{ik} \qquad i = 1, 2, \ldots, n$$

  and

$$Y_i \qquad i = 1, 2, \ldots, n$$

- Classification occur often, perhaps even more so than linear regression. e.g.

1. A bank must be able to determine instantaneously whether a transaction is fraudulent, on the basis of the user's IP address, past transaction history, etc.
2. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three illnesses. Which of the three illnesses does the individual have?

- To avoid having an estimated probability outside of $[0, 1]$, we model

$$\Pr(Y = 1 \mid X = x)$$

using a function that has the range $[0, 1]$ instead of using

$$\Pr(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x$$

- There are many functions that meet this requirement, if the logistic function

$$\Pr(Y = 1 \mid X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

is used, then we will end up with so-called logistic regression.

Q: Can you see why the logistic function is a natural choice?

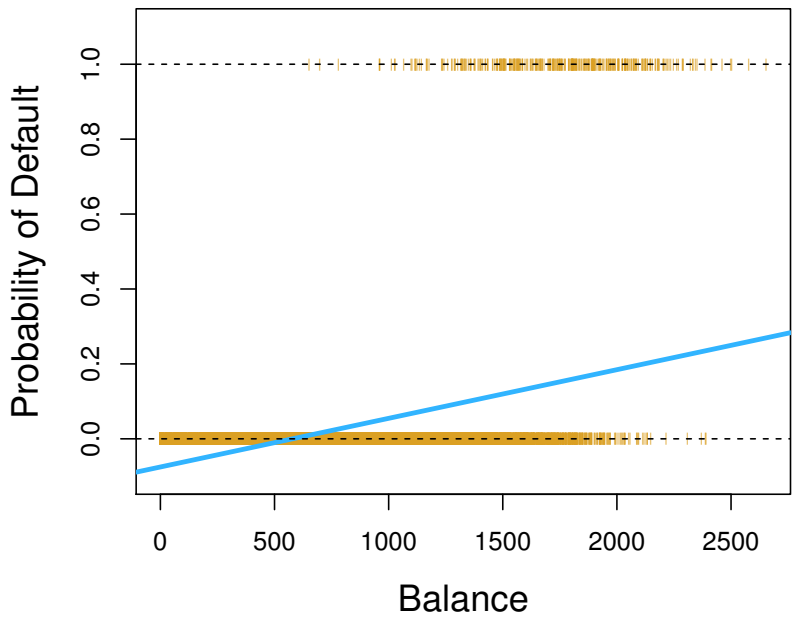
Q: Can you see why logistics regression is a generalised liner regression?

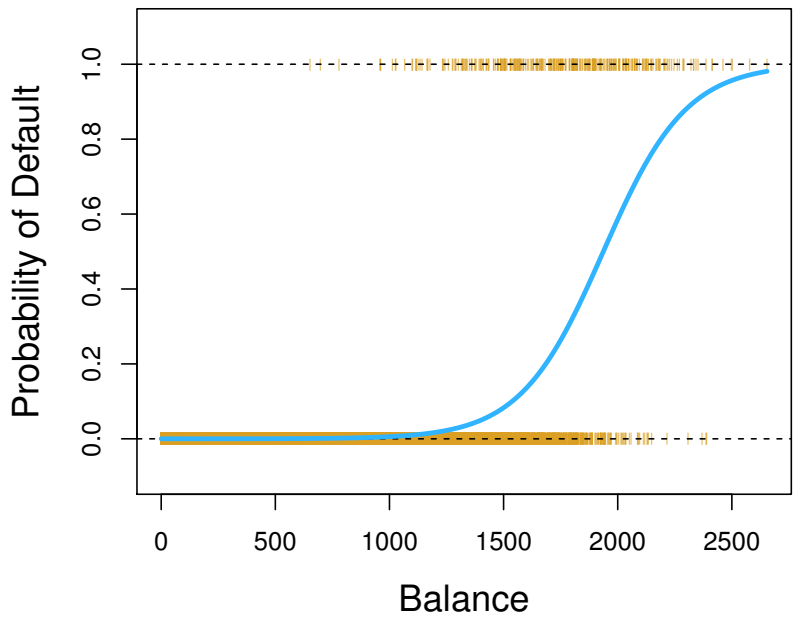
- Once this parametric form is decided, β_0 and β_1 can be estimated using the

principle of maximum likelihood

$$\hat{\boldsymbol{\beta}} \in \left\{ \arg\max_{\boldsymbol{\beta}} \mathcal{L}\left(\boldsymbol{\beta}; x, y\right) \right\}$$

- Using the maximum likelihood estimate (MLE) of β_0 and β_1, we will have





Q: How to obtain MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$? What is the likelihood function here?

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1}^{n} \Pr(Y = y_i)$$

- Under the assumption that

$$\Pr(Y = 1 \mid X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

the log-likelihood function is given by

$$\begin{aligned}
\ell(\beta_0, \beta_1) &= \log \mathcal{L}(\beta_0, \beta_1) \\
&= \log \prod_{i=1}^{n} \big(\Pr(Y = 1 \mid X = x_i)\big)^{y_i} \big(\Pr(Y = 0 \mid X = x_i)\big)^{1-y_i} \\
&= \sum_{i=1}^{n} \left[-\log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) + y_i(\beta_0 + \beta_1 x_i)\right]
\end{aligned}$$

- We obtain two nonlinear equations when setting the first derivatives to zero

$$\sum_{i=1}^{n}\left(y_i - \frac{\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) = 0$$

$$\sum_{i=1}^{n} x_i\left(y_i - \frac{x_i\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) = 0$$

- Of course, R will automatically fit and solve the nonlinear equations for us

	Coefficient	Std.error	Z-statistics	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- The outputs are similar to the linear regression output, .e.g.

$$H_0 \colon \beta_1 = 0$$

is tested by asymptotic Z-test, and we have strong evidence against $\beta_1 = 0$.

- Of course, we base our prediction on our estimates, .e.g.

$$\hat{\Pr}\left(Y = 1 \mid X = 1000\right) = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)}$$

$$= \frac{\exp\left(-10.6513 + 0.0055 \cdot 1000\right)}{1 + \exp\left(-10.6513 + 0.0055 \cdot 1000\right)}$$

$$= 0.00576$$

- Whether to classify someone with a balance of 1000 to be a high or low risk depends on the bank's preference towards risk.
- Of course, this classifier can be easily extended to more than one predictor

$$\Pr\left(Y = 1 \mid X_1 = x_1, X_2 = x_2\right) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)}$$

Q: How to access the quality of our classifier?

- One simple and common approach is to quantify the accuracy by

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

which is known as training error rate, where I is an indicator variable.

$$I\left(y_i \neq \hat{y}_i\right) = \begin{cases} 0 & \text{if} \quad y_i = \hat{y}_i; \\ 1 & \text{if} \quad y_i \neq \hat{y}_i. \end{cases}$$

- Hence the training error rate is the fraction of incorrect classifications.
- Once we think in terms of this criterion, it is natural to minimise the mean

$$I\left(y_{n+1} = \hat{y}_{n+1}\right)$$

that is, incorrectly classifying on new data, which is known as test error rate.

- It can be shown the test error rate is minimised if we simply assign a test observation with predictor $\mathbf{x}_{n+1}$ to the most likely class j, that is,

$$\arg\max_{j \in \{0,1\}} \left(\Pr\left(Y_{n+1} = j \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) \right)$$

- This simple classifier is known as the Bayes classifier, which achieves the lowest possible test error rate, known as the Bayes error rate

$$1 - \mathbb{E}\left[\max_{j \in \{0,1\}} \left(\Pr\left(Y_{n+1} = j \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) \right) \right]$$

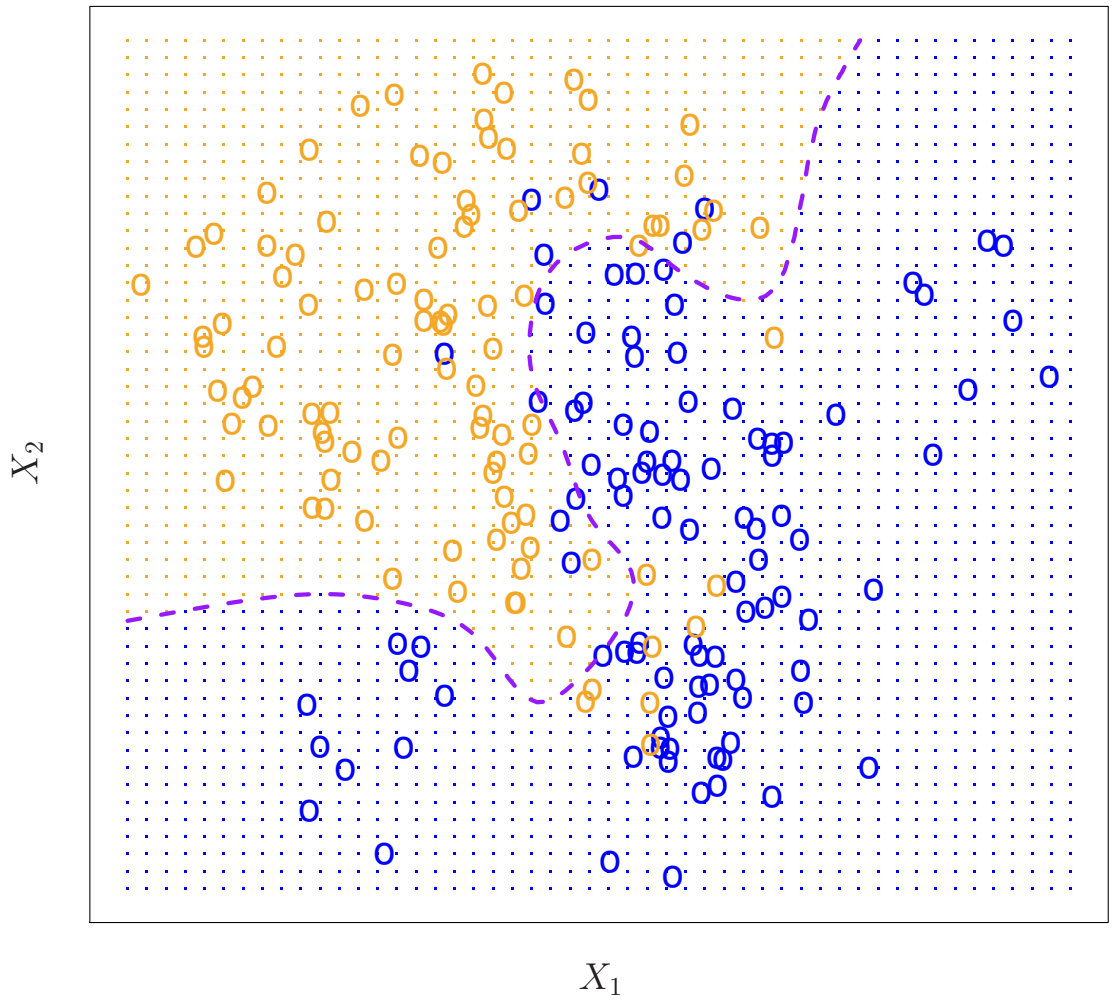
  it is the theoretical error rate that cannot be achieve in practice. Why?

- It can only be done when the conditional distribution is known,

$$\Pr\left(Y = y \mid \mathbf{X} = \mathbf{x}\right)$$

  when we have to estimate it, the extra error will prevent us from attaining this least error rate. However, it serves as the gold standard against which other methods are assessed.

- Consider the following simulation, here the conditional distribution is known



- The color of the points indicates the value that Y takes.

- For real data, we have to estimate the conditional distribution, e.g.

$$\hat{\Pr}\left(Y_{n+1} = j \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_{n+1}} I(y_i = j)$$

  which is known as the K-nearest neighbours (KNN) classifier.

- For a given K and test observation

$$\mathbf{X}_{n+1} = \mathbf{x}_{n+1}$$

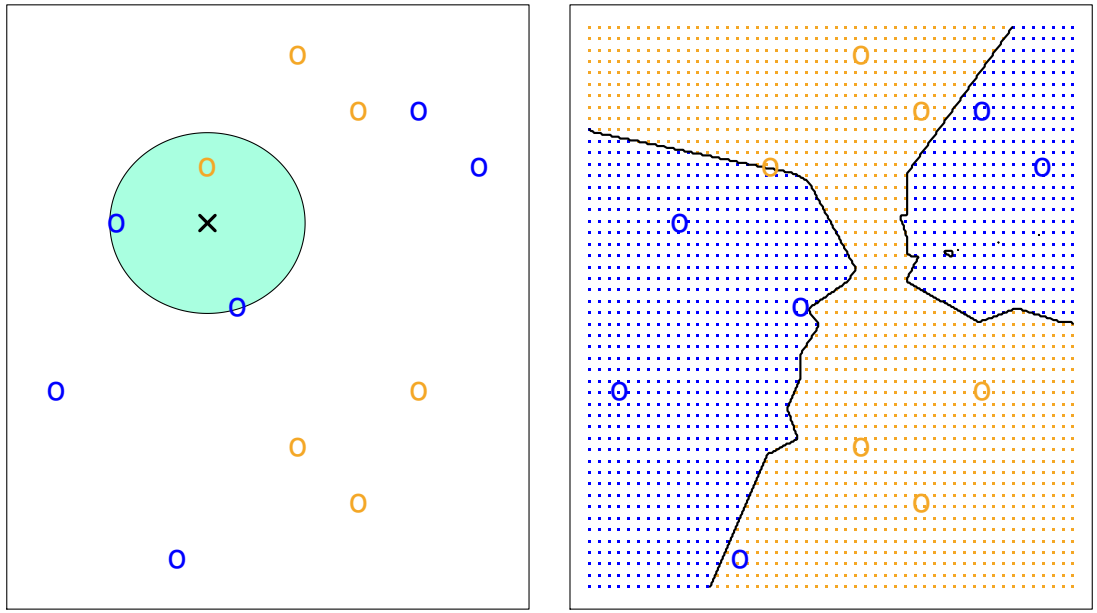
  KNN first identifies K points in the original data that are closest to $\mathbf{x}_{n+1}$

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$$

  this subset of size K is used to estimate the conditional probability

$$\mathcal{N}_{n+1} = \{\ldots, (\mathbf{x}_i, y_i), \ldots\}$$

- Consider the following scenario, and suppose K is chosen to be 3



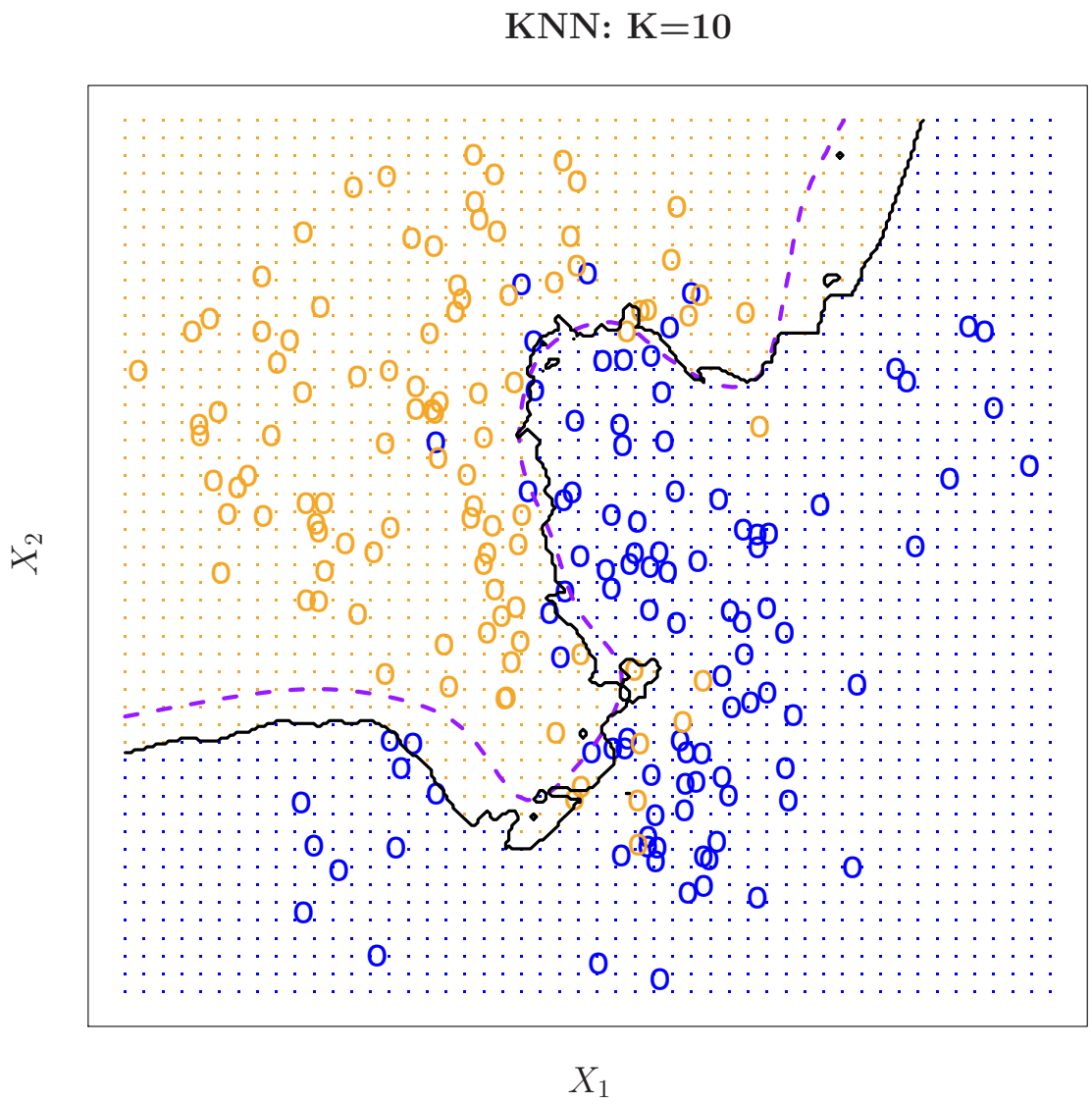
- In this case, the estimated conditional probability is

$$\hat{\Pr}\left(Y_{n+1} = \text{Blue} \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) = \frac{2}{3}$$

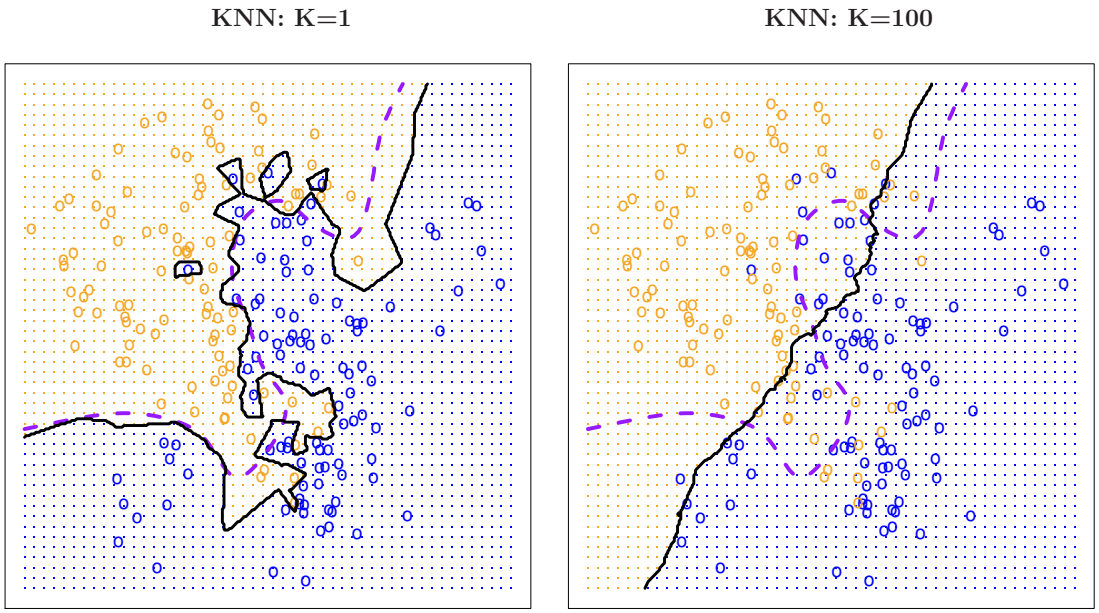
$$\hat{\Pr}\left(Y_{n+1} = \text{Orange} \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) = \frac{1}{3}$$

- Once $\hat{\Pr}$ is available, KNN would classify according to Bayes rule, thus Blue.

- Using KNN with $K = 10$ for the simulated data before

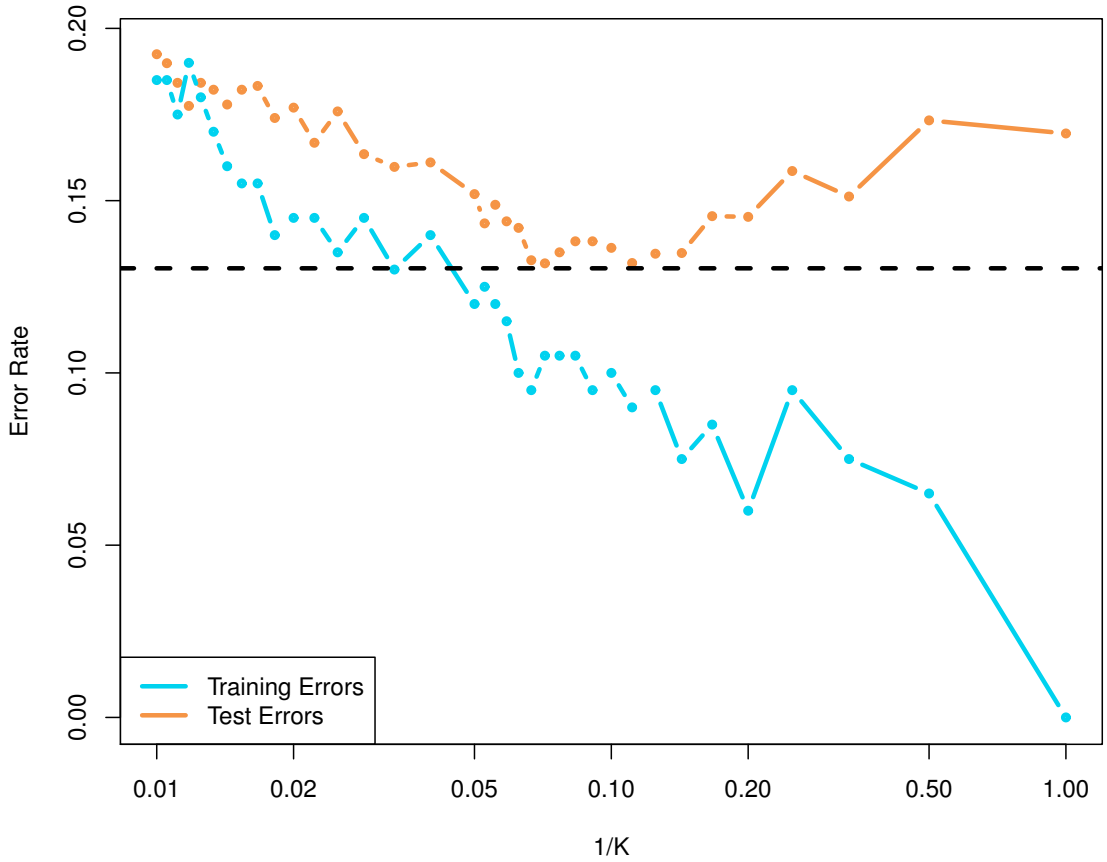


- The choice of K has a drastic effect on the result



- As K grows, the method becomes less flexible and gives a decision boundary that is close to linear.
- The quality of the classification is assessed by looking at the test error rate and the training error rate. Notice the training error rate always declines when the method becomes more flexible, but the test error rate is U-shaped.

- Just like traditional regression, the level of flexibility is crucial to the method.



- The extreme ends are to be avoid due the U-shape the test error rate has.
- Optimal flexibility depends on the data size and varies from cases to cases.