# Introduction to Data Science

DSA1101

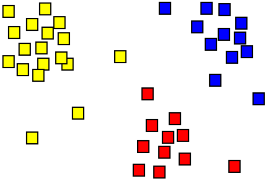
Semester 1, 2018/2019
Week 4

# Classification Methods

# Classification Methods

- Over the past few lectures, we have touched upon the $k$-means algorithm as an example of *unsupervised learning* method.
- *Unsupervised learning* is the task of inferring hidden structure based on data without the outcome $y$.
- We call such data without $y$ 'unlabeled' data.

# Classification Methods



- The $k$-means algorithm allows us to partition observations in a unlabeled data set into distinct groups so that:

(i) the observations within each group are quite similar to each other,

(ii) and observations in different groups are quite different from each other.

# Back to *supervised learning* methods

- In data science, many applications involve making predictions about the outcome $y$ based on a number of predictors $x$
- Often we assume models of the form

$$y = f(x)$$

  where $f(x)$ is a function that maps the predictor(s) to the outcome.
- In many cases, the outcome $y$ is a *categorical* variable or class membership.
  e.g: CS/CU
- We will talk about the *k-nearest neighbor* classification, a popular *supervised learning* method for class membership prediction in data science.

# Classification Methods

| **Supervised** | **Unsupervised** |
|---|---|
| Linear regression | $k$-means |
| Decision trees | Association rules |
| $k$-nearest neighbor | Hierarchical clustering |
| Linear discriminant analysis | Deep belief nets |
| Naive Bayes | Self-organizing map |

# Example: Anti-spam techniques


- Based on an e-mail's content, e-mail providers use classification methods to decide whether the incoming e-mail messages are spam.

# Example: Anti-spam techniques


- Based on features such as presence of certain keywords and images ($X$), classification methods assign a given email to the "spam" or "non-spam" class ($y$).
- Here the outcome $y$ is a class membership, with only two classes.

# Example: Automated medical diagnosis

Source: The Straits Times

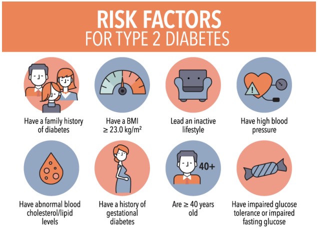
- Automated medical diagnostic methods can help with preventive screening campaigns and allow medical professionals to focus on at-risk individuals.

# Example: Automated medical diagnosis

Source: The Straits Times

- Based on clinical features such as gender, blood pressure, and presence or absence of certain symptoms ($x$), classification methods can predict whether a person has a disease or not ($y$).

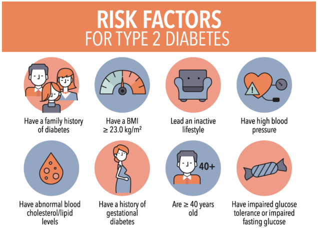
# Example: Singapore's war against diabetes




Source: `https://www.gov.sg/factually/content/can-you-develop-diabetes`

"In setting the battle scene, Health Minister Gan Kim Yong said the disease is already costing the country more than $1 billion a year. Of the more than 400,000 diabetics today, one in three do not even know they have the disease."
- *The Straits Times*, April 13, 2016

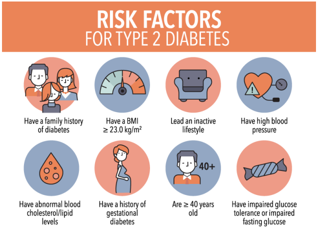
# Example: Singapore's war against diabetes




Source: `https://www.gov.sg/factually/content/can-you-develop-diabetes`

- Based on features such as gender, body mass index and lifestyle choices ($x$), an online Diabetes Risk Assessment (DRA) is developed to predict whether a person is at risk to develop diabetes or not ($y$).

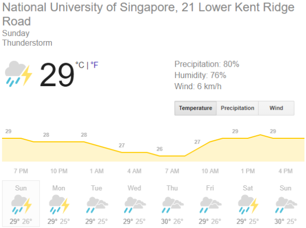
# Example: Singapore's war against diabetes




Source: `https://www.gov.sg/factually/content/can-you-develop-diabetes`

- The DRA is available at `https://www.healthhub.sg/programmes/DRA?utm_source=GovsgFactually`
- This is another example of classification technique from data science being implemented in practice and helping Singaporeans.

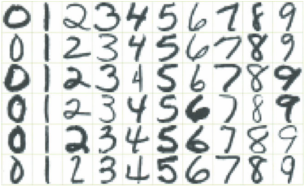
# Example: Weather forecast




Source: Google

- Predicting whether it will rain or not ($y$) based on local conditions such as humidity and temperature ($x$)

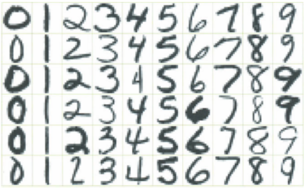
# Example: Automated handwritting recognition



Examples of handwritten digits from U.S. postal envelopes. Source: *The Elements of Statistical Learning*, Hastie et al.

- Algorithms for handwritten number recognition are important for tasks such automatic sorting procedures for postal mails and automatic check deposit systems.

# Example: Automated handwritting recognition



Examples of handwritten digits from U.S. postal envelopes. Source: *The Elements of Statistical Learning*, Hastie et al.

- The task is to predict, from the image matrix of pixel intensities ($x$), the identity of each image ($y$) quickly and accurately.
- Here the outcome $y$ takes on multiple categories $(0, 1, ..., 9)$.

# Example: finance


- Instant loan approvals online offered by banks
- Based on a loan applicant's credit history and the details on the loan ($x$), the loan can be approved or denied ($y$).

# Example: marketing


Source: The Straits Times

- Predict whether a wireless customer want to re-contract or not ($y$) based on age, number of family members on the plan, months remaining on the existing contract, and social network contacts ($x$).
- With such insight, target the customers with appropriate offers.

# $k$-nearest neighbor classification

- $k$-nearest neighbor classification involve making predictions about a categorical outcome $y$ based on a number of predictors $x$
- An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its $k$ nearest neighbors ($k$ is a positive integer, typically small).
- To determine the $k$ nearest neighbors, we will use the Euclidean distance in the feature space ($x$).
- We will illustrate how $k$-nearest neighbor classification works with a simple example shortly.

# $k$-nearest neighbor classification: simple example

- We will learn the $k$-nearest neighbor classification algorithm with a simple, hypothetical setting involving email spam detection
- Suppose we have two features ($x$) to predict whether an email is spam or not:

$x_1$: The number of occurrences of the phrase "you are a winner"

$x_2$: The number of images contained in an email

- The task is to predict whether an email is spam or not ($y$) based on $x = (x_1, x_2)$

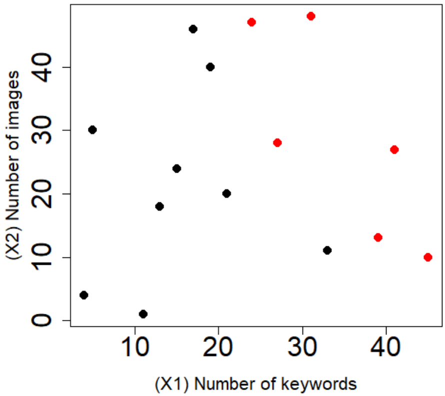
# $k$-nearest neighbor classification: simple example

- Recall that often we assume models of the form

$$y = f(x)$$

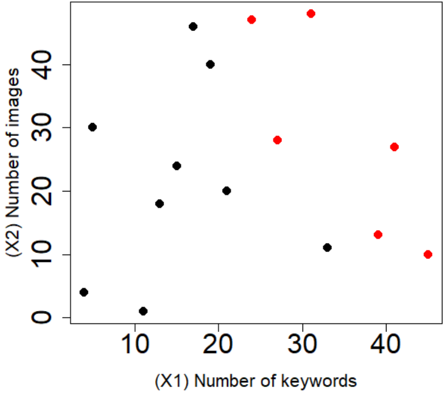
  where $f(x)$ is a function that maps the predictor(s) to the outcome.
- The task is to make predictions about the outcome $y$ based on a number of predictors $x$
- In this example, $x$ refers to the two email features, and $y$ is whether the email is spam $(y = 1)$ or not $(y = 0)$
- Since our training data contains both the features $x$ and their corresponding labels $y$, $k$-nearest neighbor classification is an example of *supervised learning*.

# $k$-nearest neighbor classification: simple example




- Since there are only two features $x_1$ and $x_2$, we can plot the data points on a 2-D graph
- Each point is labelled as spam (red, $y = 1$) or non-spam (black, $y = 0$)

# $k$-nearest neighbor classification: simple example




- Note that because the data is labelled, the classification is *supervised*
- Our task is to predict whether a new, incoming email is spam or not, based on its $(x_1, x_2)$

# $k$-nearest neighbor classification

- An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its $k$ nearest neighbors ($k$ is a positive integer, typically small).
- To determine the $k$ nearest neighbors, we will use the Euclidean distance in the feature space ($x$).
- Recall that for two given data points in $p$-dimensional feature space, $z_i$ at $(x(1)_i, x(2)_i, ..., x(p)_i)$ and $z_j$ at $(x(1)_j, x(2)_j, ..., x(p)_j)$, the Euclidean distance between $z_i$ and $z_j$ is

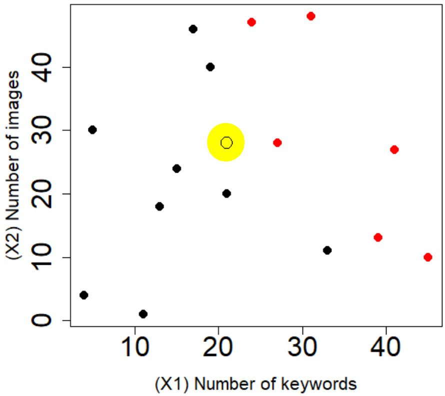
$$dist(z_i, z_j) = \sqrt{\sum_{l=1}^{p} (x(l)_i - x(l)_j)^2}$$

# $k$-nearest neighbor classification

- In the 2-dimensional feature space for our example, the Euclidean distance between $z_i$ and $z_j$ is

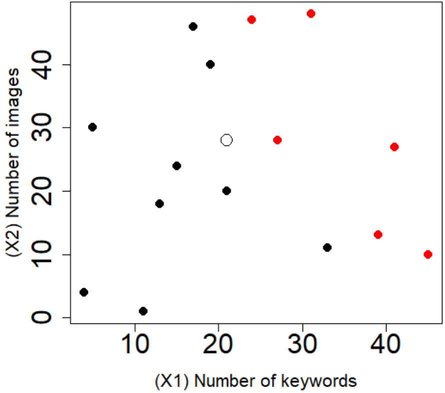
$$dist(z_i, z_j) = \sqrt{(x(1)_i - x(1)_j)^2 + (x(2)_i - x(2)_j)^2}$$

# $k$-nearest neighbor classification: simple example



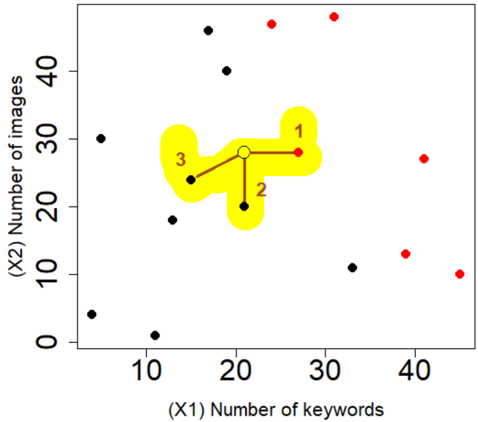

- Assume now that we want to predict whether a new, incoming email with 21 occurrences of the phrase "you are a winner" and 28 attached images, i.e. $(x(1), x(2)) = (21, 28)$

# $k$-nearest neighbor classification: simple example




- We set $k = 3$, so we need to find the three nearest neighbors to the data point (represented by the circle) in the feature space in terms of Euclidean distance

# $k$-nearest neighbor classification: simple example




Euclidean distance from the circle to

- the first point:

$$\sqrt{(21-27)^2+(28-28)^2)}$$

$$=6$$

- the second point:

$$\sqrt{(21-21)^2+(28-20)^2)}$$

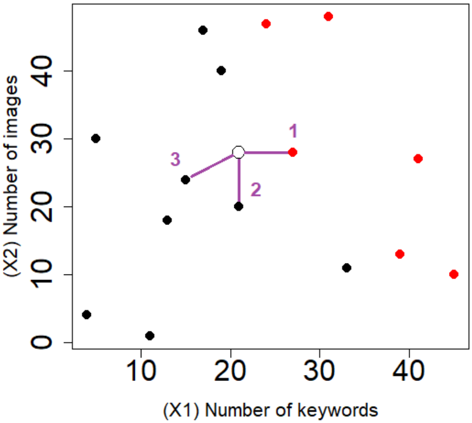
$$=8$$

- the third point:

$$\sqrt{(21-15)^2+(28-24)^2)}$$

$$\approx 7.2$$

# $k$-nearest neighbor classification: simple example




- These three data points are the closest to the circle in terms of Euclidean distance in the 2-dimensional feature space

# $k$-nearest neighbor classification: simple example

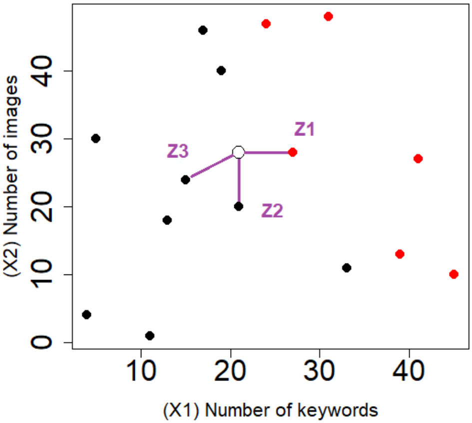
- When $k = 3$, the fitted outcome value for a new data point with feature values $x$ is

$$\hat{Y}(x) = \frac{1}{3} \sum_{z_i \in N_3(x)} y_i$$

i.e the average y of three nearest neighbors

- $N_3(x)$ is the neighborhood of $x$ defined by the 3 closest points $z_i$ in the training sample

# $k$-nearest neighbor classification: simple example




- For our example, the points $z_1$, $z_2$ and $z_3$ are the closest points in terms of Euclidean distance to the circle
- The corresponding membership values are $y_1 = 1$, $y_2 = 0$ and $y_3 = 0$

# $k$-nearest neighbor classification: simple example

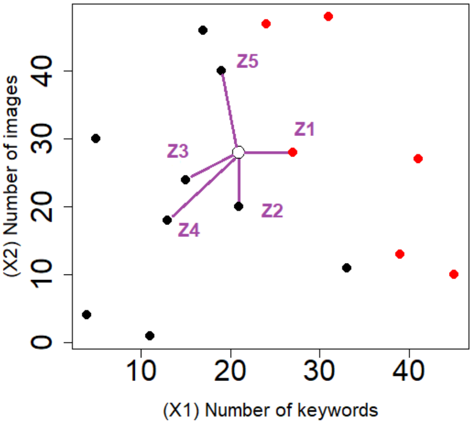
- Therefore the predicted class membership for the new data point with feature values $x^* = (x(1), x(2)) = (21, 28)$ based on $k = 3$ nearest neighbors is

$$\hat{Y}(x^*) = \frac{1}{3} \sum_{z_i \in N_3(x^*)} y_i = \frac{1}{3}(1 + 0 + 0) = \frac{1}{3}$$

# $k$-nearest neighbor classification: simple example

- There are many ways to predict class membership based on the fitted $\hat{Y}$
- One popular way is by majority vote, i.e. predict the email as spam if $\hat{Y} > 0.5$
- Therefore, for our circled data point with feature values $x^*$, since $\hat{Y}(x^*) = \frac{1}{3} < 0.5$, we predict it as non-spam

# $k$-nearest neighbor classification: simple example




- Now if we set $k = 5$, then the points $z_1$, $z_2$, $z_3$, $z_4$ and $z_5$ are the closest points in terms of Euclidean distance to the circle
- The corresponding membership values are $y_1 = 1$, $y_2 = 0$, $y_3 = 0$, $y_4 = 0$ and $y_5 = 0$

# $k$-nearest neighbor classification: simple example

- Therefore the predicted class membership for the new data point with feature values $x^* = (x(1), x(2)) = (21, 28)$ based on $k = 5$ nearest neighbors is

$$\hat{Y}(x^*) = \frac{1}{5} \sum_{z_i \in N_5(x^*)} y_i = \frac{1}{5}(1 + 0 + 0 + 0 + 0) = \frac{1}{5}$$

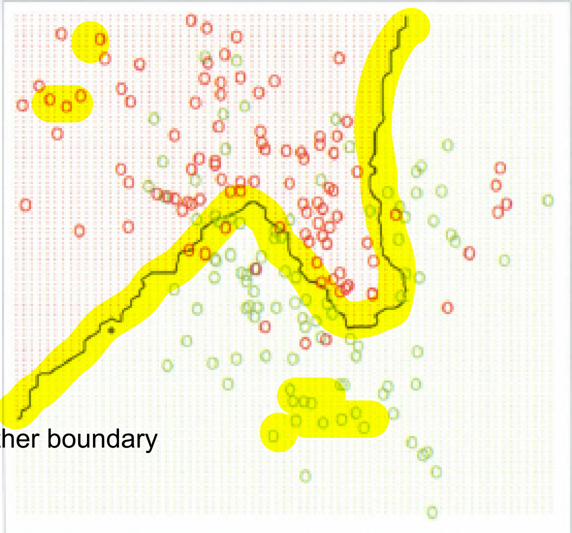
- By majority vote, for our circled data point with feature values $x^* = (x(1), x(2)) = (21, 28)$, since $\hat{Y}(x^*) = \frac{1}{5} < 0.5$, we predict it as non-spam

# $k$-nearest neighbor classification: simple example

- In general, the fitted $\hat{Y}$ with $k$ nearest neighbors for a new data point with feature values $x$ is

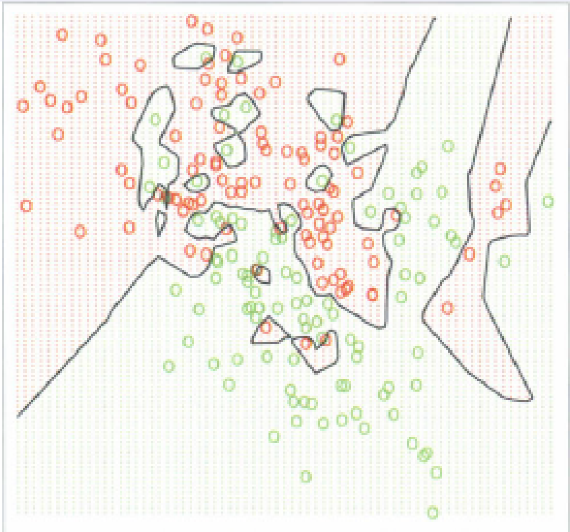
$$\hat{Y}(x) = \frac{1}{k} \sum_{z_i \in N_k(x^*)} y_i$$

# $k$-nearest neighbor classification: a few more examples




Prediction by majority vote with 15 nearest neighbors. Source: *The Elements of Statistical Learning*, Hastie et al.

# $k$-nearest neighbor classification: a few more examples



Prediction by majority vote with one nearest neighbor. Source: *The Elements of Statistical Learning*, Hastie et al.

# $k$-nearest neighbor classification: simple example

- In general, the prediction error for a model can be decomposed into

$$\text{error} = \text{bias}^2 + \text{variance} + \text{ irreducible error}$$

- Notice that in our example, for the larger value of $k = 5$, we take the average of five $y$ values as our fitted value
- So the "variance" of our fitted value $\hat{Y}$ is smaller than when $k = 3$
- However, when $k = 5$, we are also
For larger k, we are
taking data points further away from the circle to compute our fitted value. This may lead to greater "bias" in our fitted value $\hat{Y}$ compared to when $k = 3$.

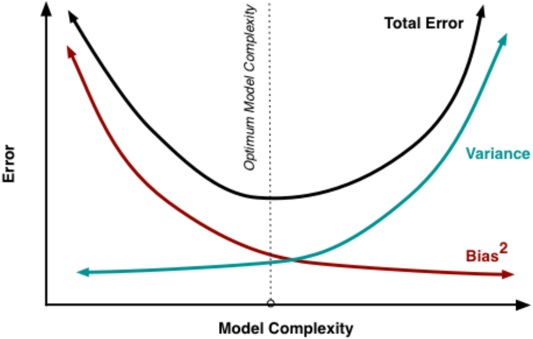
# $k$-nearest neighbor classification: simple example

i.e more data points

- So when $k$ increases, the variance decreases, but bias increases
- This is known as the *bias-variance tradeoff* and is a general property of predictive models

  inverse relation between variance and bias

# *k*-nearest neighbor classification




value of k decreases with Model Complexity***

<<<<<<<<<<<<<<<<k<<<<<<<<<<<<<<<<<

Bias-variance tradeoff. Source: *http://scott.fortmann-roe.com*

larger k -> smoother boundary b/w labels -> MODEL COMPLEXITY reduces