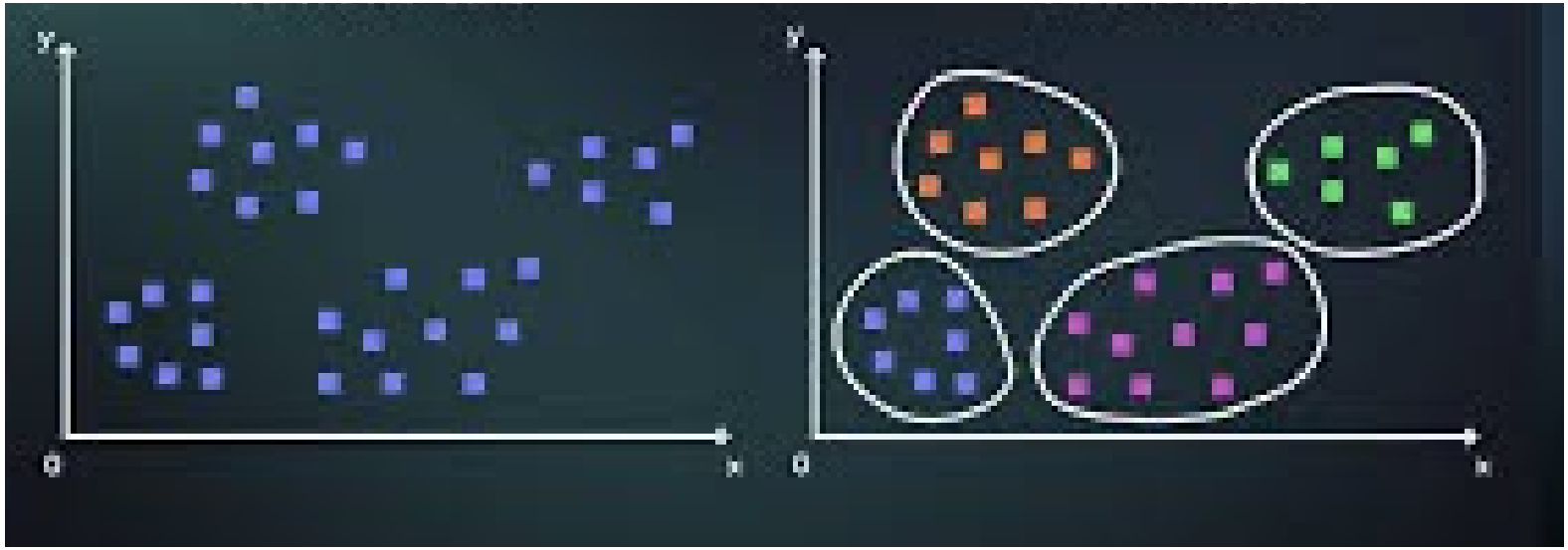
Clustering

Which one of these is not like the others?

This seeming easy task can actually become very difficult



Principal Component Analysis

- Principal component analysis (PCA) is a dimensional reduction technique technique used to
- The goal is to emphasize variation and bring out strong patterns in a dataset
- It helps explore and visualize data

Principal Component Analysis

- A mathematical procedure that transforms a large number of correlated variables into a smaller number of uncorrelated variables called principal components

- The first principal component accounts for as much of the variability in the data as possible

- Each successive component accounts for as much of the remaining variability as possible

Principal Component Analysis

- PCA reduces attribute space from a larger number of variables to a smaller number of factors

- PCA is a dimensionality reduction or data compression method. The goal is dimension reduction and there is no guarantee that the dimensions are interpretable

- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component

Principal Component Analysis

Construct new variables that are linear combinations of the measured traits. This is done to create optimal linear combination of the data to account for the most variation in the original data.

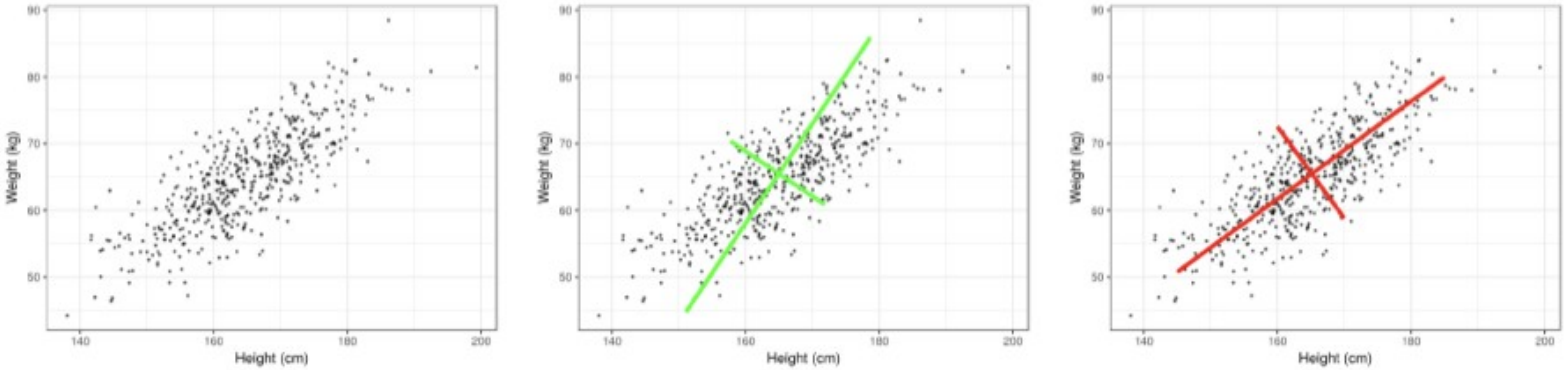


Figure 2.16: Simulated height versus weight for 500 females in the US (left panel) with two orthogonal projections (center and right panels). The right panel shows the optimal linear principal component projection.

$a_1 = w_{11}x_1 + w_{12}x_2$ and $a_2 = w_{21}x_1 + w_{22}x_2$.

Principal Component Analysis

- PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables
- It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance
- This is called the principal axis method and results in orthogonal (uncorrelated) factors

Principal Component Analysis

- Eigenvectors
  - Principal components reflect both common and unique variance of the variables and may be seen as a variance-focused approach seeking to reproduce both the total variable variance with all components and to reproduce the correlations

- The principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular orthogonal dimension

Principal Component Analysis

- Eigenvalues
  - The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.

- The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables

- If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors

Principal Component Analysis

- Eigenvalues measure the amount of variation in the total sample accounted for by each factor

- A factor's eigenvalue may be computed as the sum of its squared factor loadings for all the variables

Principal Component Analysis

- Factor loadings (factor or component coefficients)
  - The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns)

- The squared factor loading is the percent of variance in that variable explained by the factor (like R squared)

- To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables

Principal Component Analysis

- These scores are the scores of each case (row) on each factor (column)

- To compute the factor score for a given case for a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding factor loading of the variable for the given factor, and sums these products

K-means clustering

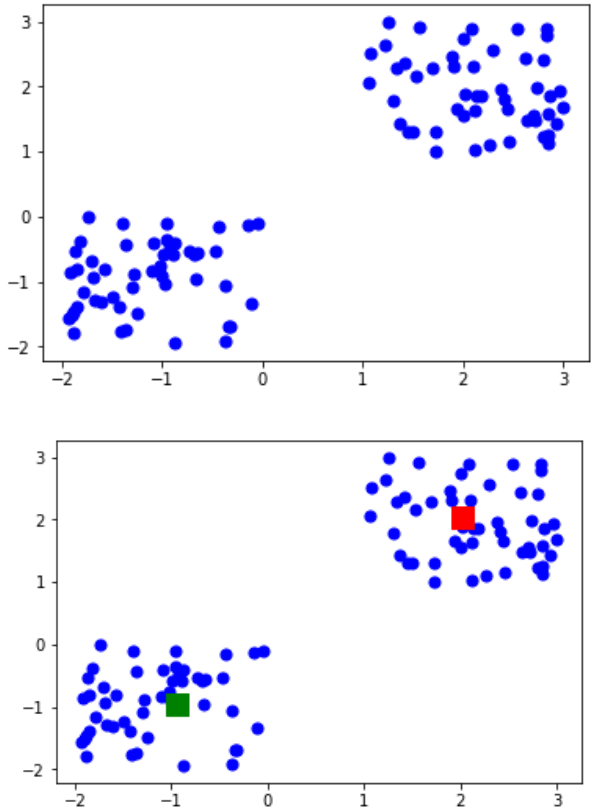
- Group similar data points together and discover underlying patterns.

- K-means looks for a fixed number (k) of clusters in data

- A cluster refers to a collection of data points aggregated together because of certain similarities.

K-means clustering

- Define a target number *k*,
- The number of K refers to the number of centroids you need in the dataset
- A centroid is the imaginary or real location representing the center of the cluster



K-means clustering

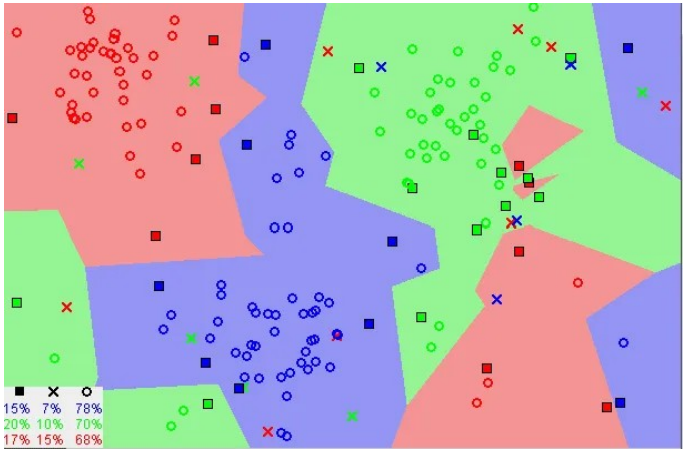
- Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

- In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

- The *'means'* in the K-means refers to averaging of the data
    - finding the centroid.

K-nearest Neighbors

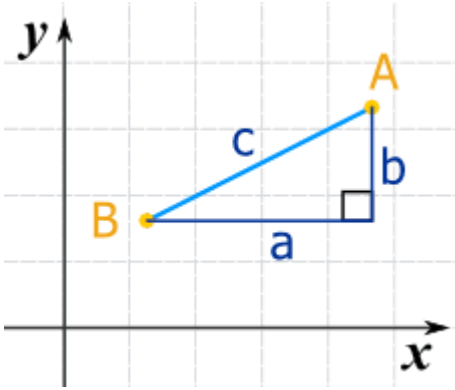
- Supervised learning classifier
- Uses proximity to make classifications
- Predictions about the grouping of an individual data point



KNN

- What is distance between points?

- distance = $\sqrt{a^2 + b^2}$



How does KNN work

- For each point in the data
- Calculate the distance between the query example and the current example from the data.
- Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries
- If regression, return the mean of the K labels
- If classification, return the mode of the K labels

- The right number of K minimizes error

- As we decrease the value of K to 1, our predictions become less stable

- As K increases predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions

Advantages

- The algorithm is simple and easy to implementation
- There’s no need to build a model, tune several parameters, or make additional assumptions
- The algorithm is versatile. It can be used for classification, regression, and search

Disadvantages

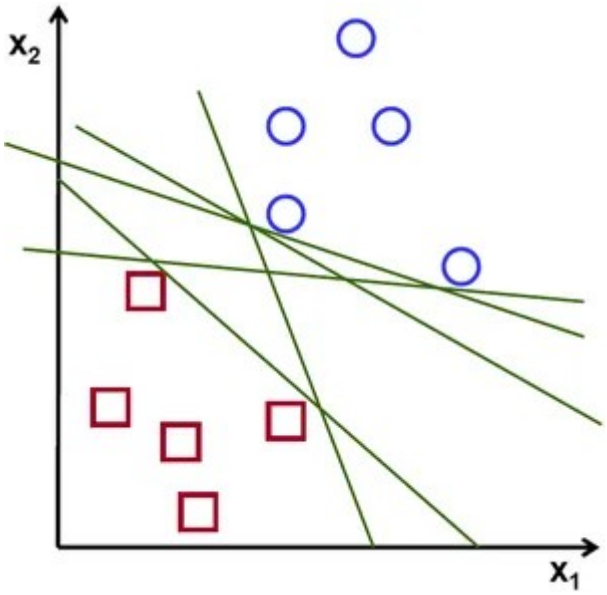
The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

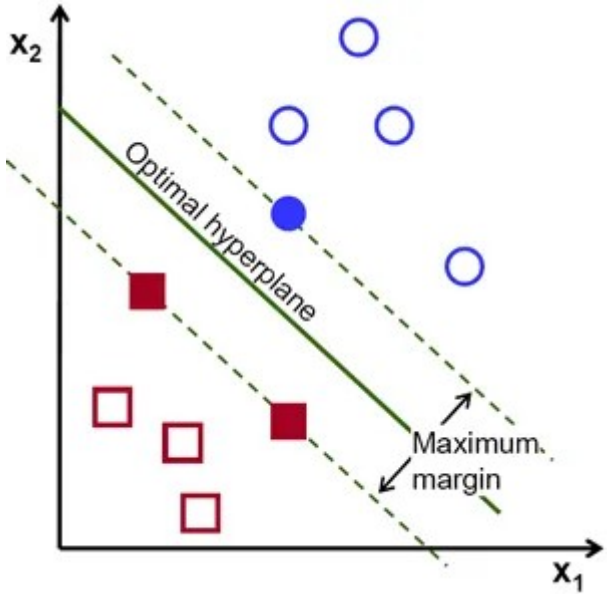
Support Vector Machines

- The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points

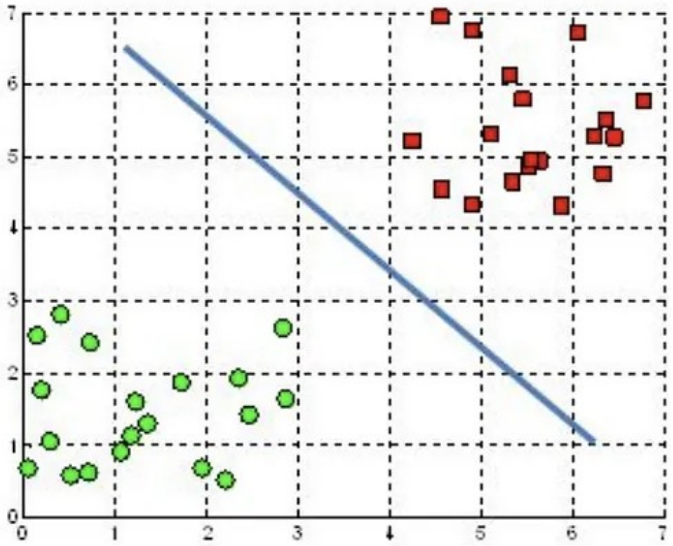
- Hyperplanes are decision boundaries that help classify the data points.

- Data points falling on either side of the hyperplane can be attributed to different classes.

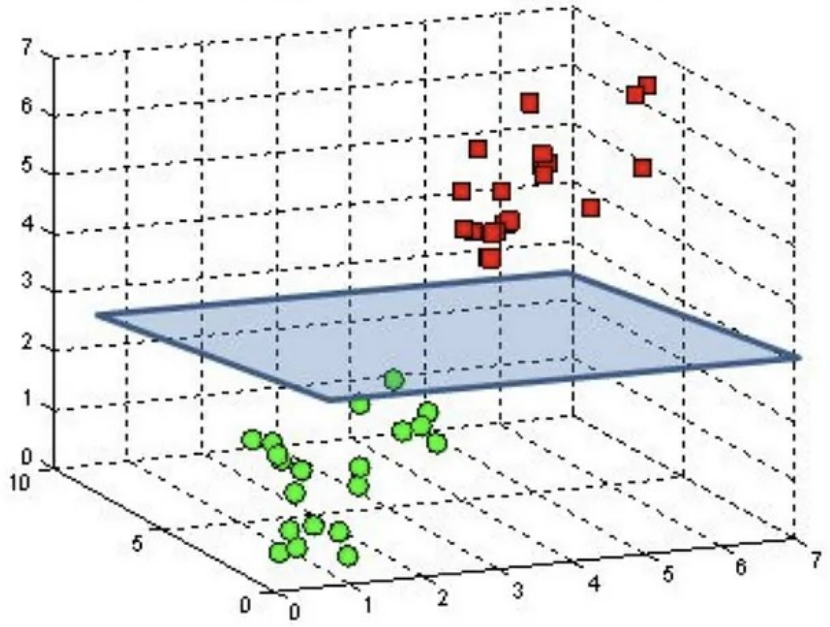

x2
x1


x2
Optimal hyperplane
Maximum
margin
x1

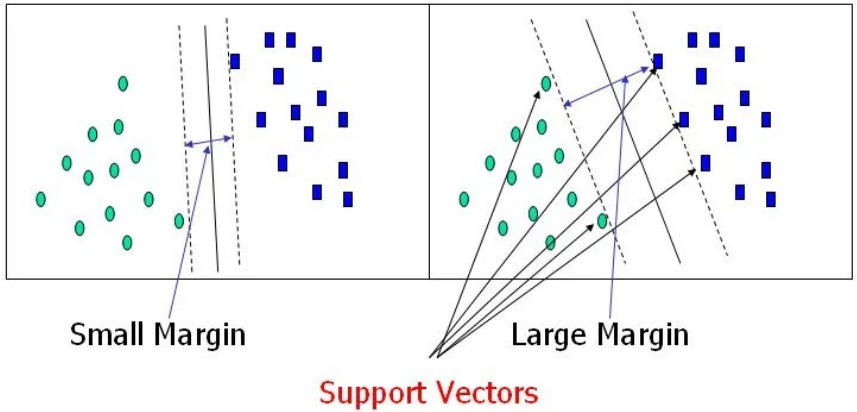
A hyperplane in $\mathbb{R}^2$ is a line

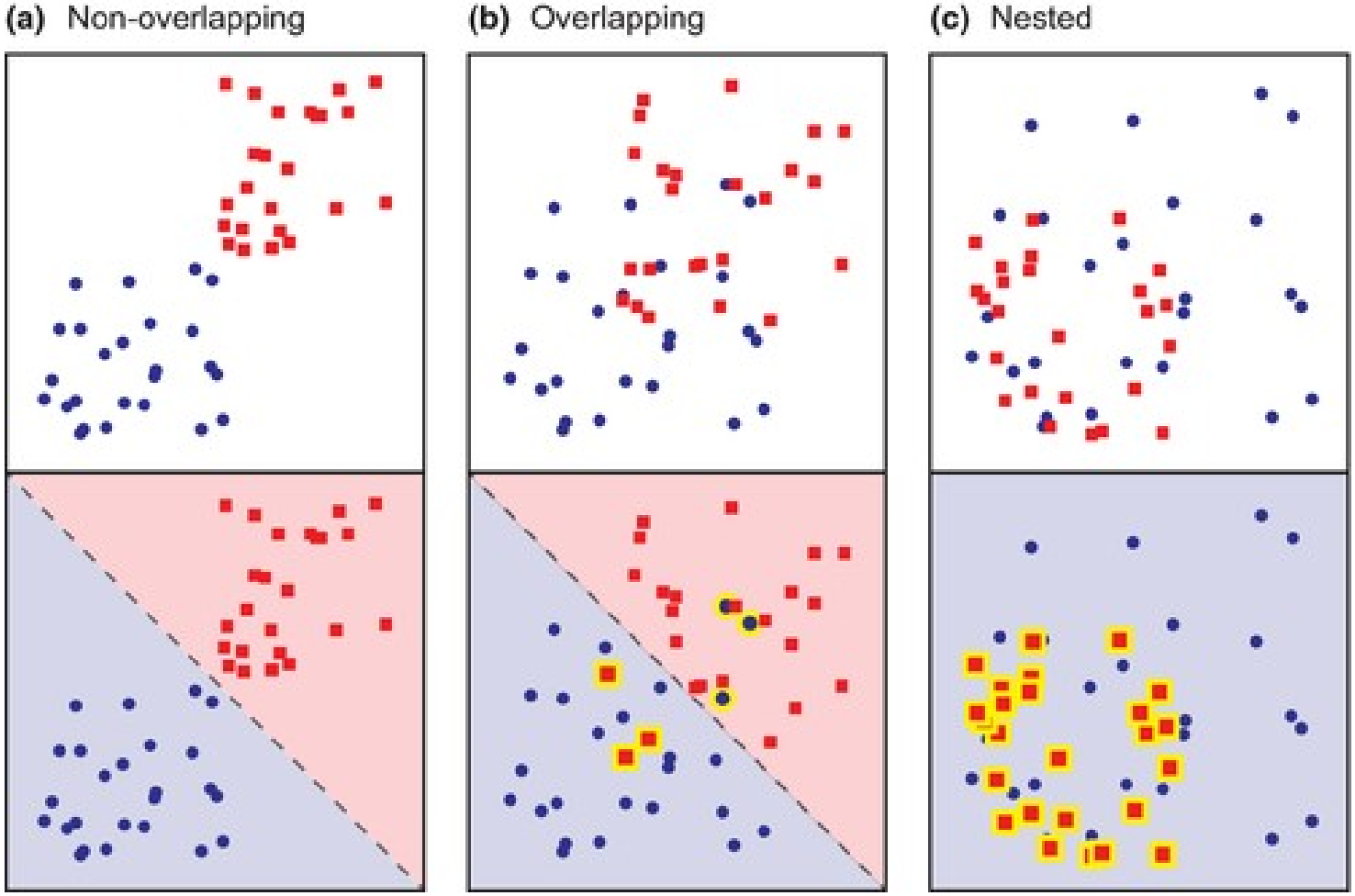


A hyperplane in $\mathbb{R}^3$ is a plane



- Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.




(a) Non-overlapping
(b) Overlapping
(c) Nested

Questions