

Lesson 4

Text Classification Using Support Vector Machines Supervised Machine Learning Method

Support vector machines (SVMs)

- A set of related supervised learning methods (the presence of training data) that analyze data, recognize patterns, classify text, recognize hand-written characters, classify images, as well as bioinformatics and bio sequence analysis.

Data Point in n-dimensional Vector Space

- A vector has in general n components, x_2, x_3, \dots, x_n . A data point represents by (X_1, X_2, \dots, X_n) in n -dimensional space, with metrics and non-metric (feature) variables. Assume for the sake of simplicity, that a vector has two components, X_1 and X_2 (Two sets of words in text analysis).

Data Point in n-dimensional Vector Space

- A bag-of-words represent a set of data-points (subset of words) in text analytics
- Terms (subset of words) represents a text-vector in text analytics

Data Point in 2-dimensional Vector Space

- Assume for the sake of simplicity, that a vector has two components, X_1 and X_2 (Two sets of words in text analysis).

Hyperplane

- A subspace of one dimension less than its ambient space in geometry, hyper plane is just a line
- If a space is 3-dimensional then its hyperplanes are 2-dimensional planes
- Hyperplane may a curved plane represented by a Kernel function

Hyperplane for Classification

- Separates the two classes most appropriately
- Has maximum distance (margin) from closest data points of the distinct classes

Figure 9. 3 Support vectors, separating hyperplane (B) and margins

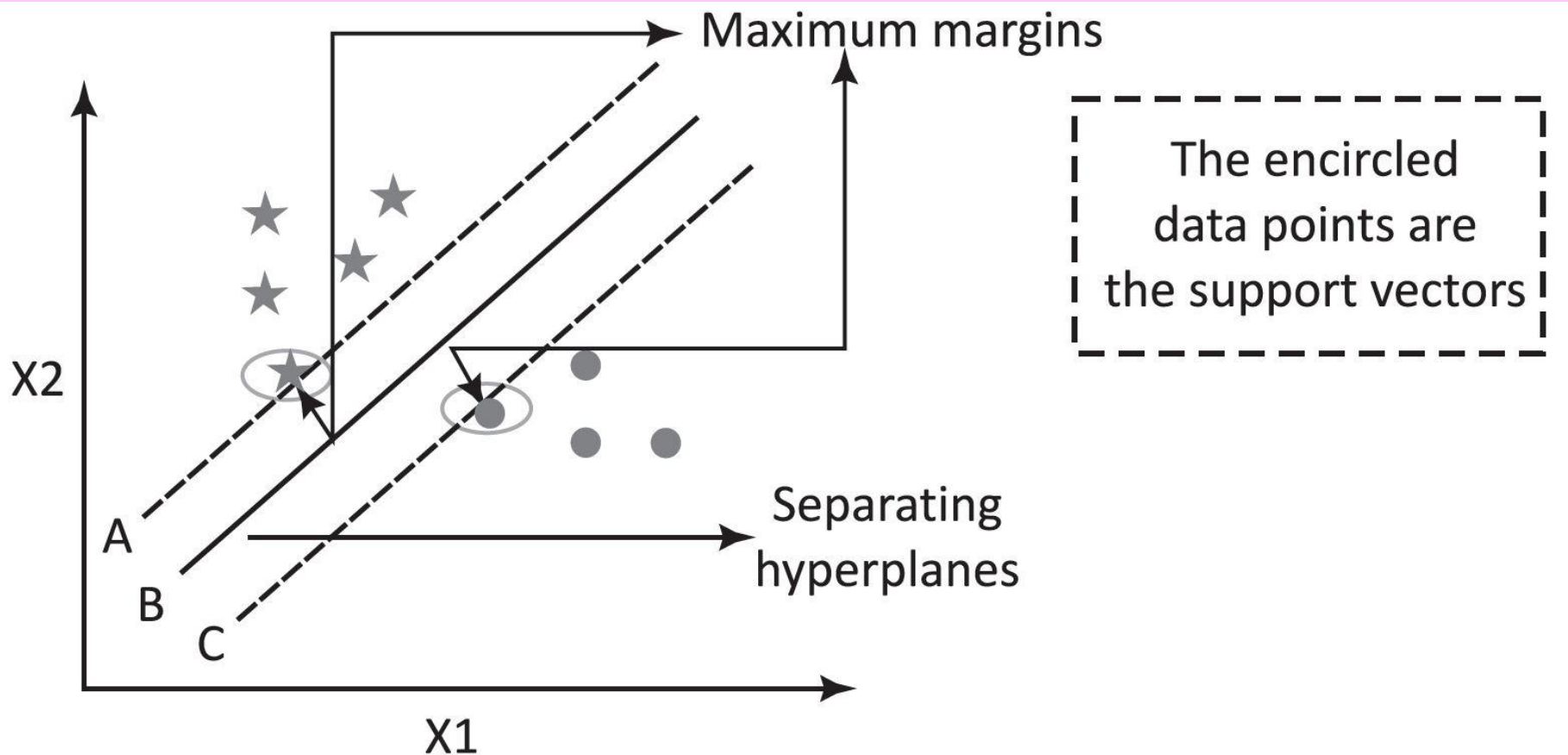


Figure 9. 4 Concept of training set data using support vectors

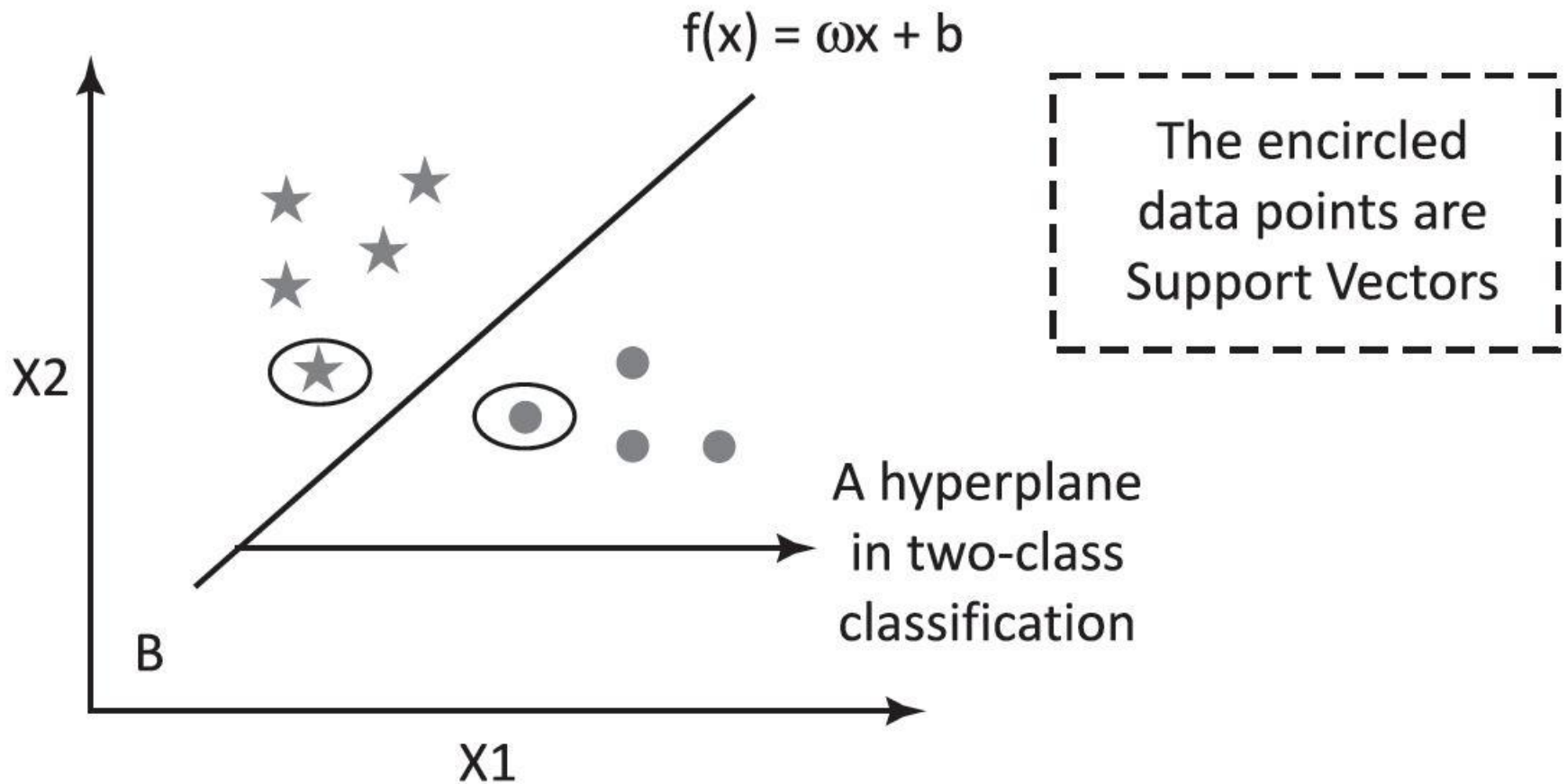
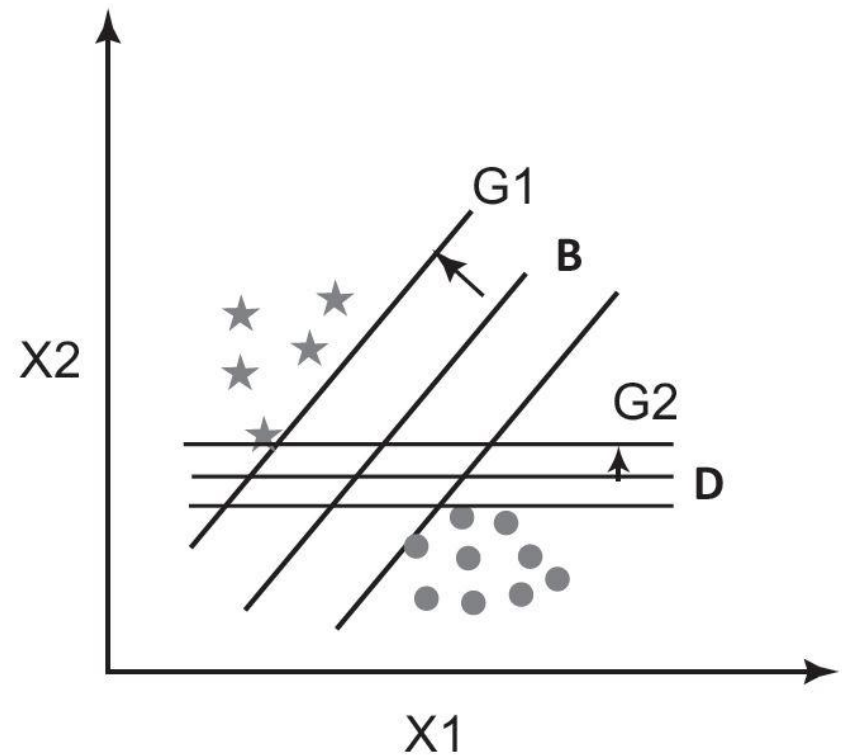
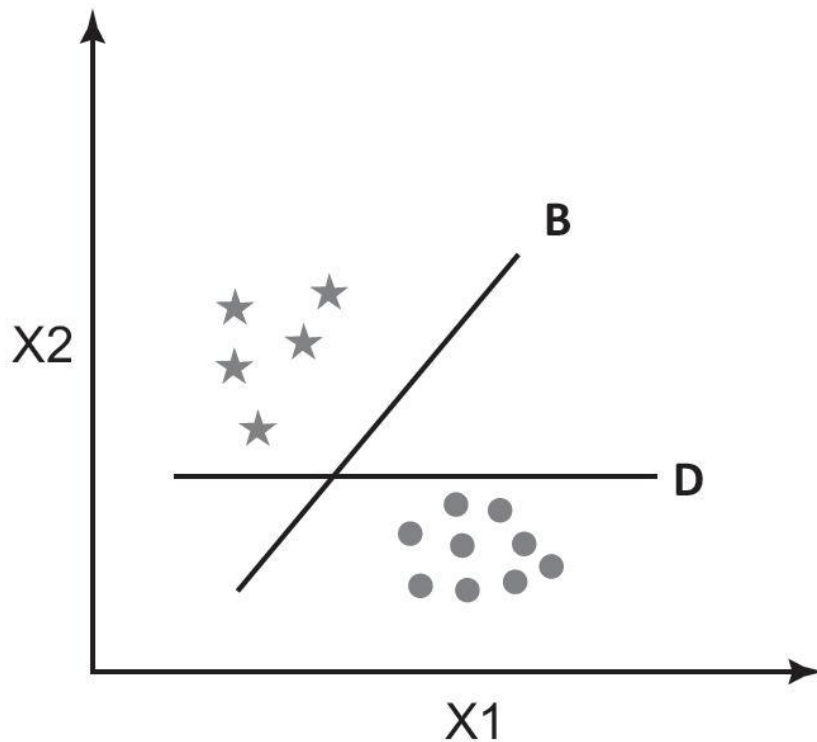


Figure 9.5 Method of selecting the right hyperplane



SVM classifier

- A discriminative classifier formally defined by a separating hyperplane. The concept applies extensively in number of application areas of ML.

Applications of SVMs

1. Classification based on the outputs taking discrete values in a set of possible categories,
 - SVM can be used to separate or predict if something belongs to a particular class or category. SVM helps in finding a decision boundary between two categories.

Applications of SVMs

- 2. Regression analysis, if learning problem has continuous real-valued output (continuous values of x , in place discrete n values, $(x_1, x_2, x_3, \dots, x_n)$)
- 3. Pattern recognition

Applications of SVMs

4. Outliers detection. Assumes that close-by objects are more probable in the same category
5. Finds k objects in the large number of text documents, which have most similar query responses

Summary

We learnt:

- Support vectors in SVMs classify the data points in n-dimensional space
- Hyperplanes discriminate one class of data-points from another
- Hyperplane may a curved plane represented by a Kernel function

End of Lesson 4 on

Text Classification Using Support
Vector Machines Supervised
Machine Learning Method