Document Similarity Measure

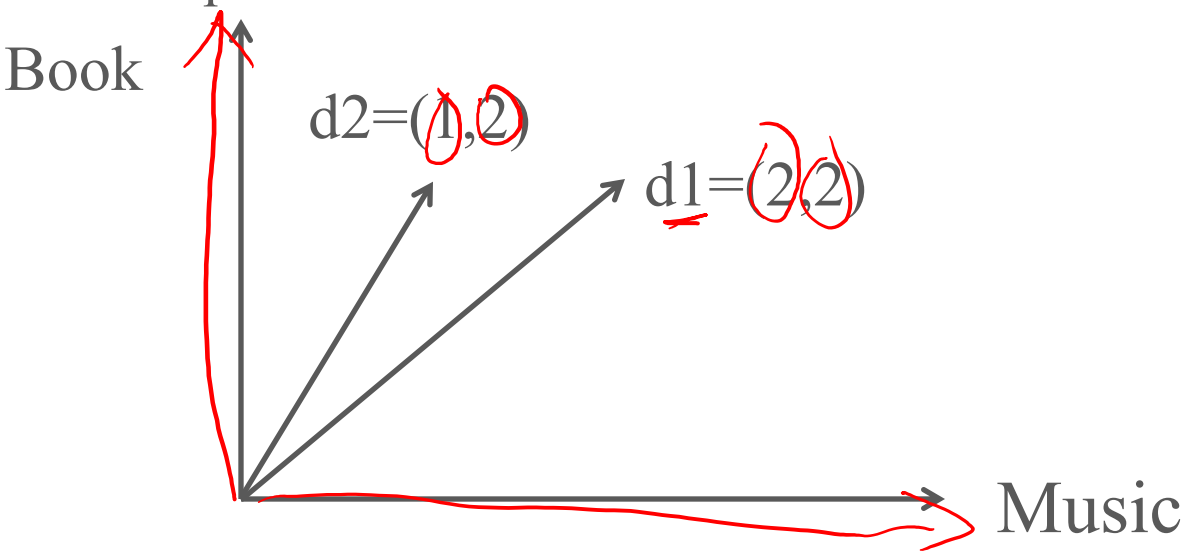
School of Information Studies
Syracuse University

Vector Representation

Bag-of-Words documents

- D1: "book, book, music, music"
- D2: "music, book, book"

Vectors in 2D-space



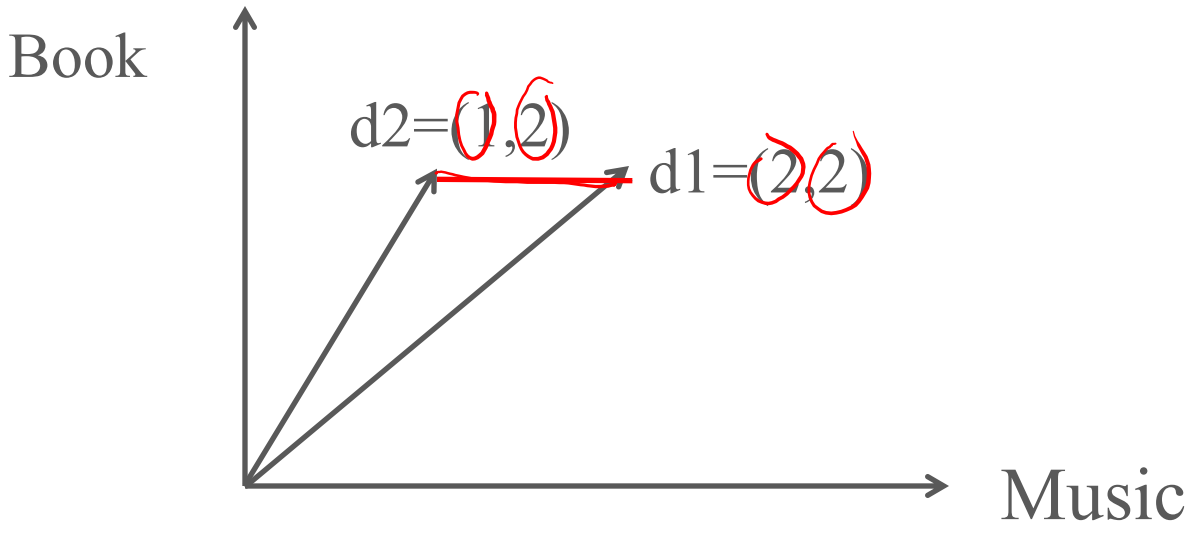
Distance/Similarity Between Two Documents

Distance/similarity measures

- Euclidean distance

$$d = X - Y = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

$$= \sqrt{(1-2)^2 + (2-2)^2} = 1$$



Distance/Similarity Between Two Documents

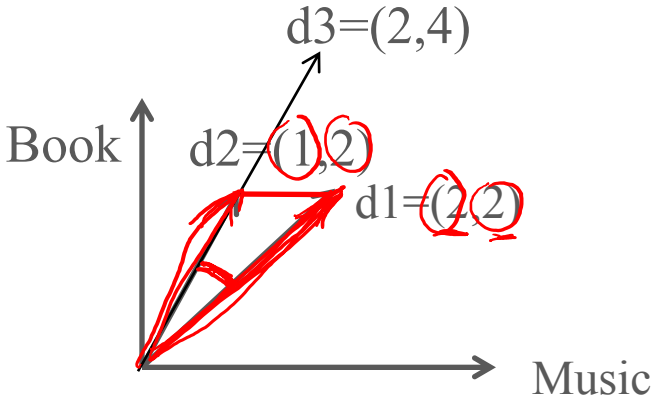
Distance/similarity measure

- Cosine similarity: the cosine value of the angle between the two vectors (0–90°)

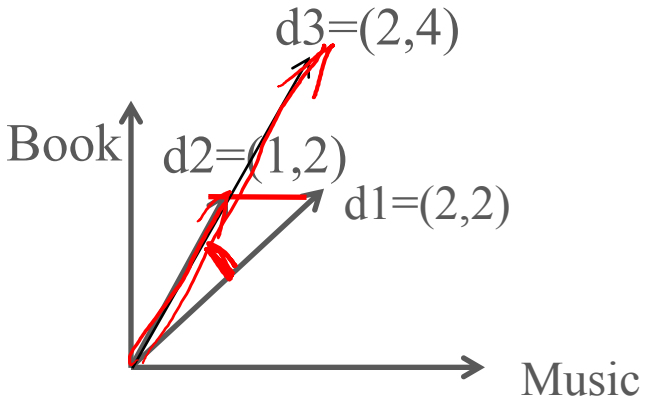
$$\cos(d_1, d_2) = \frac{x \cdot y}{|x||y|} = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2}\ \sqrt{y_1^2 + y_2^2}}$$

$$= \frac{1 \cdot 2 + 2 \cdot 2}{\sqrt{1^2 + 2^2}\ \sqrt{2^2 + 2^2}} = \frac{6}{\sqrt{5}\ \sqrt{8}} = 0.95$$

The greater the angle (distance), the smaller the cosine similarity



Does Vector Length Matter?



- (d1, d2) and (d1, d3) have the same angle
- The cosine similarity has normalized by vector norm
- Therefore, cos_sim(d1,d2) = cos_sim(d1,d3)

Why Are Linear Text Classifiers Popular?

School of Information Studies
Syracuse University

K-Nearest Neighbor (k-NN)

Training process:

- Add in all training examples.

Classification process:

- Given a new example x, compare the similarity between x and all training examples, and choose the majority-voted category label in the k nearest training examples.

K-Nearest Neighbor

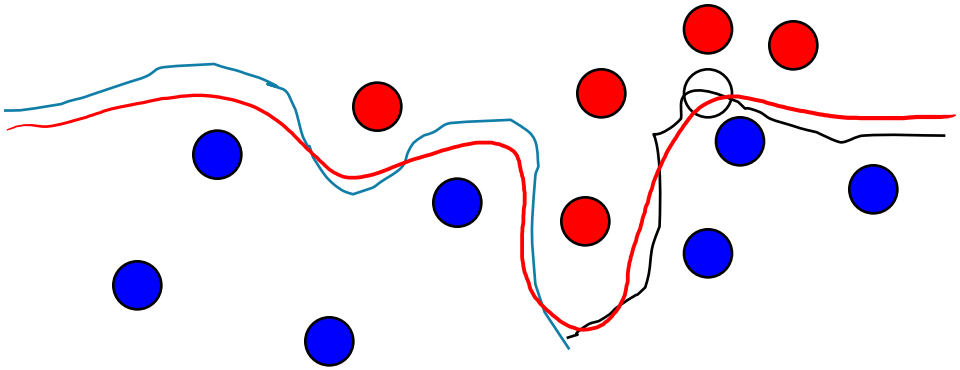
Lazy learner: no learning, just predicting

Instance-based learning

Advantages of k-NN

No assumptions: non-parametric method

When the target function to be learned is very complex



Disadvantages of k-NN

Sensitive to noisy training data

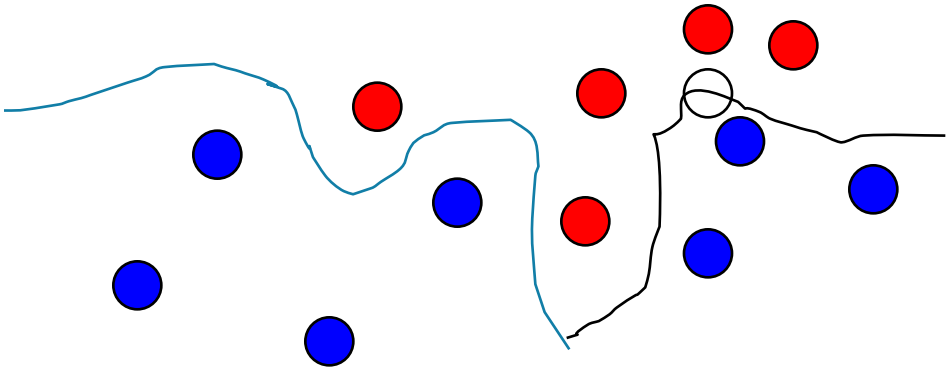
- All attributes participate in classification.
- If only a few attributes are relevant to prediction, the participation of those irrelevant attributes would harm the prediction performance.

High computational cost

- Algorithms like Naïve Bayes would create a model on the training data, and then use the model only to predict new data, which is fast.
- For k-NN, since there is no training step, nearly all computation takes place in the prediction step. If there are many examples and many attributes in the data set, computational cost is high.

The Shape of Decision Function

No regular shape: k-NN

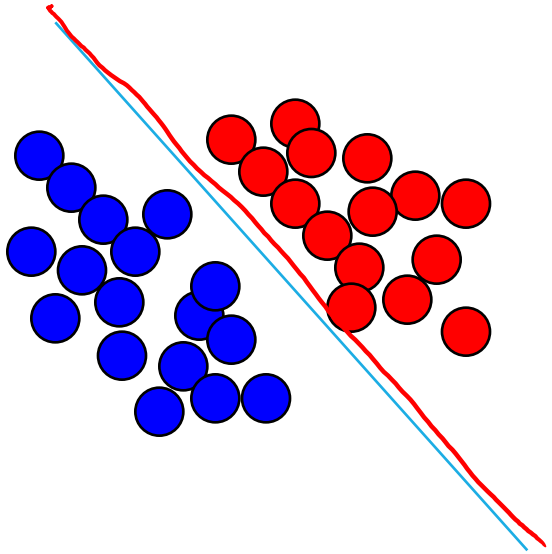


The Shape of Decision Function

Linear: Naïve Bayes, SVM

How many parameters to determine a line in 2D space?

- Y = ax+b
- Weight
- Intercept



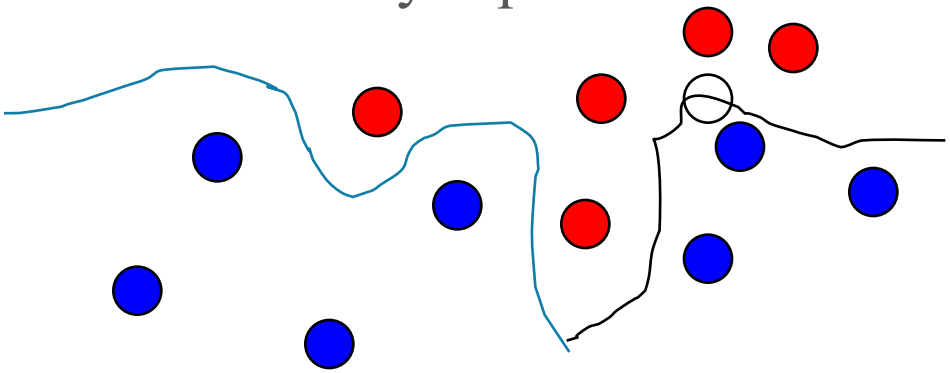
Why Are Linear Classifiers So Popular for Text Classification?

Most text classification problems are linearly separable.

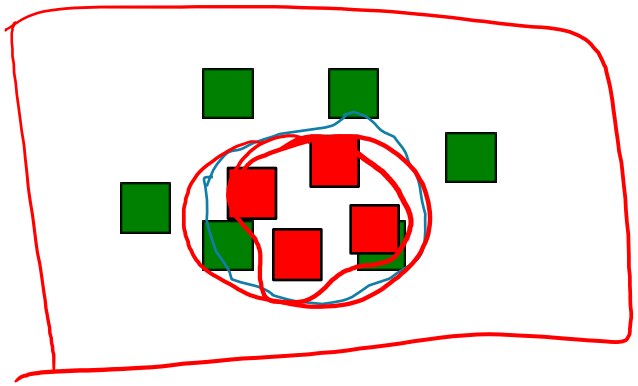
- Large number of features
  - Example: 16K word features in the movie review data
- Usually, a linear boundary can be found to separate the data

The Shape of Decision Function

Some data are not linearly separable



Use non-parametric method like k-NN

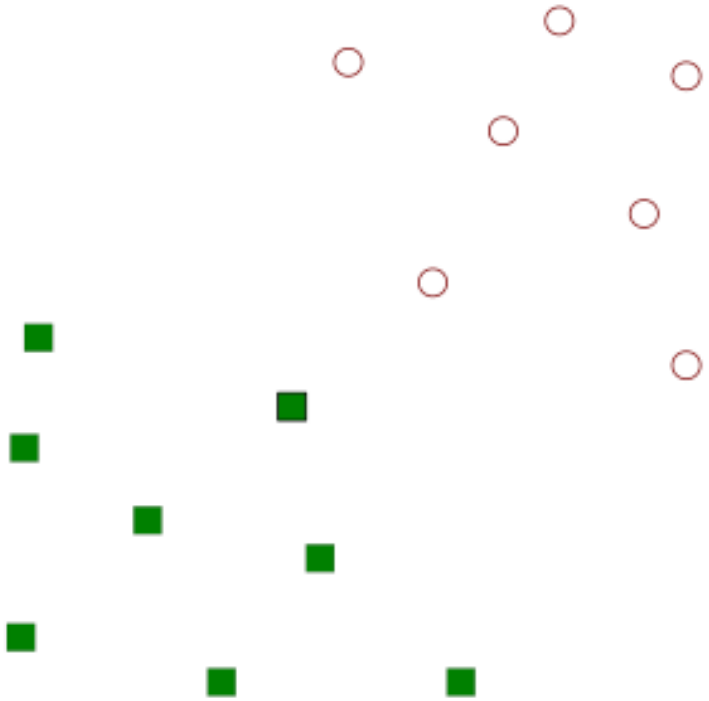


SVMs for Text Categorization

School of Information Studies
Syracuse University

Support Vector Machines

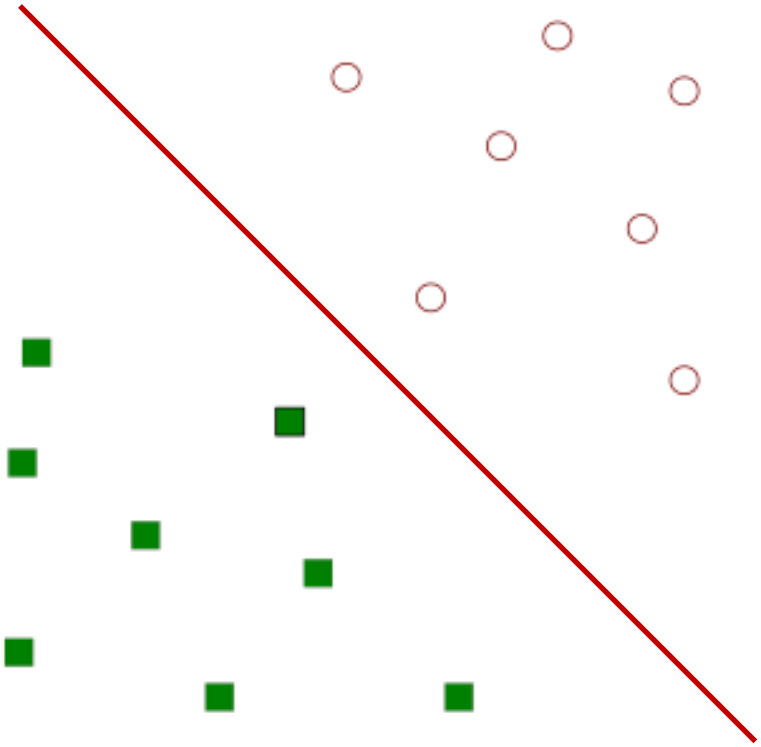
Find a linear hyperplane (decision boundary) that will separate the data



Support Vector Machines

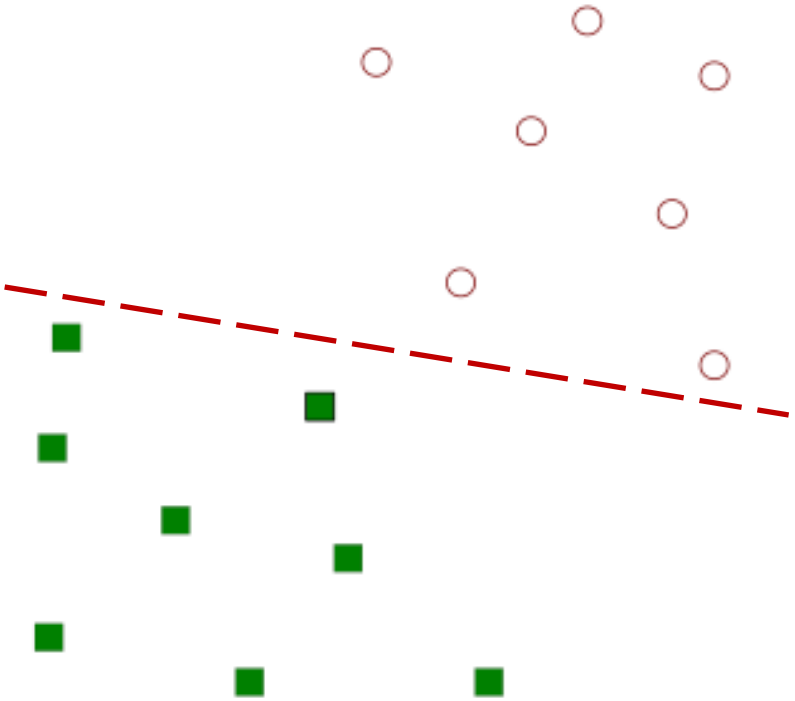
One possible solution

- Decide unknown instance by which side of the line they fall on



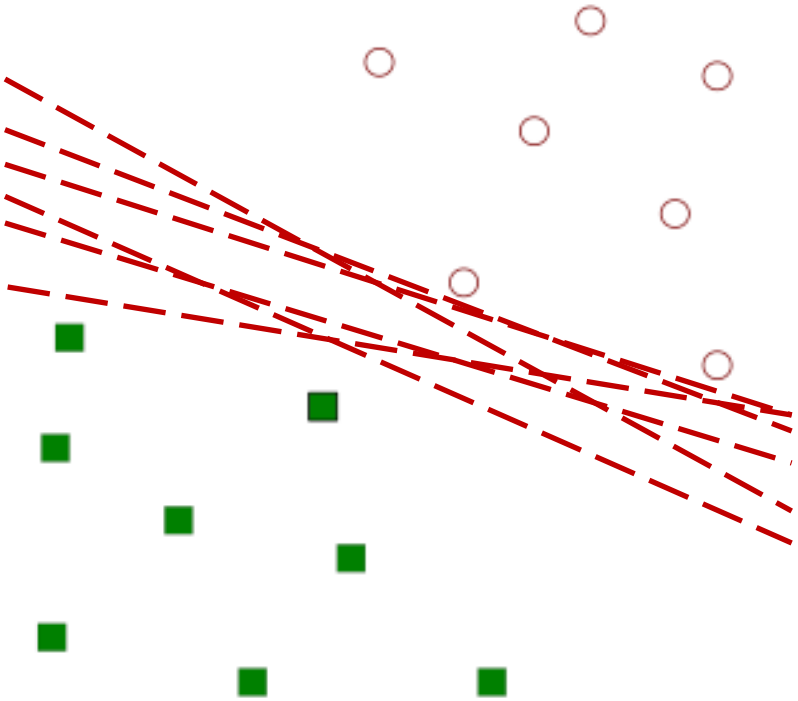
Support Vector Machines

Another possible solution



Support Vector Machines

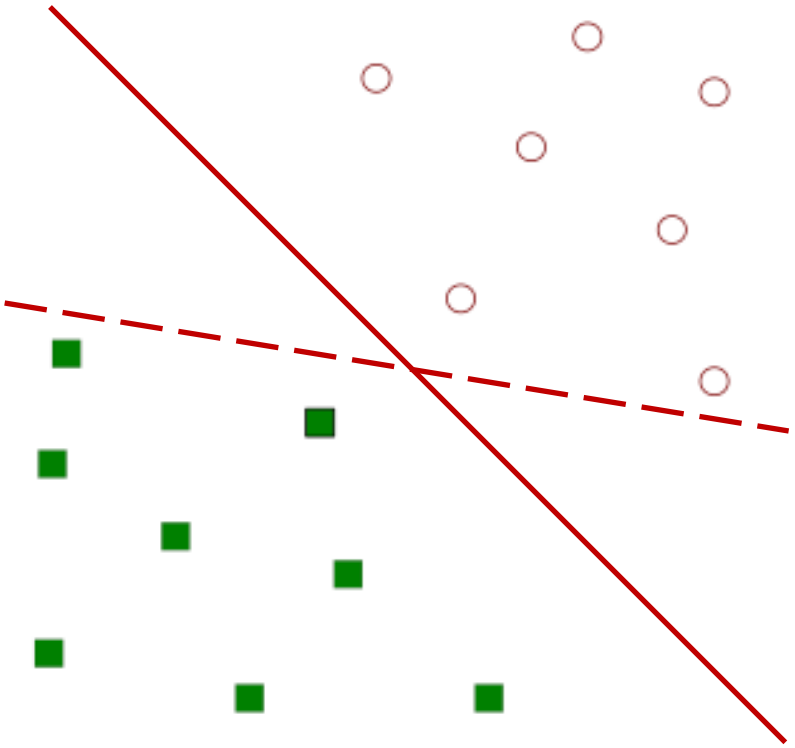
Other possible solutions



Support Vector Machines

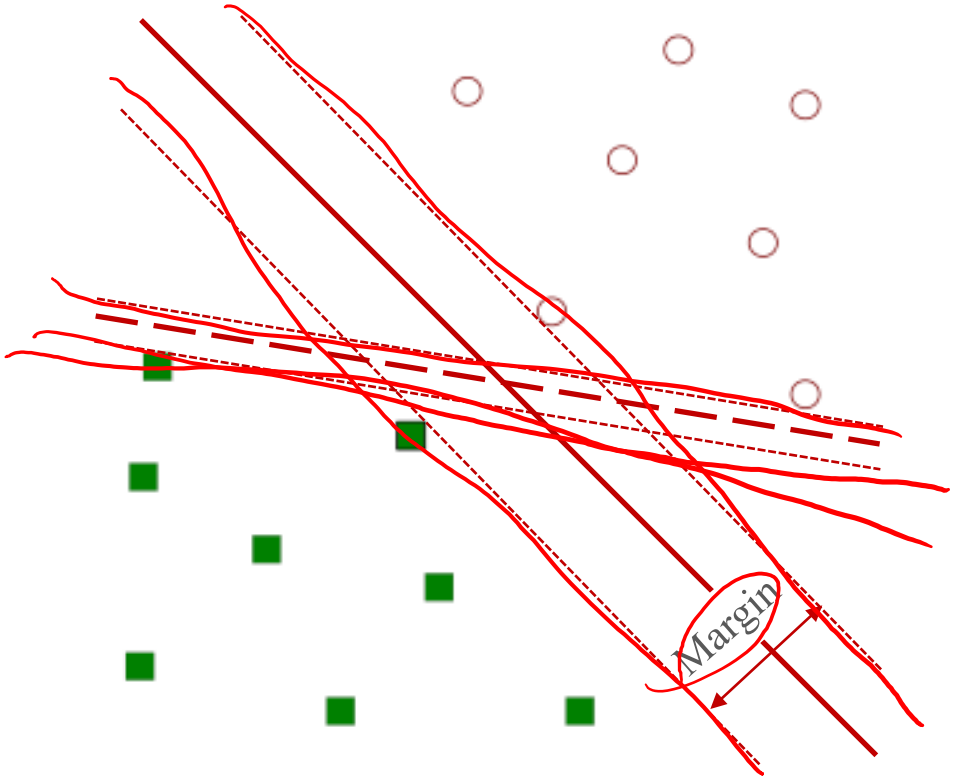
Which one is better? B1 or B2?

How do you define better? (e.g., least square-fitting)

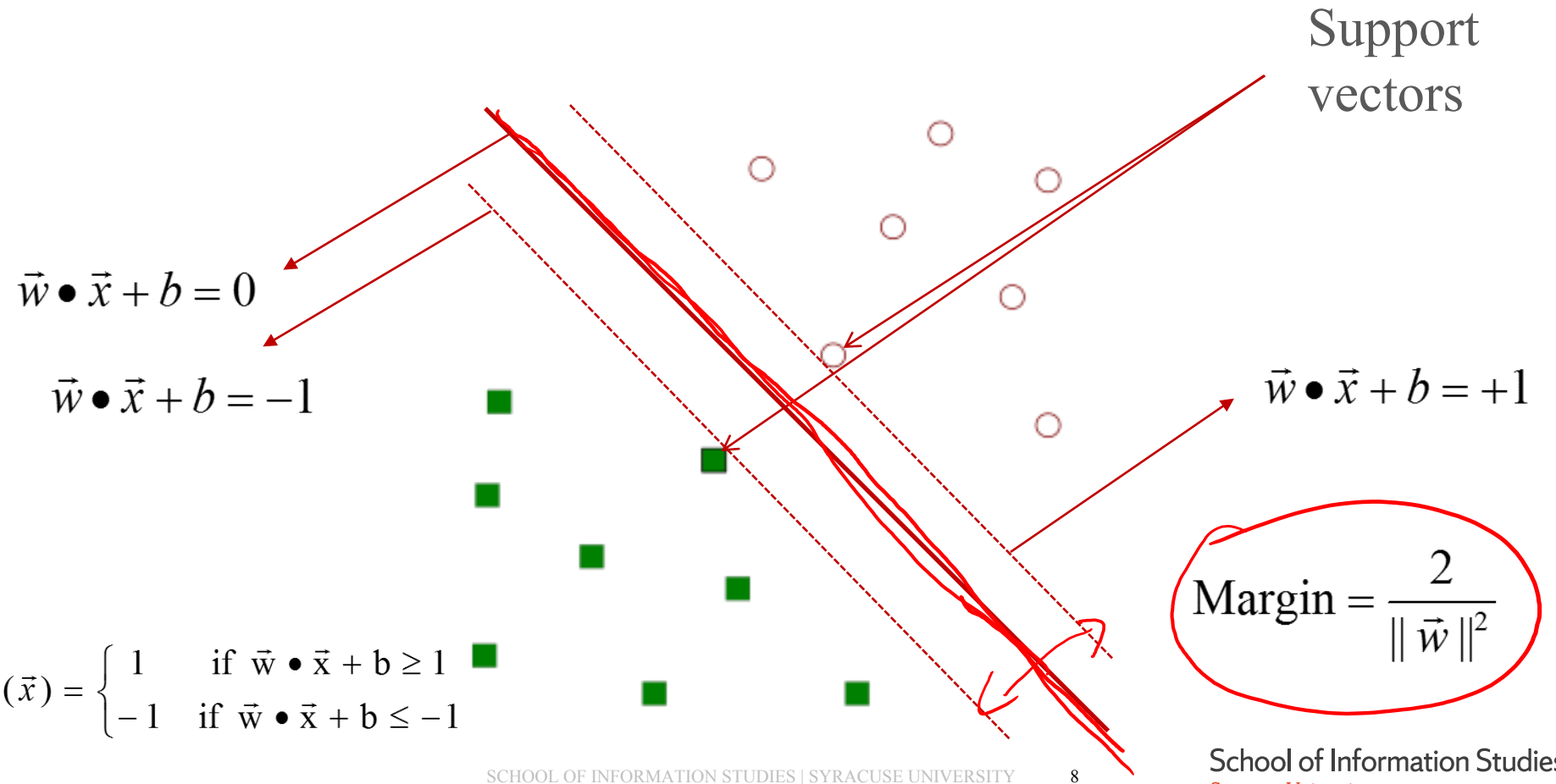


Support Vector Machines

Find a hyperplane that **maximizes** the margin => B1 is better than B2



Support Vector Machines



Support Vectors

Support vectors are the training examples that are located on the margins. They determine the decision boundary.

Training examples which are not support vectors do not participate in prediction.

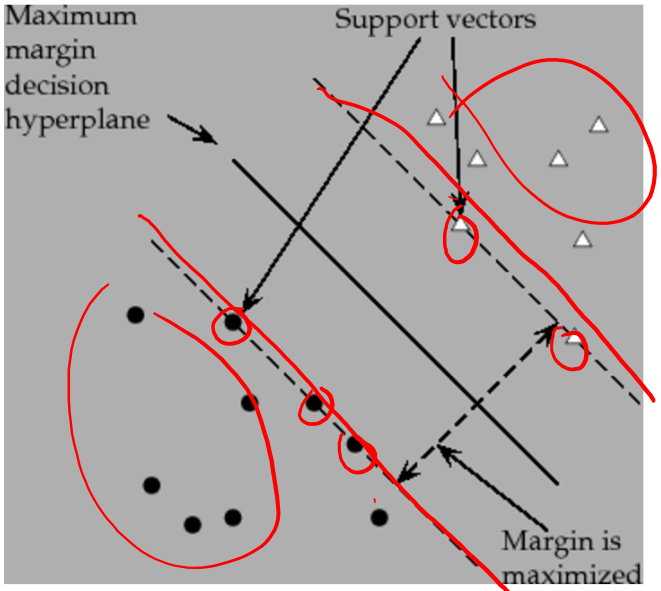
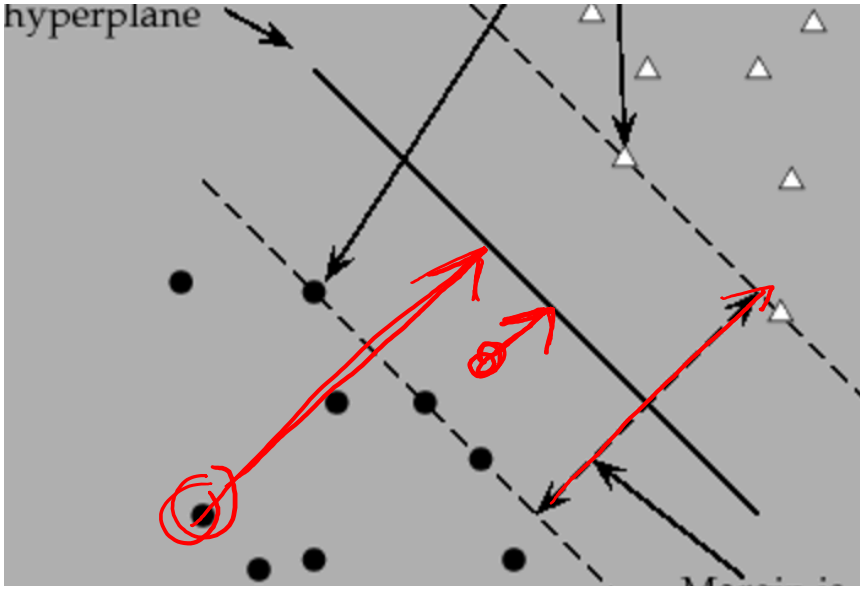


Figure 15.1: The support vectors are the 5 points right up against the margin of the classifier.

Support Vectors

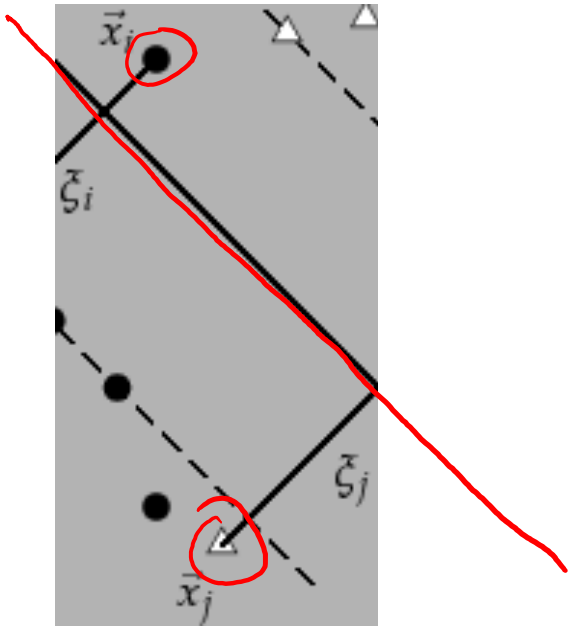
The number of support vectors is an indicator of the complexity of the trained SVM model.

The distance between the example and the decision boundary is an indicator of prediction confidence: the farther the better.



What If the Data Are Not Linearly Separable?

No linear boundary can be found between the two classes.



Soft-Margin SVMs

Introduce a slack variable ξ to pay a cost for each misclassified example.

(N) Find $\vec{w}$, b, and $\xi_i \geq 0$ such that:

- $\frac{1}{2}\vec{w}^T\vec{w} + C\sum_i \xi_i$ is minimized
- and for all $\{(\vec{x}_i, y_i)\}$, $y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i$

Regularization

Tune the regularization parameter C

Default value C = 1

When C (penalty for misclassification) is large, the classifier tries not to make errors on training data, and thus the margin is narrow, more likely to overfit

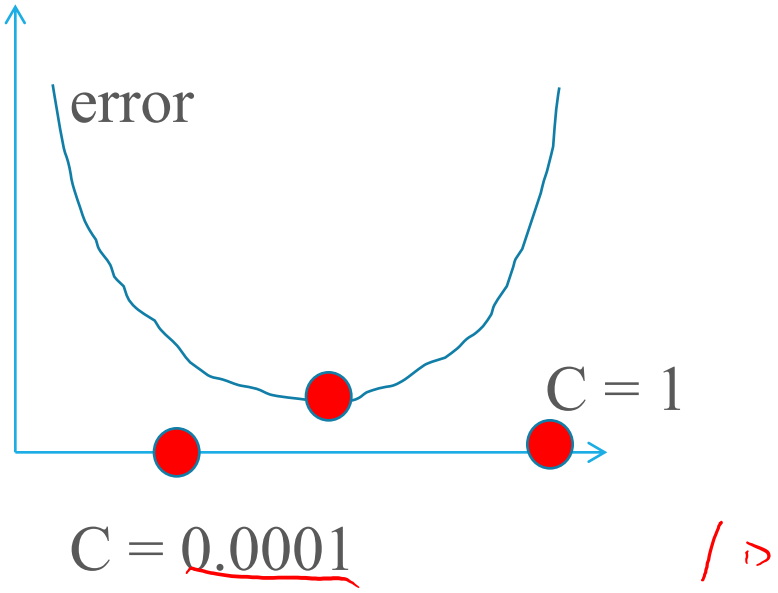
When C is small, wider margin, more robust

However, C cannot be too small, or else it does not respect the data at all

Use manual tuning or gradient descent search to find the best C

Regularization

Use manual tuning or gradient descent search to find the best C.



Kernels in SVMs

School of Information Studies
Syracuse University

Kernel Functions

SVM algorithm maximizes the margin between the two separating hyperplanes by finding the maximum of the functional:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

Subject to the constraints

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2,...,l$$

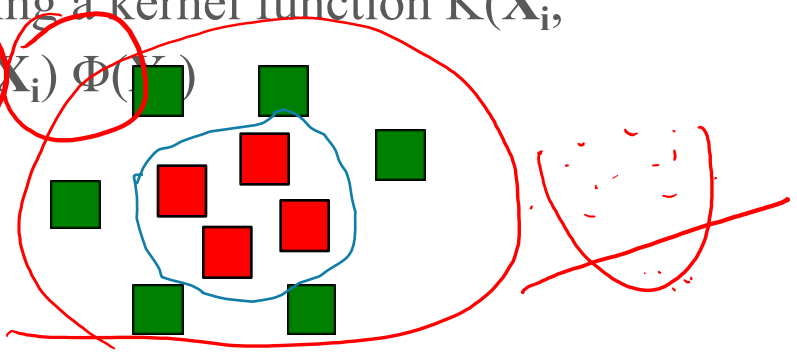
Linear Kernel

K(x_i, x_j) is the dot product of two examples.

Linear kernel is the most commonly used in text classification.

SVM—Kernel Functions

Instead of computing on the transformed data tuples, it is mathematically equivalent to instead applying a kernel function $K(\mathbf{X_i}, \mathbf{X_j})$ to the original data; i.e., $K(\mathbf{X_i}, \mathbf{X_j}) = \Phi(\mathbf{X_i})\ \Phi(\mathbf{X_j})$



Typical Kernel Functions

$$\text{Polynomial kernel of degree } h: \quad K(X_i, X_j) = (X_i \cdot X_j + 1)^h$$

$$\text{Gaussian radial basis function kernel}: \quad K(X_i, X_j) = e^{-\|X_i - X_j\|^2/2\sigma^2}$$

$$\text{Sigmoid kernel}: \quad K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta)$$

Why Is the Kernel Trick Rarely Used in Text Classification?

Most textual data are linearly separable.

- Large number of features **n**
  - Example: ~16K word features for the movie review data

Higher dimensional decision boundaries need more data to fit accurately, otherwise are more likely to overfit than linear decision boundaries.

SVMs for Multi-Class Classification

School of Information Studies
Syracuse University

Extend Binary Classification to Multi-Class

Given **n** classes, for example:

- Sentiment = {positive, negative, neutral, no opinion}

One-vs.-one (pairwise) strategy:

- Create **n(n-1)** classifiers: pos|neg, pos|neu, pos|np, neg|neu, neg|np, neu|np
- Conflicting results?

One-vs.-all strategy:

- Create **n** classifiers: positive or not, negative or not, neutral or not, np or not
- Pick the class with largest prediction value

SVMs Strength and Weakness

School of Information Studies
Syracuse University

SVMs as Classifiers

Weakness

- Require a number of parameters for each kernel type
- Interpretability
  - Easy interpretation for linear kernel
  - Difficult to interpret the model generated by nonlinear kernels

Strength

- High tolerance to noisy data
- Flexibility in data representation: well-suited for continuous- or discrete-valued inputs and outputs
- Probabilistic prediction result
- Scalability: successful on extremely large problems
- Successful on a wide array of real-world data