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

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

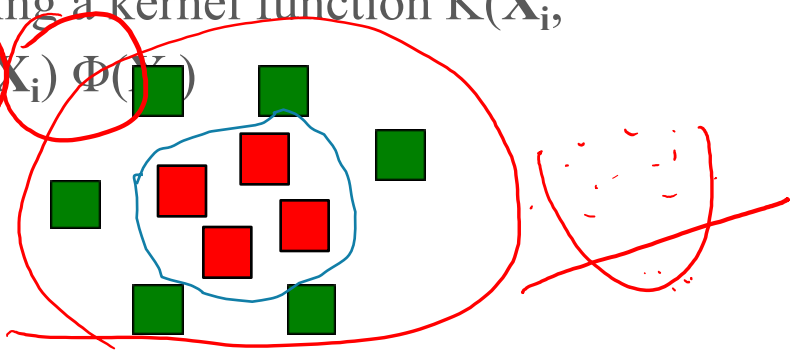
Linear Kernel

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.