Tufts

Class #11:
Kernel Functions & SVMs, II

Machine Learning (COMP 135): M. Allen, 09 Oct. 19

Review: Support Vector Machines (SVMs)

1. Start with labeled data-set:

   $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\} \qquad [\forall i,\ y_i \in \{+1, -1\}]$

2. Solve constrained quadratic optimization problem:

   $$\text{Maximize:} \quad W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

   while satisfying constraints:

   $$\forall i,\ \alpha_i \geq 0$$

   $$\sum_i \alpha_i y_i = 0$$

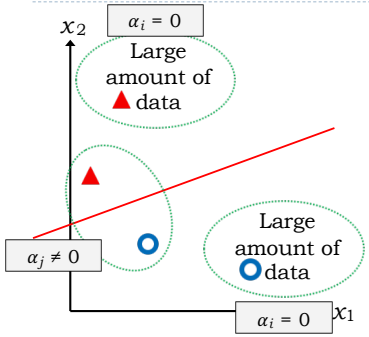
3. Derive necessary weights and biases for decision separator when and if needed:

   $$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \qquad b = -\frac{1}{2}\left(\max_{i \mid y_i = -1} \mathbf{w} \cdot \mathbf{x}_i + \min_{j \mid y_j = +1} \mathbf{w} \cdot \mathbf{x}_j \right)$$

Retaining the Support Vectors

- After computing the various optimizing α values, the SVM typically ends up with:
  1. A large number of data points $\mathbf{x}_i$ with $\alpha_i = 0$
  2. A few special data points $\mathbf{x}_j$ with $\alpha_j \neq 0$
- These special points, the support vectors, can be used by themselves to compute necessary weights and biases
  - Often, the SVM keeps a list of these vectors, for computation of later classification functions, rather than the weights defining the classification boundary directly

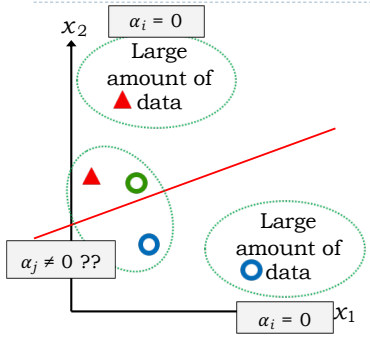
Why Retain the Support Vectors?



Sometimes, retaining only the support vectors comes in handy if we ever want to ***update*** the decision boundary as new data comes in for classification.

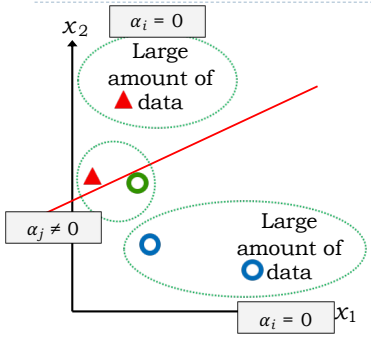
- The α_i values are 0 ***everywhere except*** at the support vectors (the points closest to the separator)

Why Retain the Support Vectors?



If the new data remains close to the old boundary, then we can compute the new α-values using only the new data and the previous support vectors

Why Retain the Support Vectors?



In such scenarios, the data for which $\alpha = 0$ before ***remains*** that way, and never needs to be reconsidered when solving the (compute-intensive) optimization step of the SVM

Only three points have their α-values re-computed

Why Retain the Support Vectors?

- Another reason to retain vectors rather than weights is that SVMs are often used with kernel functions that:
  1. Transform the data
  2. Compute necessary dot-products of points

$$k(\mathbf{x}, \mathbf{z}) = \varphi(\mathbf{x}) \cdot \varphi(\mathbf{z}) \qquad (\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m)$$

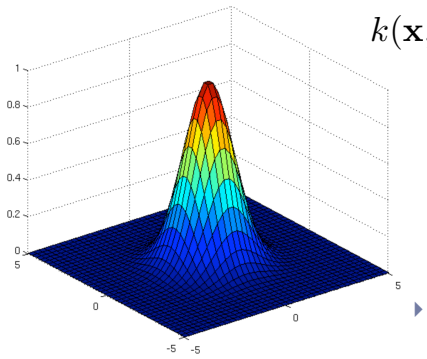
- Furthermore, there are some popular such functions where the data transform translates n-dimensional to m-dimensional data with $n << m$
  - In such cases, storing the original n-dimensional data, and then computing the transformation when necessary, can be much more efficient than trying to store the m-dimensional weight information
  - This is especially true in cases where $m = \infty$ (!!)

Pros and Cons of SVMs

- **[+]** Compared to linear classifiers like logistic regression, SVMs:
  1. Are insensitive to outliers in the data (extreme class examples)
  2. Give a robust boundary for separable classes
  3. Can handle high-dimensional data, via transformation
  4. Can find optimal α-values, with no local maxima
- **[–]** Compared to linear classifiers like logistic regression, SVMs:
  1. Are less applicable in multi-class ($c > 2$) instances
  2. Require more complex tuning, via hyper-parameter selection
  3. May require some deep thinking or experimentation in order to select the appropriate kernel functions

Gaussian Radial Basis Function (RBF)

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$

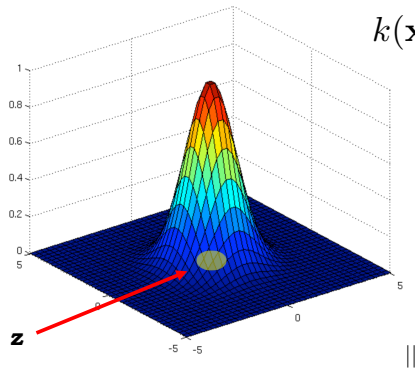


- A popular kernel with many uses is the Gaussian RBF

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$



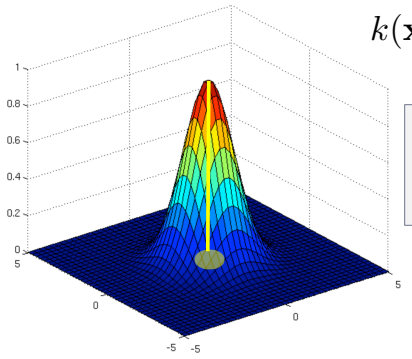
- The RBF is based on a distance from a central focal point, $\mathbf{z}$
- The can be measured in a variety of ways, but is often Euclidean:

$$||\mathbf{x} - \mathbf{z}|| = \sqrt{\sum_{i=1}^{n}(x_i - z_i)^2}$$

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$



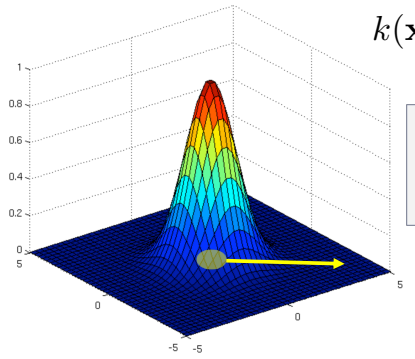
$$||\mathbf{x} - \mathbf{z}|| = 0$$
$$k(\mathbf{x}, \mathbf{z}) = e^0 = 1$$

- The value of the function is highest at point $\mathbf{z}$ itself

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$



$$||\mathbf{x} - \mathbf{z}|| \to \infty$$
$$k(\mathbf{x}, \mathbf{z}) \to e^{-\infty} = 0$$

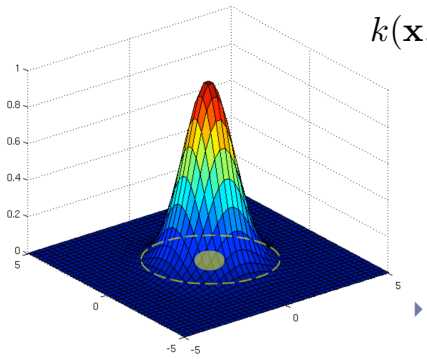
- The value drops to 0 as we get further from $\mathbf{z}$

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$

Tuning Parameter

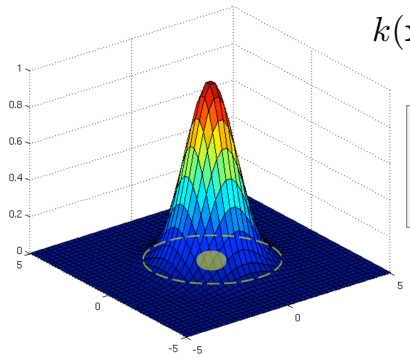


- σ controls the diameter of the non-zero area

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$



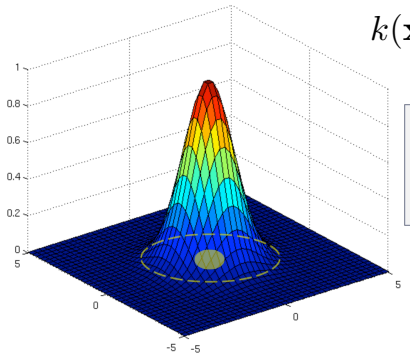
$$\sigma \to \infty$$
$$k(\mathbf{x}, \mathbf{z}) \to e^0 = 1$$

- If σ gets larger, the non-0 area will become wider

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$



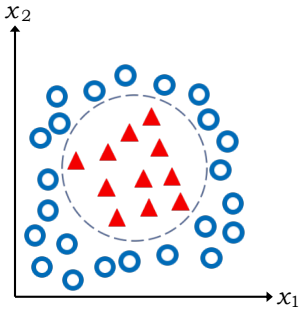
$$\sigma \to 0$$
$$k(\mathbf{x}, \mathbf{z}) \to e^{-\infty} = 0$$

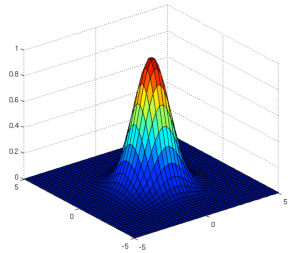
- If σ gets smaller, non-0 area will become narrower

Image source: https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Radial Basis Function

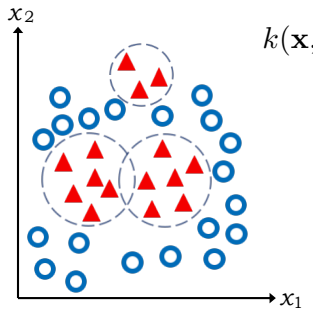
$$k(\mathbf{x}, \mathbf{z}) = e^{-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}}$$





- The radius around the focal point $\mathbf{z}$ at which the function becomes 0 corresponds to the decision boundary in our data

Gaussian Radial Basis Function



$$k(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \sum_{j=1}^{3} e^{-\frac{||\mathbf{x}-\mathbf{z}_j||^2}{2\sigma^2}}$$

- We can deal with multiple clusters in the data by using a combination of multiple RBFs

Next Week

- **Topics**: SVMs and Feature Engineering
- **Meetings**: Tuesday and Wednesday, usual time
- **Readings**: Linked from class website schedule page
  - Includes original paper (Brown, et al.) for discussion
- **Homework 03**: due Wednesday, 16 October, 9:00 AM
- **Project 01**: out Tuesday; due Monday, 04 November, 9:00 AM
- **Office Hours**: 237 Halligan, Tuesday, 11:00 AM – 1:00 PM
  - TA hours can be found on class website as well