Kernel Methods for Machine Learning

Michael Rabadi

New York University

michael.rabadi@nyu.edu

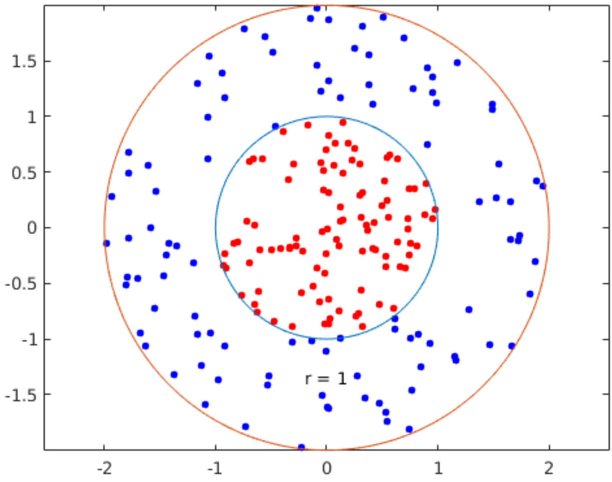
June 30, 2015

Outline

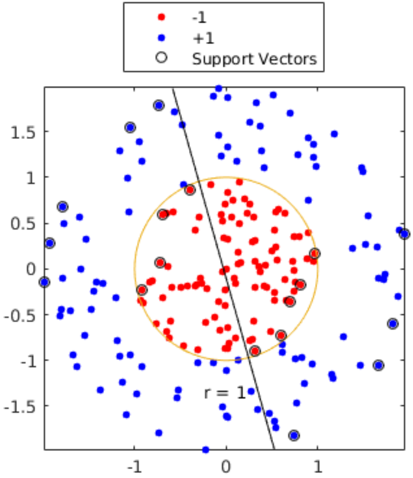
Review - SVMs

- Linear classifier that uses support vectors on margin
- Strong generalization guarantees based on margin
- But what if data is not linearly separable?

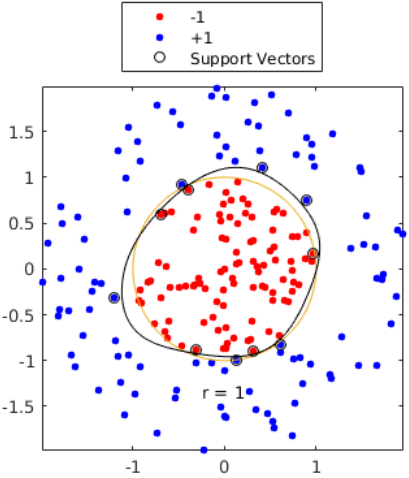
Example



SVM - Linear



SVM - Gaussian Kernel



Problem

- If data not linearly separable, must change space.
- Non-linear function ϕ maps input space to high-dimensional space $\mathbb{H}$.
- SVM generalization doesn't depend on dimension of feature space
- BUT - determining hyperplane in high-dimensional space requires computing multiple inner products.

Kernels

- A kernel is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- Define kernels so that given two points $x, x' \in \mathcal{X}$, $K(x, x')$ is equivalent to an inner product of vectors $< \phi(x), \phi(x') >$
- $\phi : \mathcal{X} \to \mathbb{H}$, where $\mathbb{H}$ is a Hilbert space called a feature space.
- Note: $\mathbb{H}$ can be infinite dimensional, yet the Kernel can be computed in finite time.
- The Kernel K is arbitrary, as long as ϕ exists
- ϕ is guaranteed as long as K is positive definite symmetric.

Outline

PDS Kernels

- A kernel K is pds if a kernel matrix $\mathbf{K} = [K(x_i, x_j)]_{ij}$ is symmetric positive semidefinite (SPSD).
- $\mathbf{K}$ is SPSD if it is symmetric and its eigenvalues are non-negative
- $\mathbf{K}$ is also known as the Gram matrix

Polynomial Kernels

- $\forall \mathbf{x}, \mathbf{x} \in \mathbb{R}^N, K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^d$
- $c > 0$ is a constant and $d \in \mathbb{N}$ is the degree
- Example: $N = 2$ and for second degree polynomial

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^2, K(\mathbf{x}, \mathbf{x}') = (x_1 x_1' + x_2 x_2' + c)^2$$
$$= [x_1^2, x_2^2 \sqrt{2} x_1 x_2, \sqrt{2c} x_1, \sqrt{2c} x_2, c]^\top \cdot [x_1^2, x_2^2 \sqrt{2} x_1 x_2, \sqrt{2c} x_1, \sqrt{2c} x_2, c]^\top$$

Gaussian Kernels (RBF)

- $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^n, K(\mathbf{x}, \mathbf{x}') = \exp\left(- \frac{\|\mathbf{x}' - \mathbf{x}\|^2}{2\sigma^2} \right)$
- The power series expansion of K shows that the corresponding $\mathbb{H}$ is inifite dimensional:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{n=0}^{+\infty} \frac{(\mathbf{x} \cdot \mathbf{x}')^n}{\sigma^n n!}$$

Sigmoid Kernels

- $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^N, K(\mathbf{x}, \mathbf{x}') = \tanh(a(\mathbf{x} \cdot \mathbf{x}') + b)$ for $a, bgeq0$
- SVMs with sigmoid kernels coinside with single layer perceptrons

Outline

Kernel-SVMs

- Recall the dual form of the constrained optimization problem of SVMs:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

subject to : $0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$.

- PDS kernels implicitly define an inner product in $\mathbb{H}$, so replace inner products $x \cdot x'$ with $K(x, x')$:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to : $0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$.

- This leads to: $h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i)\Big)$

Learning guarantees

Rademacher Complexity:
Given a sample $S \subseteq \{x : K(x,x) \leq r^2\}$ of size m and $H = \{x \mapsto \mathbf{w} \cdot \phi(x) : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ for $\Lambda \geq 0$:

$$\hat{\mathfrak{R}}_S(H) \leq \frac{\Lambda\sqrt{\mathrm{Tr}[\mathbf{K}]}}{m} \leq \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Margin bounds: Now if $r^2 = \sup_{x \in X} K(x,x)$ and $\rho > 0$ is the margin, then with probability at lest $1 - \delta$, for any $h \in H$:

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Outline

MATLAB Tutorial

MATLAB Tutorial

Conclusion

- Kernel methods are used to extend linear classifiers to non-linear spaces
- PDS Kernels implicitly define inner products in high-dimensional space
- Generalization bounds for Kernels depends on trace of Kernel matrix

Resources

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.