Support Vector Machines (SVMs)

RBF: Some interesting properties

MIT, 2009 fall, Tommy Jaakkola, HW2, pr. 1

We can write the radial basis kernel in the following form:

$$K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right),$$

where x and x' belong to $\mathbb{R}^d$, and σ is a width parameter specifying how quickly the kernel vanishes as the points move further away from each other.

We will show that this kernel has some remarkable properties:
It can perfectly separate *any* finite set of *distinct* training points. Moreover, this result holds for *any* positive finite value of σ.

[**However, while the kernel width does not affect whether we'll be able to perfectly separate the training points, it does affect generalization performance.**]

Let's proceed in stages.

We'll first show that the optimisation problem

$$\text{minimize } \frac{1}{2}\|\theta\|^2 \text{ subject to } y_i\theta \cdot \phi(x_i) = 1, \ldots, n$$

has a solution regardless of how we set the ± 1 **training labels** y_i**.**
Here $\phi(x_i)$ **is the feature vector (function actually) corresponding to the radial basis kernel** K**.**

Note**: Our formulation here is a bit non-standard for two reasons:**

1. We try to find a solution where all the points are support vectors.
2. We also omit the bias term since it is not needed for the result.

a. Introduce Lagrange multipliers for the constraints [similarly to finding the SVM solution] and show the form that the solution θ^* has to take, i.e. express $\hat{\theta}$ as a function of the Lagrange multipliers. (This should not involve lengthy calculations.)

***Notes*:**

***i*. The Lagrange multipliers here are no longer constrained to be positive. (Since you are trying to satisfy equality constraints, the Lagrange multipliers can take any real value.)**
***ii*. You can assume that θ and $\phi(x_i)$ are finite vectors for the purposes of these calculations.**

b. Put the resulting solution back into the classification (margin) constraints and express the result in terms of a linear combination of the radial basis kernels.

Solution

a. The *Lagrangian* for this optimization problem is:

$$L(\theta, \alpha) \quad = \quad \frac{1}{2}\|\theta\|^2 - \sum_{i=1}^{n} \alpha_i (y_i \theta \cdot \phi(x_i) - 1) = \frac{1}{2}\|\theta\|^2 - \theta \cdot \left(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \right) + \sum_{i=1}^{n} \alpha_i$$

Here each α_i is ***unconstrained***, because we have equality constraints rather than inequality constraints. As usual, the *dual optimization problem* is

$$\max_{\alpha} g(\alpha) = \max_{\alpha} \underbrace{\min_{\theta} L(\theta, \alpha)}_{\text{not.: } g(\alpha)} .$$

For a fixed α, the expression $L(\theta, \alpha)$ is positively quadratic in θ. We can obtain the optimal θ^* from the first-order condition $\dfrac{\partial L(\theta, \alpha)}{\partial \theta} = 0$:

$$\theta^* = \sum_{j=1}^{n} \alpha_j^* y_j \phi(x_j).$$

For convenience, we will use the short-hand $\theta^* = \Phi(y \bullet \alpha^*)$. Here $\bullet$ represents an element-wise product and Φ is a $m \times n$ matrix, where the i^{th} column is $\phi(x_i)$. (Of course, $m = \infty$ for the RBF kernel.)

b. Our constraints are equivalent to:

$$\phi(x_i)^\top \theta = y_i,\ i = 1, \ldots, n.$$

Using matrix short-hand notation and substituting $\theta^* = \Phi(y \bullet \alpha^*)$**, we obtain:**

$$\begin{array}{rcl} \Phi^\top \theta^* & = & y \\ \Phi^\top \Phi(y \bullet \alpha^*) & = & y \\ K(y \bullet \alpha^*) & = & y, \end{array}$$

where $K \overset{not.}{=} \Phi^\top \Phi$ **denotes the Gram matrix.**

c. Indicate briefly how we can use the following Michelli theorem to show that any n by n RBF kernel matrix $K_{ij} = K(x_i, x_j)$ for $i, j = 1, \ldots, n$ is invertible.

***Theorem (Michelli)*: If $\rho(t)$ is a monotonic function in $t \in [0, \infty)$, then the matrix $\rho_{ij} = \rho(\|x_i - x_j\|)$ is invertible for any distinct set of points $x_i, i = 1, \ldots, n$.**

d. Based on the above results put together the argument to show that we can indeed find a solution where all the points are support vectors.

c. Note that $\rho(t) = \exp\left(-\frac{1}{2\sigma^2}t^2\right)$ is a monotonic function in $t \in [0, \infty)$. Using the Michelli theorem, for any distinct set of points $x_i, i = 1, \ldots, n$, the matrix K, with entries $K_{ij} = \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$, is invertible.

d. As we have a distinct set of points, K is invertible. Then the linear system $K(y \bullet \alpha^*) = y$ is feasible, and has a unique solution given by $\alpha^* = y \bullet (K^{-1}y)$. Therefore, $\theta^* = \Phi(y \bullet \alpha^*) = \Phi K^{-1}y$.

e. Of course, the fact that we can in principle separate any set of training examples does not mean that our classifier does well (on the contrary). So, why do we use the radial basis kernel? The reason has to do with margin that we can attain by varying σ. Note that the effect of varying σ on the margin is not simple rescaling of the feature vectors. Indeed, for the radial basis kernel we have

$$\|\phi(x)\|^2 = \phi(x) \cdot \phi(x) = K(x, x) = 1.$$

Let's begin by setting σ to a very small positive value. What is the margin that we attain in response to any n distinct training points?

f. Provide a 1-dimensional example to show how the margin can be larger than the answer to part e. You are free to set σ and the points so as to highlight how they might "contribute to each other's margin".

e. As $\sigma \to 0$, the points become very far apart with respect to σ, and our kernel matrix $K \to I$, the identity matrix. Because our constraints dictate that $K(y \bullet \alpha^*) = y$, then $\alpha^* \to \bar{1}$, the all-ones vector. Therefore, $\|\theta^*\|^2 = {\alpha^*}^\top K \alpha^* \to \bar{1}^\top I \bar{1} = n$, and we obtain a margin of $\sqrt{\frac{1}{n}}$ in the limit.

f. The simplest example to create is a set of 2 distinct points x and x', both labeled +1. Denote $k = K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$. The Gram matrix can be written as:

$$K = \begin{bmatrix} 1 & k \\ k & 1 \end{bmatrix}$$

Solving the system $K(\bar{1} \bullet \alpha) = \bar{1}$ yields the solution $\alpha^* = \left[\frac{1}{k+1} \; \frac{1}{k+1}\right]^\top$ and $\|\theta^*\|^2 = {\alpha^*}^\top K \alpha^* = \frac{2}{k+1}$. Therefore, the margin is $\frac{1}{\|\theta^*\|} = \sqrt{\frac{k+1}{2}}$. As long as x and x' are distinct, we have that $k > 0$, so the margin is always greater than $\sqrt{\frac{1}{2}}$.

Remark

As we take the 2 points arbitrarily close together (or alternatively, imagine that $\sigma \to +\infty$), then $k \to 1$, and we obtain a margin of 1.

Is 1 always the largest possible margin that we can obtain?

For the RBF kernel, one can think of the corresponding infinite-dimensional feature vectors $\phi(x_i)$ as lying on the unit ball, as they are unit-normalized: $\|\phi(x_i)\|^2 = K(x_i, x_i) = 1$. So yes, the largest possible margin must be 1.

Intuitively, as $\sigma \to +\infty$, kernels centered at distinct points gradually become indistinguishable. In effect, all the feature vectors collapse onto each other (to a single point) on the unit ball. As they do, the margin goes to 1 in the limit.