Alin Tomescu
6.867 Machine learning | Week 3, Thursday, September 19th, 2013| Lecture 5

Lecture 5: Kernels continued

Kernels are a way to compare training examples. Using kernels, we can cast the perceptron algorithm in terms of comparing examples, as opposed to looking at the data.

$K(x, x') = x \cdot x'$ is called a linear kernel (means linear classification)

$K(x, x') = (x \cdot x')^p, p = 1,2,..$, is a p^{th} order polynomial kernel

$K(x, x') = (1 + x \cdot x')^p$ is a general polynomial kernel (this compiles all orders up to p and including p)

$K(x, x') = e^{-\beta\|x-x'\|^2}, \beta > 0$, is a radial basis kernel

⋮

SVMs with kernels

$\min \frac{1}{2}\|\theta\|^2$ s.t. $y^{(i)}\theta\phi\left(x^{(i)}\right) \geq 1, i = 1,2, \dots n$

The **Lagrangian** for this quadratic optimization problem is:

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 - \sum_{i=1}^{n} \alpha_i \left(y^{(i)}\theta\phi\left(x^{(i)}\right) - 1\right), \alpha_i \geq 0$$

Primal problem: we are trying to find θ that *minimizes* $\max_{\alpha \geq 0} L(\theta, \alpha)$:

$$\theta = \min_{\theta} \max_{\alpha \geq 0} L(\theta, \alpha)$$

$$\max_{\alpha \geq 0} L(\theta, \alpha) = \begin{cases} \frac{1}{2}\|\theta\|^2, & \text{if all the constraints are satisfied: } y^{(i)}\theta\phi\left(x^{(i)}\right) \geq 1, \forall i \\ \infty, otherwise \end{cases}$$

Dual problem: We are trying to find the α_i values that *maximizes* $\min_{\theta} L(\theta, \alpha)$:

$$\max_{\alpha \geq 0} \left\{\min_{\theta} L(\theta, \alpha)\right\}$$

Note that $L(\theta, \alpha)$ is an **increasing function** and a **convex function** (shaped like a U) with respect to θ, so we can get its minimum if we set the derivate equal to 0.

$$\frac{\partial L(\theta, \alpha)}{\partial \theta} = \theta - \sum_{i=1}^{n} \alpha_i y^{(i)} \phi\left(x^{(i)}\right) = 0 \Leftrightarrow \theta_{min} = \sum_{i=1}^{n} \alpha_i y^{(i)} \phi\left(x^{(i)}\right)$$

Now, if we replace this minimum θ_{min} in the original constraint, we get:

Alin Tomescu

$$L\left(\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right),\alpha\right)=\frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right)\right\|^2-\sum_{i=1}^{n}\alpha_i\left(y^{(i)}\left(\sum_{j=1}^{n}\alpha_j\, y^{(j)}\phi\left(x^{(j)}\right)\right)\phi\left(x^{(i)}\right)-1\right)$$

Since $\|\theta\|^2=\theta\cdot\theta$, we have:

$$L\left(\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right),\alpha\right)$$

$$=\frac{1}{2}\left[\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right)\right]^T\left[\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right)\right]+\sum_{i=1}^{n}\alpha_i-\sum_{i=1}^{n}\alpha_i\left(y^{(i)}\left(\sum_{j=1}^{n}\alpha_j\, y^{(j)}\phi\left(x^{(j)}\right)\right)\phi\left(x^{(i)}\right)\right)$$

$$=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi\left(x^{(i)}\right)\phi\left(x^{(j)}\right)+\sum_{i=1}^{n}\alpha_i-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi\left(x^{(i)}\right)\phi\left(x^{(j)}\right)$$

$$=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi\left(x^{(i)}\right)\phi\left(x^{(j)}\right)=$$

Thus we have **reformulated the dual** as:

$$\max_{\alpha\geq 0}\left\{\min_{\theta}L(\theta,\alpha)\right\}=\max_{\alpha\geq 0}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi\left(x^{(i)}\right)\phi\left(x^{(j)}\right)\right\},s.t.\ \alpha_i\geq 0$$

Is $\min_{\theta}\left\{\max_{\alpha\geq 0}L(\theta,\alpha)\right\}=\max_{\alpha\geq 0}\left\{\min_{\theta}L(\theta,\alpha)\right\}$? Only under Slater conditions that hold.

KKT conditions for primal dual optimality

$$\frac{\partial L(\theta,\alpha)}{\partial\theta}=\theta-\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right)=0\Leftrightarrow\theta_{min}=\sum_{i=1}^{n}\alpha_i\, y^{(i)}\phi\left(x^{(i)}\right)$$

Note that if $\alpha_i=0$, then the i^{th} example does not matter, since it does not affect the value of θ.

Complementary slackness: $\alpha_i\left(y^{(i)}\theta\phi\left(x^{(i)}\right)-1\right)=0$ **(this says that only SVs can have non-zero α_i)**

Primal feasibility: $y^{(i)}\theta\phi\left(x^{(i)}\right)\geq 1$, or in the dual form: $y^{(i)}\left(\sum_{j=1}^{n}\alpha_j\, y^{(j)}K\left(x^{(i)},x^{(j)}\right)\right)\geq 1$

Dual feasibility: $\alpha_i\geq 0$

These must hold true.

If $\alpha_i>0$ then the complementary slackness tells me that:

$$y^{(i)}\theta\phi\left(x^{(i)}\right)-1=0\Leftrightarrow y^{(i)}\left(\sum_{j=1}^{n}\alpha_j\, y^{(j)}K\left(x^{(i)},x^{(j)}\right)\right)=1$$

Alin Tomescu

These are the **support vectors**.

If $\alpha_i = 0$, then the $x^{(i)}$ will be not be a support vector, which means it'll sit outside the margin so:

$$y^{(i)}\theta_{min}\phi(x^{(i)}) \geq 1 \Leftrightarrow y^{(i)}\left(\sum_{j=1}^{n} \alpha_j\, y^{(j)}\phi(x^{(j)})\right)(x^{(i)}) \geq 1 \Leftrightarrow y^{(i)}\sum_{j=1}^{n} \alpha_j y^{(j)}K(x^{(i)}, x^{(j)}) \geq 1$$

The primal solution is unique, but its representation in terms of the Lagrange multipliers is not unique in general.

- **What does this mean? Does it mean that you can come up with different α_i values that give you the same θ_{min}?**

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y^{(i)}y^{(j)}K(x^{(i)}, x^{(j)})\,, s.t.\, \alpha_i \geq 0$$

$$sign(\theta(\alpha)\phi(x)) = sign\left(\sum_{i=1}^{n} \alpha_i y^{(i)}K(x^{(i)}, x)\right)$$

Let's include slack. The generalized Lagrangian is:

$$L(\theta, \theta_0, \xi, \alpha, r) = \frac{1}{2}\|\theta\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i - \sum_{i=0}^{n} \alpha_i \left(y^{(i)}\left(\theta\phi(x^{(i)}) + \theta_0\right) - 1 + \xi_i\right) - \sum_{i=0}^{n} r_i\xi_i$$

$$\nabla_\theta L = \frac{\partial L}{\partial \theta} = \theta - \sum_{i=1}^{n} \alpha_i y_i\, \phi(x^{(i)}) = 0$$

$$\frac{\partial L}{\partial \theta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = \frac{c}{n} - \alpha_i - r_i = 0 \text{ or } \alpha_i = \frac{c}{n} - r_i$$

Note: This last minimum constraint is what ends up constraining the value of α to $\frac{c}{n}$ and what ends up dropping all the ξ_i terms in the dual since the linear combination $\frac{c}{n} - \alpha_i - r_i = 0$. (see http://www.youtube.com/watch?v=XUj5JbQihlU at 60 minutes in)

Without offset:

$$\min_{\theta, \xi} \frac{1}{2}\|\theta\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \text{ s.t. } y^{(i)}\theta\phi(x^{(i)}) \geq 1 - \xi_i, \xi_i \geq 0$$

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y^{(i)}y^{(j)}\phi(x^{(i)})\phi(x^{(j)}),$$

$$\text{such that } \alpha_i \geq 0 \text{ and } \alpha_i \leq \frac{c}{n}$$

Alin Tomescu

With offset:

$$\min_{\theta,\theta_0,\xi} \frac{1}{2}\|\theta\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \text{ s.t. } y^{(i)}\left(\theta\phi\left(x^{(i)}\right)+\theta_0\right) \geq 1-\xi_i, \text{such that } \xi_i \geq 0$$

$$\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi\left(x^{(i)}\right)\phi\left(x^{(j)}\right),$$

$$\text{such that } \alpha_i \in \left[0,\frac{c}{n}\right] \text{ and } \sum_{i=1}^{n}\alpha_i y^{(i)} = 0$$

Then we can build the classifier as:

$$sign(\theta(\alpha)\phi(x)+\theta_0) = sign\left(\sum_{i=1}^{n}\alpha_i y^{(i)}K\left(x^{(i)},x\right)+\theta_0\right)$$

$$\alpha_i = 0 \Rightarrow y^{(i)}\left(\sum_{j=1}^{n}\alpha_j y^{(j)}K\left(x^{(j)},x^{(i)}\right)+\theta_0\right) \geq 1$$

$$\alpha_i = \frac{c}{n} \Rightarrow y^{(i)}\left(\sum_{j=1}^{n}\alpha_j y^{(j)}K\left(x^{(j)},x^{(i)}\right)+\theta_0\right) \leq 1$$

$$\alpha_i \in \left(0,\frac{c}{n}\right) \Rightarrow y^{(i)}\left(\sum_{j=1}^{n}\alpha_j y^{(j)}K\left(x^{(j)},x^{(i)}\right)+\theta_0\right) = 1$$

Support vectors are those with $\alpha_i = \frac{c}{n}$ and $\alpha_i \in \left(0,\frac{c}{n}\right)$

Classification problem where we have only one class

I have a bunch of points and I want to know if another point is part of them?

Find the minimum enclosing ball. Search over where to place the center of the circle and how long should be the radius.

$$\min r^2 \;\; s.t. \left\|\theta - \phi\left(x^{(i)}\right)\right\|^2 \leq r^2$$

Is x an outlier? Once I have the solution $\hat{\theta},\hat{r}$ I can answer that question by seeing if x fits in the circle of radius r.