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)

\[ K(x, x') = (x \cdot x')^p, p = 1, 2, \ldots \] 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

\[
\min \frac{1}{2} \|\theta\|^2 \text{ s.t. } y^{(i)} \theta \phi(x^{(i)}) \geq 1, i = 1, 2, \ldots n
\]

The Lagrangian for this quadratic optimization problem is:

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

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

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

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

**Dual problem:** We are trying to find the \( \alpha_i \) 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 \( \theta \), so we can get its minimum if we set the derivate equal to 0.

\[
\frac{\partial L(\theta, \alpha)}{\partial \theta} = \theta - \sum_{i=1}^{n} \alpha_i y^{(i)} \phi(x^{(i)}) = 0 \iff \theta_{\min} = \sum_{i=1}^{n} \alpha_i y^{(i)} \phi(x^{(i)})
\]

Now, if we replace this minimum \( \theta_{\min} \) in the original constraint, we get:
If these must hold true.

Dual feasibility:

Primal feasibility:

Complementary slackness

Note that if KKT conditions for primal dual optimality

Thus we have reformulated the dual as:

Thus we have reformulated the dual as:

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

KKT conditions for primal dual optimality

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

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

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

Dual feasibility: \( \alpha_i \geq 0 \)

These must hold true.

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

\[
y^{(i)} \theta \phi(x^{(i)}) - 1 = 0 \Leftrightarrow y^{(i)} \left( \sum_{j=1}^{n} \alpha_j y^{(j)} K(x^{(i)}, x^{(j)}) \right) = 1
\]
If $\alpha_i = 0$, then the $x^{(i)}$ will be not be a support vector, which means it’ll sit outside the margin so:

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

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 $\alpha_i$ values that give you the same $\theta_{\min}$?

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

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

Let’s include slack. The generalized Lagrangian is:

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

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

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

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

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

Without offset:

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

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

such that $\alpha_i \geq 0$ and $\alpha_i \leq \frac{c}{n}$
With offset:

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

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

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

Then we can build the classifier as:

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

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

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

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

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.

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

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 $$r$$. 