Statistical Machine Learning | 11 — Kernel SVM | Oct. 9, 2014

Lecturer: Prof. Homrighausen | Scribe: Miranda Fix

0 A little machine learning humor

Colonel Hilbert was one of the finest officers of his time. Mortally wounded in battle, army scientists devised a plan to clone him so that his leadership could persist throughout the war. "How is this possible? Aren't there an infinite number of parameters?" asked an army senior officer. "Don't worry" one of the army scientists replied, "we've built a Reproducing Colonel Hilbert Space." (Courtesy of Matt Wytock.)

1 Kernel SVM

Let me start with the basic idea. If the class boundary is nonlinear, we could enlarge the feature space using functions of the predictors (e.g. polynomials or interactions). However, we could end up with a huge (or even infinite) number of features this way. So instead we can use kernel SVM which is a computationally efficient approach to enlarging the feature space. Recall that the dual Lagrangian is

$$\ell_D(\gamma) = \sum_i \gamma_i - \frac{1}{2}\sum_i \sum_{i'} \gamma_i \gamma_{i'} y_i y_{i'} x_i^\top x_{i'}$$

where $x_i^\top x_{i'} = \langle x_i, x_{i'} \rangle = \sum_{j=1}^p x_{ij} x_{i'j}$ is an inner product. So SVMs depend on the covariates only through an inner product. The kernel method essentially replaces this inner product with a generalization of the inner product, i.e. a kernel $K(x_i, x_{i'})$. A kernel is a function that quantifies the similarity of two observations. We want it to be symmetric and positive definite. Note that we can write $K(x, x') = \langle \phi(x), \phi(x') \rangle$ where ϕ is the feature map, so the kernel function computes inner products in the transformed space.

Some common choices of kernels for SVMs are the **polynomial kernel** of degree d:

$$K(x, x') = (1 + \langle x, x' \rangle)^d$$

and the **radial kernel**:

$$K(x, x') = \exp(-\lambda ||x - x'||^2).$$

Notice that the radial kernel has very local behavior. If a test observation x^* is far away from training observation x_i in terms of Euclidean distance, then $K(x^*, x_i) = \exp(-\lambda ||x^* - x_i||^2)$ will be small, so x_i will not contribute much to $f(x^*)$.

The solution can then be written $\hat{f}(x) = \sum_{i=1}^N \hat{\gamma}_i y_i K(x, x_i) + \hat{\beta}_0$, with classifier $\hat{G}(x) = \text{sign}(\hat{f}(x))$ as usual.

2 Example of the polynomial kernel in action

Consider a feature space with two inputs X_1 and X_2 and a polynomial kernel of degree 2. Then

$$\begin{aligned} K(X,X') &= (1+\langle X,X'\rangle)^2 \\ &= (1+X_1X_1'+X_2X_2')^2 \\ &= 1+2X_1X_1'+2X_2X_2'+(X_1X_1')^2+(X_2X_2')^2+2X_1X_1'X_2X_2' \end{aligned}$$

Let $\phi_1(X)=1$, $\phi_2(X)=\sqrt{2}X_1$, $\phi_3(X)=\sqrt{2}X_2$, $\phi_4(X)=X_1^2$, $\phi_5(X)=X_2^2$, $\phi_6(X)=\sqrt{2}X_1X_2$.

Then we could equivalently write

$$\begin{aligned} K(X,X') &= \langle\phi(X),\phi(X')\rangle \\ &= \sum_{j=1}^{6}\phi_j(X)\phi_j(X') \\ &= 1+\sqrt{2}X_1\sqrt{2}X_1'+\sqrt{2}X_2\sqrt{2}X_2'+X_1^2X_1'^2+X_2^2X_2'^2+\sqrt{2}X_1X_2\sqrt{2}X_1'X_2' \\ &= 1+2X_1X_1'+2X_2X_2'+(X_1X_1')^2+(X_2X_2')^2+2X_1X_1'X_2X_2' \end{aligned}$$

3 More on kernel methods

When we replace the inner product with a kernel we are implicitly generating a feature map (via the eigenfunctions of the kernel) and vice versa. We are also implicitly doing regularization with a penalty on smoothness. These concepts are all equivalent.

From the regularization perspective, we can define an estimator via

$$\min \hat{\mathbb{P}}\ell_f \text{ such that } f\in\mathcal{H}_k \Rightarrow \min_{f\in\mathcal{H}_k}\hat{\mathbb{P}}\ell_f+\lambda||f||_{\mathcal{H}_k}$$

We are basically regularizing by things that look like our kernel. This is potentially an infinite dimensional optimization problem which is difficult to compute. However, if we have a Reproducing Kernel Hilbert Space (RKHS, i.e. a Hilbert space $\mathcal{H}_k$ with a reproducing kernel, $k(\cdot,\cdot)$ s.t. $f(x)=\langle k(\cdot,x),f\rangle\forall f\in\mathcal{H}_k$, whose span is dense in $\mathcal{H}_k$), then it can be shown that the solution has the form $\hat{f}(x)=\sum_{i=1}^{n}\beta_i k(x,x_i)$. This exists in finite dimensional space!

Now we can rewrite the optimization problem as something much easier:

$$\min_{\beta}\hat{\mathbb{P}}\ell_{\boldsymbol{K}}\beta+\lambda\beta^\top\boldsymbol{K}\beta \text{ where } \boldsymbol{K}=[k(x_i,x_{i'})]$$

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* New York: Springer, 2009.