CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Linear & Ridge Regression and Kernels

Lecturer: Michael I. Jordan *Scribes: Dave Latham*

1 Kernel Definitions Reviewed

Let us review the definition of a kernel function.

The definition given before is that a function $K(x, y)$ is a *kernel function* if

$$\int K(x, y)g(x)g(y)dxdy \geq 0$$

for all functions g in L_2. Also, by Mercer's theorem we have

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

for some $\Phi : X \to H$, where H is some (possibly high or even infinite dimensional) inner-product space. Furthermore, for any finite collection of points, forming a Gram matrix K, we have

$$x^T K x \geq 0$$

We haven't proven the equivalence of any of these definitions, but we will use them for now.

2 Linear Regression

We now take a look at linear regression in order to see how the XX^T matrix shows up, allowing us to use a kernel function instead. Note that we are used to using seeing the product $X^T X$ as a sufficient statistic, but we need XX^T instead.

For data we are given a set of n tuples (x_i, y_i), $x_i \in \Re^d, y_i \in \Re$. We will use matrix notation:

$$X = \begin{pmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_n^T — \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

We wish to fit a model in which $y = X\beta + \varepsilon$. We will assume that ε has a Gaussian distribution with zero mean and variance σ^2. Setting the derivative of the log likelihood to zero then produces the normal equations for the maximum likelihood estimate of β,

$$(X^T X)\widehat{\beta}_{ML} = X^T y$$

This suggests that $X^T X$ and $X^T y$ are the sufficient statistics for β. (To estimate β, one can calculate these values and throw away the original data.) The dimensionality of X is $n \times d$, so $X^T X$ is $d \times d$, and $X^T y$ is

$d \times 1$. Neither of these depend on n, so the dimension of the sufficient statistic does not grow as the data grows. (This is the spirit of a *parametric* method.) This is generally a good thing. The amount of data is quite often larger than its dimension, and it also allows one to set aside a fixed amount of space to store the sufficient statistics regardless of how much data is given.

However, the Gram matrix $K = XX^T$ of all inner products of pairs of x values is $n \times n$. This is *non-parametric*. It grows with the data.

Let us suppose that $(X^TX)^{-1}$ exists. Then we have,

$$\begin{aligned} \widehat{\beta}_{ML} &= (X^TX)^{-1}X^Ty \\ &= X^TX(X^TX)^{-2}X^Ty \\ &\triangleq X^T\alpha \end{aligned}$$

where $\alpha = X(X^TX)^{-2}X^Ty$. So $X^T\alpha = \sum_i \alpha_i x_i$.

Now, if we want to predict the y values from X values, we have

$$\widehat{y}_{ML} = x\widehat{\beta}_{ML} = XX^T\alpha$$

Aha! There's the Gram matrix XX^T (which would suffice if someone gave us α).

2.1 Ridge Regression

Now we will turn to ridge regression (also known as regularized regression) which is a slight generalization of linear regression. It can be viewed in a couple of ways.

- From a frequentist perspective, it is linear regression with the log-likelihood penalized by a $-\lambda\|\beta\|^2$ term. ($\lambda > 0$)
- From a Bayesian perspective, it can be viewed as placing a prior distribution on β:

  $$\beta \sim N(0, \lambda^{-1})$$

  and computing the mode of the posterior.

In either case, ridge regression favors smaller β values. The estimate can be obtained by minimizing the following,

$$\widehat{\beta}_{MAP} = \arg\min_{\beta} \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Some calculus brings us equations similar to the normal equations:

$$(X^TX + \lambda I)\widehat{\beta}_{MAP} = X^Ty$$

Note that the first factor, $(X^TX + \lambda I)$, is now always invertible because we added a positive diagonal matrix. Also note that X^TX and X^Ty are still the sufficient statistics for β. Now we have,

$$\widehat{\beta}_{MAP} = (X^TX + \lambda I)^{-1}X^Ty$$

Alternatively,

$$(X^TX + \lambda I)\widehat{\beta}_{MAP} = X^Ty$$

$$\begin{aligned}
X^T X\widehat{\beta}_{MAP} + \lambda\widehat{\beta}_{MAP} &= X^T y \\
\lambda\widehat{\beta}_{MAP} &= X^T y - X^T X\widehat{\beta}_{MAP} \\
&= X^T(y - X\widehat{\beta}_{MAP}) \\
\widehat{\beta}_{MAP} &= \lambda^{-1}X^T(y - X\widehat{\beta}_{MAP}) \\
&\triangleq X^T\alpha
\end{aligned}$$

where $\alpha = \lambda^{-1}(y - X\widehat{\beta}_{MAP})$. Just as before with linear regression, we could use the Gram matrix if someone gave us α. Now, from the definition of α we have,

$$\begin{aligned}
\lambda\alpha &= y - X\widehat{\beta}_{MAP} \\
&= y - XX^T\alpha \\
XX^T\alpha + \lambda\alpha &= y \\
(XX^T + \lambda I)\alpha &= y \\
\alpha &= (XX^T + \lambda I)^{-1}y \\
&= (K + \lambda I)^{-1}y
\end{aligned}$$

So we can do ridge regression based only on K and y. In particular, we can throw away the X data once we have K. To predict a new y value from a new x value we have,

$$\begin{aligned}
\widehat{y}_{new} &= \widehat{\beta}^T x_{new} \\
&= \sum_i \alpha_i x_i^T x_{new}
\end{aligned}$$

Again, x only enters into the prediction only via inner products.

Why would we want to keep around an $n \times n$ matrix K instead of the original $n \times d$ matrix X, or even the $d \times d$ matrix $X^T X$? We wouldn't if we were doing actual linear or ridge regression. However, we can now replace the matrix K with a kernel function to do ridge regression in some other space, giving us a non-linear (and non-parametric) regression! We can do this in cases where there is no finite dimensional sufficient statistic. So β can now even be infinite dimensional.

3 Closure Properties of Kernel Functions

There exists an algebra of kernel functions that allows us to build up more complex kernels from simpler ones.

Let $K_1(x, y)$ and $K_2(x, y)$ be kernel functions (choose your favorite definition). Then the following are all kernel functions:

(a) $K_1(x, y) + K_2(x, y)$

(b) $\alpha K_1(x, y)$ $(\alpha > 0)$

(c) $K_1(x, y)K_2(x, y)$

(d) $f(x)f(y)$ $\forall f$

(e) $x^T Ay$ for positive semidefinite A

Some proof intuition:

For (a), sums of positive semidefinite matrices or functions are also positive semidefinite. Parts (b) and (d) are trivial. Part (c) is called a Hadamard product, its proof is tricky (look at a Kronecker product). For part (e), consider Mercer's theorem:

$$x^T Ay = \langle \Phi(x), \Phi(y) \rangle$$

Using Cholesky decomposition gives us $x^T Ay = x^T LL^T y$, so we can define $\Phi(x) = L^T x$.

3.1 Examples

Using these closure properties, we can first design simple, basis kernels, then combine them to create more complex kernels. In particular, let us look at the polynomial kernel which can be built up using these properties:

$$K(x, y) = (\langle x, y \rangle + R)^d = \sum_{s=0}^{d} \begin{pmatrix} d \\ c \end{pmatrix} R^s \langle x, y \rangle^{d-s} \qquad (R \geq 0)$$

by the binomial theorem. This can be seen as a weighted sum of monomials, where the weights are $\begin{pmatrix} d \\ c \end{pmatrix} R^s$, and the monomials are $\langle x, y \rangle^{d-s}$. Consider the simpler case, where R is zero.

$$\langle x, y \rangle = (x_1 y_1 + x_2 y_2 + \ldots + x_m y_m)^r$$

When one distributes the product out, each term will be a combination of sum $x_i y_i$ from each of r factors. This gives a sum of monomials each of the same degree, r. We can write it as:

$$\langle x, y \rangle = x_1^{i_1} y_1^{i_1} x_2^{i_2} y_2^{i_2} \ldots x_m^{i_m} y_m^{i_m}$$

where $\sum_j i_j = r$. This gives us the intuition to come up with a function Φ related to the kernel by Mercer's theorem.

$$\Phi(x) \triangleq \begin{pmatrix} \vdots \\ x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m} \\ \vdots \end{pmatrix}$$

where $\Phi(x)$ is a long vector with a coordinate for every monomial such that $\sum_j i_j = r$. The more general polynomial kernel (with $R > 0$) would have coordinates with additional weights associated to them as well, as given by the binomial theorem. Note that the dimensionality of the vector is huge, and so it would be very expensive to actually form them in order to take inner products. It is much cheaper to just use the kernel function $K(x, y) = (\langle x, y \rangle + R)^d$ without having to visit that high dimensional space.

In our study of kernels it will turn out to be useful to go in both directions of Mercer's theorem, from kernel function to feature space and vice versa.

Another interesting and widely used kernel is the Gaussian kernel:

$$K(x, y) = exp\{-\frac{1}{2}\|x - y\|^2\}$$

One can show that it is a kernel function because it is the limit of a series of polynomials, and limits of kernels are also kernels. Its feature space is infinite dimensional!