CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Regression, the SVD, and PCA

Lecturer: Michael I. Jordan *Scribes: Brian Vogel*

1 Support Vector Regression

Before discussing support vector (SV) regression, let us first consider a few possible loss functions for the linear regression problem

$$y_i = w^T \phi(x_i) + \epsilon_i$$

Least squares regression minimizes the quadratic loss:

$$\min_w \sum_{i=1}^n \epsilon_i^2$$

Ridge regression adds a penalization term for the $\mathcal{L}_2$ norm of the parameter vector w:

$$\min_w \sum_{i=1}^n \epsilon_i^2 + \lambda w^T w$$

Lasso regression adds a penalization term for the $\mathcal{L}_1$ norm of the parameter vector:

$$\min_w \sum_{i=1}^n \epsilon_i^2 + \lambda \sum_{j=1}^n |w_j|$$

For SV regression, we will make use of the ϵ insensitive loss, which does not penalize errors below some $\epsilon \geq 0$ Figure 1 shows the ϵ insensitive loss function. We introduce slack variables ξ_i and $\tilde{\xi}_i$. We have the following optimization problem:

$$\min_w w^T w + C \sum_{i=1}^n (\xi_i + \tilde{\xi}_i)$$

such that

$$y_i - w_T \phi(x_i) - b \leq \epsilon + \xi_i$$
$$w^T \phi(x_i) + b - y_i \leq \epsilon + \tilde{\xi}_i$$
$$\tilde{\xi}_i \geq 0, \xi_i \geq 0$$

Here, C controls the amount of penalization for points lying outside the ϵ tube.

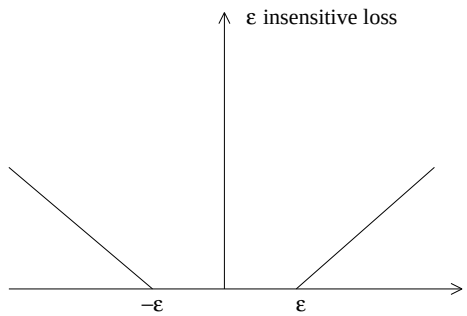

Figure 1: The ϵ insensitive loss function for SV regression.

1.1 Dual Problem

Writing the Lagrangian, we get the following optimization problem:

$$\max_{\alpha_i, \tilde{\alpha}_i} \sum_i (\hat{\alpha}_i - \alpha_i) y_i - \epsilon \sum_i (\hat{\alpha}_i + \alpha_i) - \frac{1}{2} \sum_{i,j} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) \langle \phi(x_i), \phi(x_j) \rangle$$

such that

$$0 \leq \alpha_i, \tilde{\alpha_i} \leq C$$
$$\sum_i (\tilde{\alpha} i - \alpha_i) = 0$$

Note that this is a quadratic program. Note also that this is kernalized, so we can use $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ instead of explicitly computing the inner product in the feature space.

1.2 KKT conditions

$$\alpha_i(\langle w, \phi(x_i) \rangle + b - y_i - \epsilon + \xi_i) = 0$$
$$\tilde{\alpha_i}(\langle w, \phi(x_i) \rangle + b - y_i - \epsilon + \tilde{\xi_i}) = 0$$

From the KKT conditions, we also have that

$$\xi_i = \tilde{\xi_i} = 0$$
$$\alpha_i = \tilde{\alpha_i} = 0$$
$$(\alpha_i - C)\xi_i = 0$$
$$(\tilde{\alpha_i} - C)\tilde{\xi_i} = 0$$

Note that if the ξ_i are strictly positive, then $\alpha_i = C$.

1.3 Kernel PCA

Question: Is finding eigenvectors a convex problem? Ans: No, but eigenvector problems can be efficiently solved nonetheless. PCA is an eigenvector problem.

We would like to perform PCA in the feature space. Recall that for PCA, we need to find the eigenvectors of $X^T X$.

1.3.1 The Singular Value Decomposition (SVD)

A symmetric matrix has real eigenvalues and its eigenvectors can be chosen to be orthonormal. Any symmetric matrix A can be factorized as $A = M\Lambda M^T$. Here, M is an orthogonal matrix containing the eigenvectors of A as its columns. Λ is the diagonal matrix containing the eigenvalues of A.

To see this, right-multiply by M to obtain

$$AM = M\Lambda$$
$$Ax = \Lambda x$$

where x is a column of M.

Now if A also happens to be psd, then all of its eigenvalues will be non negative (i.e., $\Lambda \geq 0$).

Any m by n matrix A can be factored into the *singular value decomposition*

$$A = U\Sigma V^T$$

where the columns of the m by m orthogonal matrix U are the eigenvectors of AA^T. The columns of the n by n orthogonal matrix V are the eigenvectors of A^TA. The m by n diagonal matrix Σ has the singular values along its diagonal. The singular values, σ_i, are all nonnegative, and are the square roots of the eigenvalues of both AA^T and A^TA.

Consider the design matrix X, with rows consisting of the feature vectors.

Let $X^T = U\Sigma V^T$. Recall that our kernel matrix is given by

$K = XX^T = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$

So, the columns of V are the eigenvectors of K. Moreover, the elements of Σ^2 are eigenvalues of K. What is the relationship between U and XX^T?

$$C \triangleq XX^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T$$

So, $col(U)$ are the eigenvectors of C and $diag(\Sigma^2)$ are the eigenvalues of C. Note that K and C share the same eigenvalues. We have that $U\Sigma = X^TV$. So, the j'th column of U can be expressed as

$$u_j = \lambda^{-1/2} X^T V = \sum_{i=1}^{n} \alpha_i^j \phi(x_i), \alpha^j = \lambda_j^{-1/2} v_j$$

So, we only need to calculate the eigenvectors of one of these matrices. Note also that the eigenvectors of the correlation matrix can be expressed as a linear combination of the feature vectors, $\phi(x_i)$.

2 Principal Components Analysis (PCA)

The idea of PCA is to find the direction such that the variance of the data projected onto that direction is maximized. Let's suppose that we are given centered data x_i. That is, $\sum_i x_i = 0$ If we were not given centered data, then the first step would be to translate the data to the mean.

The variance of the projected values is maximized:

$\max_{||w=1||} Var(w^T x)$
$\Leftrightarrow \max_{||w1||=1} E(w^T x - w^T \mu)^2$
$\Leftrightarrow \max_{||w||=1} E(2^T x - w^T \mu)(x^T w - \mu^T w)$
$\Leftrightarrow \max_{||w||=1} w^T E(x-\mu)(x-\mu)^T w$
$\Leftrightarrow \max_{||w||=1} w^T \Sigma W$

Which is quadratic with the constraint that $||w|| = 1$.

The eigenvector with the biggest eigenvalue is the solution. To see this, write the Lagrangian.

$\mathcal{L}(w^T, \lambda) = w^T \Sigma w + \lambda(1 - w^T w)$

$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow 2\Sigma w - 2\lambda w = 0$

$\Rightarrow \Sigma w = \lambda w$

So the eigenvector with the biggest eigenvalue is the solution.