UNIVERSITY OF CALIFORNIA, BERKELEY

EE:127 Connecting Least Squares, SVD, Regularization, and PCA

Oscar Ortega

July 16, 2021

1 TIKHONOV REGULARIZATION

What if its the case that we know that our 'x' is not necessarily something that is close to 0? Well in this case we use Tikhonov Regularizaition:
Recall from last lecture:

$$x^*_{\text{Tikhonov}} = argmin_x \|W_1(Ax-b)\|_2^2 + \|W_2(x-x_0)\|_2^2$$

eg: if we allow $W_1 = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ s.t $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ then our regularizer would take more importance on the first few measurements in our A matrix. would puts more weight on measurements that apriori are deemed to be more 'accurate' so to speak. Moreover, if defined $W_2 = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ s.t $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, this would allow us to have individual regularizers the individual components of our vector x. Can compute a similar value to x^* by computing the derivative and minimizing.

What if we allow $W_1 = I$ and $W_2 = \lambda I$?
This reduces to the following:

$$x^*_{\text{Tikhonov}} = argmin_x \|I(Ax-b)\|_2^2 + \|\lambda I(x-x_0)\|_2^2$$

$$= \|Ax-b\|_2^2 + \lambda\|(x-x_0)\|_2^2 = x^*_{Ridge}$$

2 SVD $\iff$ LEAST SQUARES $\iff$ PCA

Recall that in Least Squares we are trying to minimize the following:

$$\min_w \|Xw - y\|_2^2$$

$$w_{LS}^* = (X^T X)^{-1} X^T y$$

Furthermore, recall any A can be decomposed into $U\Sigma V^T$ where U and V are orthonormal matrices, where $U \in \mathbb{R}^{m,m}, V \in \mathbb{R}^{n,n}, and \Sigma \in \mathbb{R}^{n,m}$
Note that if we plug the SVD decomposition to the projection formula, $X = U\Sigma V^T$ we get the following:

$$\begin{aligned} w^* &= (V\Sigma U^T U\Sigma V^T)^{-1} X^T y \\ &= (V\Sigma^2 V^T)^{-1} V\Sigma U^T y \\ &= V\Sigma^{-2} V^T V\Sigma U^T y \\ &= V\Sigma^{-1} U^T y \end{aligned}$$

In a scalar setting: this reduces to finding the optimal w for $y = xw + n$ where the x is known.
Now, consider $X = \begin{vmatrix} x_1 & x_2 & .... & x_n \end{vmatrix}^T = U\Sigma V^T$ and assume we seek to use ridge regression to find

$$z_{\text{ridge}}^* = argmin_z \|Xz - y\|_2^2 + \lambda \|z\|_2^2$$

Now, let $z = Vw$:
This would imply our minimization becomes the following:

$$argmin_w \|XVw - y\|_2^2 + \lambda \|Vw\|_2^2$$

$$argmin_w \|XVw - y\|_2^2 + \lambda \|w\|_2^2$$

Recall: $\|Vw\| = \|w\|$ because in SVD decomposition, V is an orthonormal matrix. We substitute the value $X = U\Sigma V^T$ and after some derivations arive at the following:

$$\begin{aligned} w_{ridge}^* &= ((XV)^T XV + \lambda I)^{-1} (XV)^T y \\ &= (\Sigma^2 + \lambda I)^{-1} \Sigma U^T y \\ &= (\Sigma^2 + \lambda I)^{-1} \Sigma U^T y \end{aligned}$$

and recall that because $z = Vw^*$

$$\rightarrow z_{ridge}^* = V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y$$

Key point: Note the structure of $(\Sigma^2 + \lambda I)^{-1}\Sigma$:

$$= \begin{bmatrix} \frac{1}{\sigma_1 + \lambda} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\sigma_n^2 + \lambda} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma_1}{\sigma_1+\lambda} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\sigma_n}{\sigma_n^2+\lambda} \end{bmatrix}$$

Consider the effect of the regularizer λ on the diagonal entries of the matrix:

$$\lim_{\lambda} \to \infty : \frac{\sigma}{\sigma^2+\lambda} \to 0$$

$$\lim_{\lambda} \to 0 : \frac{\sigma}{\sigma^2+\lambda} \to \frac{1}{\sigma}$$

As one can see, increasing the value of our regularizer λ will first begin 'zeroing' out the smaller singular values of our matrix, in essence, providing a 'soft' dimensionality reduction to our data sort of like PCA.

Now, Lets contrast this with our PCA based least squares minimization:

$$min_w \|XV_k w - y\|_2^2$$

where $X \in \mathbb{R}^{m,n} : V_k \in \mathbb{R}^{n,k} : w \in \mathbb{R}^k$

$$w^*_{PCA} = ((XV_k)(XV_k)^T)^{-1}(XV_k)^T y$$

$$= (V_k^T V \Sigma^2 V^T V_k)^{-1} V_k^T V \Sigma V^T y$$

Now similarly to before, we use the decomposition $X = U\Sigma V^T$

$$= (\Sigma_k^{-1}\Sigma_k^2)^{-1} U_k^T y$$

Note that due to the 'hard' sort of cutoff by the matrix V_k our solution provides a dimensionality reduction similar to least squares, where we simply toss the dimensions whose singular values are not one of the k largest.

3 Min Norm Problems $\Longleftrightarrow$ Least Squares

Related to the concepts of kernels and why we care about them (will learn more in CS189) Recall once again our formulations for least squares and minimum norm problems:

$$w^*_{LS} = argmin_w \|Xw - y\|_2^2 = (X^T X)^{-1} X^T y$$

$$w^*_{MN} = argmin_x \|w\|_2^2 \text{ subject to: } (Xw = y) = X^T (XX^T)^{-1} y$$

As noted in lecture before, we can recall that Least Squares is usually for overdetermined system, and the reverse for min norm, however, these are a lot more related if we consider how they are related under ridge regression.

$$w^*_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

lets express this in a seperate perspective.
Claim: $w^*_{ridge} \in \mathscr{R}(X^T) \iff$ 'is a linear combination of the data points'
Proof:

$$X^T X w + \lambda w = X^T y$$

$$\lambda w = X^T y - X^T X w = X^T (y - X w)$$

$$w = \frac{1}{\lambda}(X^T y - X^T X w) = \frac{X^T}{\lambda}(y - X w)$$

$$\rightarrow w = X^T v$$

This implies we can find v in the basis of X^T to solve for w^*_{Ridge} How to fnd v?

$$X^T X X^T v + \lambda X^T v = X^T y$$

or:

$$(X X^T + \lambda X^T) v = y$$

$$\rightarrow v = (X X^T + \lambda I)^{-1} y$$

$$w = X^T (X X^T + \lambda I)^{-1} y$$

note:

$$X X^T \in \mathbb{R}^{m,m}$$

Note the similarity with te minimum norm solution! The ridge paramater thus yields a continuum between the minimum norm solution and the least squares solutions to minimization problems.

$$x^*_{LS} : \lambda_{\text{small}} \iff x^*_{MN} : \lambda_{\text{large}}$$

Note how $X X^T$inner product of the data points for a data matrix X. These are known as kernel methods and might come to a more efficient computation.

4 Total Least Squares

Consider a system where we not only have dependent noisy measurements y and our independent noisy measurements X. In other words we now have the following system:

$$\left[X + \tilde{X}\right] w = \left[y + \tilde{y}\right]$$

We define the following: $\tilde{X} + X = \hat{X}$ and $\tilde{y} + y = \hat{y}$

Where $\tilde{X}$ and $\tilde{y}$ are the errors in the measurements and $\hat{X}$ and $\hat{y}$ are the our best estimates for the true system. Our goal thus becomes to minimize $\tilde{X}$ and $\tilde{y}$. This can be described as the following minimization:

$$min \| \left[\tilde{X} | \tilde{y}\right] \|_F \text{ subject to } \hat{X} \hat{w} = \hat{y}$$

Note: $\|A\|_F = tr(A^T A) = \sum_{i,j} |A_{i,j}|^2$, $y \in R^m$ and that $\left[\bar{X}|\bar{y}\right] \in \mathbb{R}^{m,n+1}$ and $m > n$
This constraint equation can be rewritten as the following:

$$\left[X + \tilde{X}|y + \tilde{y}\right] \begin{bmatrix} w \\ -1 \end{bmatrix} = 0 : 0 \in \mathbb{R}^m.$$

For there to be a unique solution for w, it must be that the matrix $\left[\hat{X}|\hat{y}\right]$ must have rank exactly n; however it has $n+1$ columns. Thus, we want to find the smallest perturbation to the matrix $\left[X|y\right]$ with rank $n+1$ such that the resulting matrix $\left[\hat{X}|\hat{y}\right]$ has rank n. For this we want to minimize the *total least squares cost* $min\|\left[\tilde{X}|\tilde{y}\right]\|_F$, which is the sum of the squares of all the perturbations.
The Eckart-Young theorem allows us to know that this perturbation can be calculated by dropping the smallest singular value in the SVD of the matrix $\left[X|y\right]$. We won't be proving this in this class (for now), but maybe later.