CS 189 Introduction to Machine Learning
Spring 2018 DIS6

1 OLS, Ridge Regression, TLS, PCA and CCA

In this discussion, we will review several topics we have learnt so far. We emphasize on their basic attributes, including the objective functions, the generative models as well as the explicit form of solutions. You will also learn the connection and distinction between those methods.

(a) What problem does each of the methods trying to solve? What are their objective functions? Can you write out their solutions in a closed form? What are the probablistic perspectives for OLS, ridge regression and total least squares?

Solution: OLS assumes a linear model with noisy $\mathbf{y}$, its objective function is: $\arg\min_w ||\mathbf{X}\mathbf{w}-\mathbf{y}||^2$. Its underlying generative model is $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, where $\epsilon \sim N(0, \mathbf{I})$. The closed form solution is $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Ridge regression adds extra regularization terms on top of OLS. In the generative model aspect, it puts an Gaussian prior on $\mathbf{w}$. I.e. $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, where $\epsilon \sim N(0, \lambda\mathbf{I})$ and $\mathbf{w} \sim N(0, \mathbf{I})$. The corresponding optimization problem is $\arg\min_w ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2$. The closed form solution is $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

TLS is another modification of OLS by no longer assuming we have a noiseless $\mathbf{X}$. To be specific, the generative model is $\mathbf{y} + \epsilon_y = (\mathbf{X} + \epsilon_x)\mathbf{w}$, where $\epsilon_x, \epsilon_y \sim N(0, \mathbf{I})$. The objective in this case is: $\arg\min_{\epsilon_x,\epsilon_y} ||[\epsilon_x \quad \epsilon_y]||_F^2$, subject to $(\mathbf{X} + \epsilon_x)\mathbf{w} = \mathbf{y} + \epsilon_y$. And the closed form solution is $(\mathbf{X}^T\mathbf{X} - \sigma_{d+1}^2\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$, where σ_{d+1} is the least singular value of $[\mathbf{X} \quad \mathbf{y}]$.

PCA deals with the dimension reduction of $\mathbf{X}$. For the one dimensional case, the objective is to $\arg\max_{||\mathbf{u}||=1} ||\mathbf{X}\mathbf{u}||^2$. The closed form solution is given by first do SVD on $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, then $\mathbf{u} = \mathbf{V}_1$, where $\mathbf{V}_1$ is the first column of $\mathbf{V}$.

CCA models relationship between two point sets $\mathbf{X}$ and $\mathbf{Y}$. The objective is $\arg\max_{\mathbf{u},\mathbf{v}} \rho(X_{rv}^T\mathbf{u}, Y_{rv}^T\mathbf{v})$. The solution is: $\mathbf{u} = \mathbf{W}_x\mathbf{D}_x\mathbf{u}_d$ and $\mathbf{v} = \mathbf{W}_y\mathbf{D}_y\mathbf{v}_d$, where the whitening step gives: $\mathbf{W}_x = \mathbf{V}_x\Sigma_x^{-1}\mathbf{V}_x^T$ and $\mathbf{V}_x, \Sigma_x$ given by the SVD $\mathbf{X} = \mathbf{U}_x\Sigma_x\mathbf{V}_x^T$. The whitened matrix has the property of $(\mathbf{X}\mathbf{W}_x)^T(\mathbf{X}\mathbf{W}_x) = \mathbf{I}$ Similar results hold for $\mathbf{W}_y$. The decorrelation step gives $\mathbf{D}_x = \mathbf{U}_r$ and $\mathbf{D}_y = \mathbf{V}_r$, where $\mathbf{X}_w^T\mathbf{Y}_w = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T$. After decorrelation, $(\mathbf{X}_w\mathbf{D}_x)^T(\mathbf{Y}_w\mathbf{D}_y) = \mathbf{I}$

(b) Suppose you have a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and vector $\mathbf{y} \in \mathbb{R}^{n\times 1}$. Use PCA to compute the first k principal components of $[\mathbf{X} \quad \mathbf{y}]$. Show how this solution would relate to a TLS solution to the problem.

Solution: Using what we've learned in PCA, we know that the SVD decomposition of the matrix becomes $[\mathbf{X} \quad \mathbf{y}] = \mathbf{U}\Sigma\mathbf{V}$. From there, we pick the resulting k vectors such that we start at $\mathbf{u} = \mathbf{V}_1$, where $\mathbf{V}_1$ is the first column of $\mathbf{V}$ and go up to $\mathbf{u} = \mathbf{V}_k$, where $\mathbf{V}_k$ is the kth column of $\mathbf{V}$.

How is the TLS solution found? TLS minimizes $|| \begin{bmatrix} \boldsymbol{\epsilon}_X & \boldsymbol{\epsilon}_y \end{bmatrix} ||_F^2$, subject to the constraint of:

$$\begin{bmatrix} \mathbf{X} + \boldsymbol{\epsilon}_X & \mathbf{y} + \boldsymbol{\epsilon}_y \end{bmatrix} \begin{bmatrix} \mathbf{w} & -1 \end{bmatrix}^\top = \mathbf{0}$$

Since we want to find the minimum noise, that after perturbation, has at least one zero eigenvalue. By Eckhart-Young theorem, we know that the noise perturbed $\begin{bmatrix} \mathbf{X} + \boldsymbol{\epsilon}_X & \mathbf{y} + \boldsymbol{\epsilon}_y \end{bmatrix}$ should be the best rank-d approximation to the original matrix $[\mathbf{X} \quad \mathbf{y}]$. Taking a step back, we see that TLS seeking a rank-d best approximation has the same objective function as PCA when reducing the dimensionality to d.

(c) Among OLS, Ridge and TLS, what method would you use when: (1) observation $\mathbf{X}$ is noisy (2) $\mathbf{X}$ is not noisy and $d >> n$ (3) $\mathbf{X}$ is not noisy and $d << n$?

Solution: (1) When $\mathbf{X}$ is noisy, TLS is preferred, since it models the noise explicitly in the model. It can recover the latent structure by explicitly removing the noise. However, note that when doing out of sample prediction, TLS is better only when the testing $\mathbf{X}$ is noiseless or has less noise than the training $\mathbf{X}$.

(2) In this case, the problem is under-constraint and ridge regression fits best. TLS assumes that the $\mathbf{X}$ has noise. The matrix inversion step in OLS ($(\mathbf{X}^\top\mathbf{X})^{-1}$) will be unstable due to $d > n$. Thus TLS and OLS do not fit this scenario.

(3) In this case, OLS fits the best, if there are no linear relationship among the features. The problem is over-constraint and generally the matrix inversion step ($(\mathbf{X}^\top\mathbf{X})^{-1}$) is stable. However, if $\mathbf{X}$ does not have full column rank, the matrix inversion will fail. In that case, ridge regression with small regularization is preferred over OLS.

(d) How do OLS, ridge and TLS interact with the matrix $\mathbf{X}^\top\mathbf{X}$ in the closed form solutions? What are the eigenvalues of the matrix being inverted in the closed form solutions? Do you have any intuitions of why the eigenvalues changes in those manners?

Solution: OLS invert the matrix $\mathbf{X}^\top\mathbf{X}$, ridge invert the regularized matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ and TLS invert the $\mathbf{X}^\top\mathbf{X} - \sigma_{d+1}^2\mathbf{I}$ matrix. Thus those three methods either add, subtract an identity matrix. For the OLS case, it could be interpreted as adding $0\mathbf{I}$.

The eigenvalues in the ridge regression adds λ to the eigenvalues of the OLS case, while the TLS subtract σ_{d+1}^2.

Ridge regression regularizes the model by having an extra $\lambda||\mathbf{w}||^2$ loss. It turns out that it is also helping the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ to stay away from zero, and thus improving the numerical stability. TLS on the other hand, seems to decrease the numerical stability. However, as shown below, it could be think of as removing the components generated by noise.

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\right] \\
&\mathbb{E}\left[(\mathbf{X}^* + \epsilon_x)^T(\mathbf{X}^* + \epsilon_x)\right] \\
&\mathbb{E}\left[\mathbf{X}^{T*}\mathbf{X} + \mathbb{E}\left[\epsilon_x^T\epsilon_x\right]\right] \\
&\mathbb{E}\left[\mathbf{X}^{T*}\mathbf{X} + \sigma_{noise}^2\mathbf{I}\right]
\end{aligned} \tag{1}$$

In this case, we know that even the smallest eigenvalues is not zero, which comes from the noise. The TLS explicitly remove that noise by subtracting the σ_{d+1}^2

(e) Suppose you have a multi-variate regression problem, i.e. the feature matrix is $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the regression target is $\mathbf{Y} \in \mathbb{R}^{n \times q}$ and $q > 1$. We know a prior that the number of regression targets is large and there are strong correlations between the multiple regression targets. For example, consider you have $n = 100$ samples. Each example has $p = 500$ features, and there are $q = 1000000$ regression targets.

There are two approaches you can solve the problem. The first approach is treat the multi-variate regression problem as q independent ridge regression problems. The second one is that first compute the CCA between $\mathbf{X}$ and $\mathbf{Y}$, which gives two projection matrices $\mathbf{U}_k$ and $\mathbf{V}_k$, then use q independent ridge regressions to fit $\mathbf{Y}_c \equiv \mathbf{Y}\mathbf{V}_k$ from $\mathbf{X}_c \equiv \mathbf{X}\mathbf{U}_k$, i.e. solve for $\mathbf{W}$ that satisfy $\mathbf{X}_c\mathbf{W} \approx \mathbf{Y}_c$. The final predictor is given by: $\mathbf{Y}_{predict} = \mathbf{X}(\mathbf{U}_k\mathbf{W}\mathbf{V}_k^{-1})$. What's the pros and cons of each approach?

Solution: The ridge regression is assuming that each of the fitting target is independent. On the other hand, the CCA approach has taken into account of the potential correlation among the fitting targets. Having taken advantage of the target correlation, CCA might have higher statistical efficiency than the set of independent ridge regressions. For example, the ridge regression use $n = 100$ examples to fit each target, with $p = 500$ dimensional features. It will results in severe overfitting in general. However, if we assume that the regression targets have large correlation with each other, conceptually we are using $nq = 10E8$ examples to fit each target, with the same number of features. Obviously CCA in this case will be much better than the first approach.

The independent ridge regression assumes identity Gaussian noise on each of the fitting target. On the other hand, the noise model for the CCA approach depends on the data, influenced via the computed matrices $\mathbf{U}_k$ and $\mathbf{V}_k$. It is conceptually much harder to make sense of this noise model and usually people use a noise model that does not depend on the data.