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?

(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.

(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 $d << n$?

(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?

(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?

