CS 189 Introduction to Machine Learning
Fall 2018 DIS5

1 Canonical Correlation Analysis

Assume that you have a database of images of words handwritten by two different people. $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ corresponds to the dataset of handwriting 1 and handwriting 2 respectively. We will think of the databases as being composed of n samples of random variables X and $Y \in \mathbb{R}^d$. Your goal is to use machine learning to build a text recognition of word images. Because the feature size of these images is large, your first step will be to find a lower dimensional representation.

(a) Explain why you would want to consider using CCA in this problem.

(b) Assume that X and Y are zero-mean. Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, what is the correlation coefficient of the embedding $\mathbf{u}^\top X$ and $\mathbf{v}^\top Y$? Correlation coefficient between two scalar random variables P and Q is computed by:

$$\rho(P, Q) = \frac{cov(P, Q)}{\sigma_P \sigma_Q}.$$

How do we estimate this quantity using our data in $\mathbf{X}$ and $\mathbf{Y}$?

(c) Assume that the features of matrix $\mathbf{X}$ are rescaled by a constant to have values between -1 and 1. How does this change the correlation coefficient?

(d) CCA aims to find the projection vectors $\mathbf{u}$ and $\mathbf{v}$ that maximize the correlation coefficient. Using the singular value decomposition of $\mathbf{X}$ and $\mathbf{Y}$, show that the maximization problem is equivalent to

$$\max_{\substack{\|\mathbf{a}\|_2=1 \\ \|\mathbf{b}\|_2=1}} \mathbf{a}^\top \mathbf{M} \mathbf{b},$$

for some matrix $\mathbf{M}$. Then using the singular value decomposition of $\mathbf{M}$, solve the optimization problem and find the optimal $\mathbf{u}$ and $\mathbf{v}$ in terms of the SVD of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{M}$.

2 Connections between OLS, Ridge Regression, TLS, PCA, and CCA

We will review several topics we have learned so far: ordinary least-squares, ridge regression, total least squares, principle component analysis, and canonical correlation analysis. We emphasize their basic attributes, including the objective functions and the explicit form of their solutions. We will also discuss the connections and distinctions between these methods.

(a) What are the objective functions and closed-form solutions to OLS, ridge regression, and TLS? How do the probabilistic interpretations vary?

(b) Consider the matrix inversion in the solution to OLS, ridge regression, and TLS. How do the eigenvalues compare to those of the matrix $\mathbf{X}^\top\mathbf{X}$?

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

(d) 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 $q > 1$ is potentially very large. We believe that 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 is to first compute the CCA between $\mathbf{X}$ and $\mathbf{Y}$, which gives two projection matrices $\mathbf{U}$ and $\mathbf{V}$, then use q independent ridge regressions to fit $\mathbf{Y}_c \equiv \mathbf{YV}$ from $\mathbf{X}_c \equiv \mathbf{XU}$, i.e. solve for weights $\mathbf{W}$ that satisfy $\mathbf{X}_c\mathbf{W} \approx \mathbf{Y}_c$. The final predictor is given by: $\mathbf{Y}_{predict} = \mathbf{XUW}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top$. What are the advantages and disadvantages of each approach?