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.

**Solution:** These two feature matrices share similar information. By maximizing the correlation, CCA helps to remove redundant information between the two views. At the same time, CCA is useful in eliminating noise specific to only one view. For example, it will ignore flourishes common in one person's handwriting but not the other.

(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 $\mathbf{Y}$?

**Solution:** The question is asking for $\rho(\mathbf{u}^\top X, \mathbf{v}^\top Y)$ for random variables $X$ and $Y$.

Writing out the algebra (notice that $\mathbf{u}^\top X$ and $\mathbf{v}^\top Y$ are zero mean), we have

$$\begin{aligned}
\rho(\mathbf{u}^\top X, \mathbf{v}^\top Y) &= \frac{\mathbb{E}[\mathbf{u}^\top X \mathbf{v}^\top Y]}{\sqrt{\mathbb{E}[\mathbf{u}^\top X \mathbf{u}^\top X]\mathbb{E}[\mathbf{v}^\top Y \mathbf{v}^\top Y]}} \\
&= \frac{\mathbb{E}[\mathbf{u}^\top X Y^\top \mathbf{v}]}{\sqrt{\mathbb{E}[\mathbf{u}^\top X X^\top \mathbf{u}]\mathbb{E}[\mathbf{v}^\top Y Y^\top \mathbf{v}]}} \\
&= \frac{\mathbf{u}^\top \mathbb{E}[X Y^\top]\mathbf{v}}{\sqrt{\mathbf{u}^\top \mathbb{E}[X X^\top]\mathbf{u}\mathbf{v}^\top \mathbb{E}[Y Y^\top]\mathbf{v}}}.
\end{aligned}$$

Next, we outline a way to estimate the variance and covariance from data. We assume that $\mathbf{X}$ and $\mathbf{Y}$ contain $n$ i.i.d. samples of the random variables $X$ and $Y$. That is, the rows of $\mathbf{X}$ are given by datapoints $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and each $\mathbf{x}_i$ is an independent draw of random variable $X$ (and similarly for $\mathbf{Y}$). We then have the estimates:

 

$$\Sigma_{XY} = \mathbb{E}[XY^\top] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^\top = \frac{1}{n}\mathbf{X}^\top \mathbf{Y},$$

$$\Sigma_{XX} = \mathbb{E}[XX^\top] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n}\mathbf{X}^\top \mathbf{X},$$

$$\Sigma_{YY} = \mathbb{E}[YY^\top] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^\top = \frac{1}{n}\mathbf{Y}^\top \mathbf{Y}.$$

Plugging these into the definition of $\rho$ yields our estimate of the correlation

$$\widehat{\rho} = \frac{\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}}{\sqrt{\mathbf{u}^\top \mathbf{X}^\top \mathbf{X}\mathbf{u}\mathbf{v}^\top \mathbf{Y}^\top \mathbf{Y}\mathbf{v}}}.$$

(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?

**Solution:** From the expression above, it can be seen that correlation coefficient is invariant to scalings of either datasets. For example, if we scale $\mathbf{X}$ by a constant, then the denominator and numerator of the correlation coefficient will scale equally so the correlation coefficient will not change.

(d) CCA aims to find the projection vectors $\mathbf{u}$ and $\mathbf{v}$ that maximize the empirical correlation coefficient:

$$\max_{\mathbf{u},\mathbf{v}} \ \widehat{\rho}\,.$$

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 $\mathbf{M}$.

**Solution:**

Let the compact SVD be represented as $\mathbf{X} = \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{V}_x^\top$ and $\mathbf{Y} = \mathbf{U}_y \mathbf{\Sigma}_y \mathbf{V}_y^\top$.

$$\widehat{\rho} = \frac{\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}}{\sqrt{\mathbf{u}^\top \mathbf{V}_x \mathbf{\Sigma}_x^2 \mathbf{V}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{V}_y \mathbf{\Sigma}_y^2 \mathbf{V}_y^\top \mathbf{v}}} = \frac{\mathbf{a}^\top \mathbf{V}_x \mathbf{\Sigma}_x^{-1} \mathbf{V}_x^\top \mathbf{X}^\top \mathbf{Y} \mathbf{V}_y \mathbf{\Sigma}_y^{-1} \mathbf{V}_y^\top \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{a}\mathbf{b}^\top \mathbf{b}}}$$

where we let $\mathbf{a} = \mathbf{V}_x \mathbf{\Sigma}_x \mathbf{V}_x^\top \mathbf{u}$ and $\mathbf{b} = \mathbf{V}_y \mathbf{\Sigma}_y \mathbf{V}_y^\top \mathbf{v}$. This transformation can be thought of as *whitening* the data.[1] Notice that

$$\mathbf{V}_x \mathbf{\Sigma}_x^{-1} \mathbf{V}_x^\top \mathbf{X}^\top \mathbf{Y} \mathbf{V}_y \mathbf{\Sigma}_y^{-1} \mathbf{V}_y^\top = \mathbf{V}_x \mathbf{U}_x^\top \mathbf{U}_y \mathbf{V}_y^\top\,,$$

[1] This problem can equivalently be solve with the transformations $\mathbf{a} = \mathbf{\Sigma}_x \mathbf{V}_x^\top \mathbf{u}$ and $\mathbf{b} = \mathbf{\Sigma}_y \mathbf{V}_y^\top \mathbf{v}$. However, the connection to *whitening* and *decorrelating* the data are less obvious. This convention is consistent with the course notes.

where we plug in the SVD of $\mathbf{X}$ and $\mathbf{Y}$. If we assume $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^\top \mathbf{a}} = 1$ and $\|\mathbf{b}\|_2 = \sqrt{\mathbf{b}^\top \mathbf{b}} = 1$, then the denominator is 1, so the optimization problem can be written as

$$\max_{\substack{\|\mathbf{a}\|_2=1 \\ \|\mathbf{b}\|_2=1}} \mathbf{a}^\top \mathbf{V}_x \mathbf{U}_x^\top \mathbf{U}_y \mathbf{V}_y^\top \mathbf{b}.$$

Next, defining the full SVD of $\mathbf{V}_x \mathbf{U}_x^\top \mathbf{U}_y \mathbf{V}_y^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we have

$$\begin{aligned}\max_{\substack{\|\mathbf{a}\|_2=1 \\ \|\mathbf{b}\|_2=1}} \mathbf{a}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{b} &= \max_{\substack{\|\mathbf{U}\mathbf{c}\|_2=1 \\ \|\mathbf{V}\mathbf{d}\|_2=1}} \mathbf{c}^\top \mathbf{\Sigma} \mathbf{d} \\ &= \max_{\substack{\|\mathbf{c}\|_2=1 \\ \|\mathbf{d}\|_2=1}} \sum_{i=1}^{d} \sigma_i c_i d_i \ .\end{aligned}$$

where we let $\mathbf{c} = \mathbf{U}^\top \mathbf{a}$ and $\mathbf{d} = \mathbf{V}^\top \mathbf{d}$. This transformation can be thought of as *decorrelating* the data. Without loss of generality, assume that the singular values are ordered, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$. Then the optimal arguments to the weighted sum are

$$\mathbf{c}^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \mathbf{e}_1, \quad \mathbf{d}^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \mathbf{e}_1 \ .$$

Therefore, the optimal $\mathbf{u}$ and $\mathbf{v}$ are

$$\mathbf{u}^* = \mathbf{V}_x \mathbf{\Sigma}_x^{-1} \mathbf{V}_x^\top \mathbf{u}_1, \quad \mathbf{v}^* = \mathbf{V}_y \mathbf{\Sigma}_y^{-1} \mathbf{V}_y^\top \mathbf{v}_1 \ .$$

# 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?

**Solution:**

The objective function for OLS is $\arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$, and the closed form solution is $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ if $\mathbf{X}^\top \mathbf{X}$ is full rank. If $\mathbf{X}^\top \mathbf{X}$ is not full rank, then there is not a unique solution; the least-norm solution uses the pseudo-inverse. The noise model is $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$.

Ridge regression adds a regularization term: $\arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$. The closed form solution is $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$. The noise model is the same: $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$. In Homework 2, we saw that the regularization dampens the effect of the noise on the training errors.

Total least squares assumes a different noise model, in which the measurements in $\mathbf{X}$ are also affected. The noise model is therefore $\mathbf{y} + \epsilon_y = (\mathbf{X} + \epsilon_x)\mathbf{w}$. The optimization problem is $\arg\min_{\epsilon_x, \epsilon_y} \|[\epsilon_x \quad \epsilon_y]\|_F^2$, subject to $(\mathbf{X} + \epsilon_x)\mathbf{w} = \mathbf{y} + \epsilon_y$. The closed form solution is $(\mathbf{X}^\top \mathbf{X} - \sigma_{d+1}^2 \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$, where $\sigma_{d+1}$ is the least singular value of $[\mathbf{X} \quad \mathbf{y}]$.

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

**Solution:** OLS inverts the matrix $\mathbf{X}^\top \mathbf{X}$, ridge inverts the regularized matrix $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ and TLS inverts $\mathbf{X}^\top \mathbf{X} - \sigma_{d+1}^2 \mathbf{I}$ matrix.

We can write $\mathbf{X}^\top \mathbf{X} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$ with eigenvalues $\lambda_1, \ldots, \lambda_d$. Then in OLS, we invert these eigenvalues directly. In ridge regression, $\lambda$ is added to each value before inverting, which ensures numerical stability. In TLS, $\sigma_{d+1}^2$ is subtracted from each. This may decrease the numerical stability, but it can be viewed as removing the components generated by noise.

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

In the above, we assume that the noise on the measured features $\mathbf{X}$ is independent across both features and samples, and has variance $\sigma_{noise}^2$. In this case, even the smallest eigenvalue will not be not zero. TLS tries to explicitly remove the noise by subtracting $\sigma_{d+1}^2$.

(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.

**Solution:** To write down the principle components, we first write the SVD decomposition of the matrix $[\mathbf{X} \quad \mathbf{y}] = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$, where the singular values are ordered by magnitude. Then the principle components are given by $\mathbf{v}_1, \ldots, \mathbf{v}_k$, the first $k$ columns of $\mathbf{V}$. Remember that projecting the data onto these components is equivalent to finding the best rank-$k$ approximation to the original matrix $[\mathbf{X} \quad \mathbf{y}]$.

On the other hand, recall that TLS minimizes $\| \begin{bmatrix} \boldsymbol{\epsilon}_X & \boldsymbol{\epsilon}_y \end{bmatrix} \|_F^2$, subject to the constraint of:

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

Therefore, we find the minimum perturbation of $[\mathbf{X} \quad \mathbf{y}]$ that has at least one zero eigenvalue. On the homework, we see that this is equivalent to setting $\begin{bmatrix} \mathbf{X} + \boldsymbol{\epsilon}_X & \mathbf{y} + \boldsymbol{\epsilon}_y \end{bmatrix}$ to be the best rank-$d$ approximation to the original matrix $[\mathbf{X} \quad \mathbf{y}]$. Therefore, performing TLS is like using PCA and projecting onto the first $d$ principle components.

(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 multivariate 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?

**Solution:** Ridge regression assumes that each fitting target is independent. On the other hand, the CCA approach takes the potential correlation among the targets variables into account. Having taken advantage of the correlation, CCA might have higher statistical efficiency than the set of independent ridge regressions. For example, ridge regression uses $n = 100$ examples to fit each target with $p = 500$ dimensional features potentially resulting in overfitting. However, if we assume that the regression targets have large correlation with each other, conceptually we are using $nq = 1 \times 10^8$ examples to fit each target, with the same number of features.

The independent ridge regression assumes an additive noise model on each of the targets $\mathbf{Y}_i$. On the other hand, the noise model for the CCA approach depends on the data, influenced via the computed matrices $\mathbf{U}$ and $\mathbf{V}$. Therefore, it is conceptually harder to make sense of the noise.