Machine Learning, Spring 2019 Chao Lan
EN 2105 EN 4084A
MWF 9-9:50am clan@uwyo.edu

Canonical Correlation Analysis

CCA is used to reduce the feature dimensions of two correlated data sets. Let X and Z be two feature sets of a sample, e.g., $(x_i, z_i) \in (X, Z)$ is the profile of student i, where $x_i \in \mathbb{R}^{p_1}$ contains academic features (major, GPA, etc) and $z_i \in \mathbb{R}^{p_2}$ contains personal features (age, gender, etc). CCA finds a projection vector $w \in \mathbb{R}^{p_1}$ for X and a projection vector $u \in \mathbb{R}^{p_2}$ for Z so that the correlation between projected features $w^T x$ and $u^T z$ is maximized. The correlation is

$$corr(w^T x, u^T z) := \frac{cov(w^T x, u^T z)}{\sqrt{var(w^T x)}\sqrt{var(u^T z)}} = \frac{w^T cov(x, z) u}{\sqrt{w^T var(x) w}\sqrt{u^T var(z) u}}. \tag{1}$$

It can be maximized by solving an equivalent optimization problem

$$\begin{aligned} &\max_{w,u} \; w^T cov(x, z) u \\ &\text{s.t. } w^T var(x) w = u^T var(z) u = 1. \end{aligned} \tag{2}$$

Applying the Lagrange multiplier, we have the Lagrange function

$$J(w, u) = w^T cov(x, z) u - \lambda_1 (w^T var(x) w - 1) - \lambda_2 (u^T var(z) u - 1). \tag{3}$$

Since x, z are vectors, their covariance and variance are matrices. Let $\Sigma_{x,z} = cov(x, z) \in \mathbb{R}^{p_1 \times p_2}$ be their covariance matrix, $\Sigma_x = var(x) \in \mathbb{R}^{p_1 \times p_1}$ be the variance matrix of x and $\Sigma_z = var(z) \in \mathbb{R}^{p_2 \times p_2}$ be the variance matrix of z. All matrices can be estimated from sample. Then

$$J(w, u) = w^T \Sigma_{x,z} u - \lambda_1 (w^T \Sigma_x w - 1) - \lambda_2 (u^T \Sigma_z u - 1). \tag{4}$$

Applying the critical point method, we have

$$\frac{\partial J}{\partial w} = \Sigma_{x,z} u - 2\lambda_1 \Sigma_x w = 0 \quad \Longrightarrow \quad \Sigma_{x,z} u = 2\lambda_1 \Sigma_x w, \tag{5}$$

and

$$\frac{\partial J}{\partial u} = \Sigma_{x,z} w - 2\lambda_2 \Sigma_z u = 0 \quad \Longrightarrow \quad \Sigma_{x,z} w = 2\lambda_2 \Sigma_z u. \tag{6}$$

The next analysis shows $\lambda_1 = \lambda_2$. Left-multiplying w^T on both sides of (5) gives

$$w^T \Sigma_{x,z} u = 2\lambda_1 w^T \Sigma_x w = 2\lambda_1. \tag{7}$$

Similarly, left-multiplying u^T on both sides of (6) gives

$$u^T \Sigma_{x,z} w = 2\lambda_2 u^T \Sigma_z u = 2\lambda_2. \tag{8}$$

Combining (7) and (8), we have

$$2\lambda_1 = w^T \Sigma_{x,z} u = u^T \Sigma_{x,z} w = 2\lambda_2 \quad \Longrightarrow \quad \lambda_1 = \lambda_2 =: \lambda. \tag{9}$$

Now we can write (5) and (6) jointly in a matrix form

$$\begin{bmatrix} \Sigma_{x,z} & 0 \\ 0 & \Sigma_{x,z} \end{bmatrix} \cdot \begin{bmatrix} u \\ w \end{bmatrix} = 2\lambda \cdot \begin{bmatrix} 0 & \Sigma_x \\ \Sigma_z & 0 \end{bmatrix} \cdot \begin{bmatrix} u \\ w \end{bmatrix} \tag{10}$$

(10) is a generalized eigenvalue problem, where $[u; w]^T$ is an eigenvector and based on (7) λ is the largest eigenvalue. The rest analysis is similar to PCA.

Fisher Discriminant Analysis

PCA and CCA are unsupervised feature learning methods. Fisher Discriminant Analysis (FDA) is a supervised feature learning method. It learns a subspace where different classes are more separable. It does so by minimizing the variance within each class while maximizing the variance between different classes.

Let $x \in \mathcal{R}^p$ be a random instance and $w \in \mathcal{R}^p$ be a projection vector. Let there be a sample of n instances from K classes. Let μ_k be the mean of instances from class k, and μ be the mean of all instances. The <u>between-class scatter</u> is

$$S_B = \sum_{k=1}^{K} (\mu_k - \mu)(\mu_k - \mu)^T. \tag{11}$$

The <u>within-class scatter</u> is

$$S_W = \sum_{k=1}^{K} \sum_{(x,y) \in C_k} (x - \mu_k)(x - \mu_k)^T. \tag{12}$$

Let $\tilde{x} = w^T x$ be the projected instance, $\tilde{\mu}, \tilde{\mu}_k$ be the corresponding sample means, and $\tilde{S}_B, \tilde{S}_W$ be the corresponding scatters. FDA finds a w that maximizes $\tilde{S}_B$ while minimizing $\tilde{S}_W$, i.e.,

$$\hat{w}_{fda} = \arg\max_w \frac{\tilde{S}_B}{\tilde{S}_W}. \tag{13}$$

[*Discussion*] What is the geometric interpretation of FDA's objective?

It can be shown that the optimal w satisfies the following generalized eigenvalue problem

$$S_B \cdot w = \lambda \cdot S_W \cdot w, \tag{14}$$

and the solution is an eigenvector of $S_W^{-1} S_B$ associated with the largest eigenvalue λ.

[*Exercise*] Derive (14).