Machine Learning for Data Science (CS 4786)

Lecture 4: Canonical Correlation Analysis

The text in black outlines high level ideas. The text in blue provides simple mathematical details to "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested.

1 Motivation

Lets motivate Canonical Correlation Analysis through an example. Say we are interested in the task of speech recognition, that is automatically converting the waveform received by a microphone to text. What we have at our disposal is video of people speaking, that is both the visual image data and the audio data. Of course for speech recognition we would primarily use the audio data. However consider the scenario where apart from the person speaking, we also have some background noise, perhaps some music playing in the background. Now the visual data at our disposal contains information relevant to the speech such as lip movement and facial expressions of the person speaking. It also may not contain information that relates to the background noise which is only a part of the audio data.

In this scenario, given that the information from the data we are interested in is common to both the views, the hope is that one could use the visual data to filter or clean the audio data to get rid of the noise that is specific only to the audio data. This type of scenario is exactly where techniques like Canonical Correlation Analysis are useful. More generally whenever we have two views of the same data points and the information we care about is the information that is contained in both these view (ie. we have redundancy in the views), we can hope to use this redundancy to reduce noise in either one or both the views and obtain a low dimensional representation of the data consisting of mainly the redundant information.

Another example of when these techniques are of practical use is when we have multiple choices of feature extraction techniques which we believe all work well. In this scenario, we could blindly concatenate these extracted feature to form a single feature, but alternatively if we believe that all these extracted features are individually good for the task, then we can try to use techniques like Canonical Correlation Analysis to only extract the redundant information amongst these features thus reducing noise that is individual to each of the feature extraction process.

2 Two View Compression Problem

We are provided with pairs of data points $(\mathbf{x}_1, \mathbf{x}'_1), \ldots, (\mathbf{x}_n, \mathbf{x}'_n)$ where each $\mathbf{x}_t \in \mathbb{R}^d$ is a d-dimensional vector and $\mathbf{x}'_t \in \mathbb{R}^{d'}$ is a d'-dimensional vector. Think of $\mathbf{x}_t$ and $\mathbf{x}'_t$ as being two-views of the same data. For instance, if we have video data, $\mathbf{x}_t$ might be features corresponding to the image at some point in time while $\mathbf{x}'_t$ could be the associated audio data for that same point in time.

The goal is to compress point $\mathbf{x}_1, \ldots, \mathbf{x}_n$ into vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ and points $\mathbf{x}'_1, \ldots, \mathbf{x}'_n$ into vectors $\mathbf{y}'_1, \ldots, \mathbf{y}'_n \in \mathbb{R}^K$ so that we retain the information common to both the views. Again we shall use linear transformation of data.

Of course the key question at hand is how do we find appropriate linear transformations for the two views that tend to retain as much redundant information between the two views?

3 Canonical Correlation Analysis

To start with lets say we want to find a one dimensional linear projection of the points in each view. That is we want to find $\mathbf{w}_1 \in \mathbb{R}^d$ and $\mathbf{v}_1 \in \mathbb{R}^{d'}$ such that the numbers $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{y}'_1, \ldots, \mathbf{y}'_n$ retain as much of the redundant information between views 1 and 2.

A first sketch idea would be to find these directions such that the covariance between $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{y}'_1, \ldots, \mathbf{y}'_n$ is maximized. That is to maximize

$$\mathrm{Cov}(Y, Y') = \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right) \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)$$

However note that the issue with this is that the two views are two possibly completely different types of sources. One might measure in meters and other is kilometers or worse yet we might be comparing distance measurements to amplitude etc. The problem with maximizing covariance is that it is scale sensitive. So if some coordinate has high variance or scale then this direction will dominate covariance.

So the idea in PCA is to maximize not covariance but rather covariance normalized by variance. within each projection. That is maximize correlation coefficient given by

$$\mathrm{Corr}(Y, Y') = \frac{\frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right) \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right)^2} \cdot \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)^2}}$$

Now however note that $\mathbf{y}_t$'s and $\mathbf{y}'_t$'s scale linearly with $\mathbf{w}_1$ and $\mathbf{v}_1$ respectively. That is, if we scale $\mathbf{w}_1$ to $\alpha \mathbf{w}_1$ then $\mathbf{y}_t$ scales by α too. Further, note that the term $\mathrm{Corr}(Y, Y')$ does not change if we scale all $\mathbf{y}_t$'s or $\mathbf{y}'_t$'s as the α-scaling in numerator is canceled by the scaling in denominator. Hence we can always scale $\mathbf{w}_1$ and $\mathbf{v}_1$ appropriately so that $\frac{1}{n} \sum_{t=1}^{n} \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right)^2 = 1$. Hence the first directions of CCA $\mathbf{w}_1$ and $\mathbf{v}_1$ can be written as the ones that maximize

$$\frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right) \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)$$

$$\text{such that } \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{1}{n} \sum_{s=1}^{n} y_s \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left(y'_t - \frac{1}{n} \sum_{s=1}^{n} y'_s \right)^2 = 1$$

Writing this explicitly,

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top \mathbf{x}_t - \frac{1}{n} \sum_{s=1}^{n} \mathbf{w}_1^\top \mathbf{x}_s \right) \left(\mathbf{v}_1^\top \mathbf{x}'_t - \frac{1}{n} \sum_{s=1}^{n} \mathbf{v}_1^\top \mathbf{x}'_s \right)$$
$$\text{s.t. } \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top \mathbf{x}_t - \frac{1}{n} \sum_{s=1}^{n} \mathbf{w}_1^\top \mathbf{x}_s \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{v}_1^\top \mathbf{x}'_t - \frac{1}{n} \sum_{s=1}^{n} \mathbf{v}_1^\top \mathbf{x}'_s \right)^2 = 1$$

which can be rewritten as:

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top \mathbf{x}_t - \mathbf{w}_1^\top \left(\frac{1}{n} \sum_{s=1}^{n} \mathbf{x}_s \right) \right) \left(\mathbf{v}_1^\top \mathbf{x}'_t - \mathbf{v}_1^\top \left(\frac{1}{n} \sum_{s=1}^{n} \mathbf{x}'_s \right) \right)$$
$$\text{s.t. } \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top \mathbf{x}_t - \mathbf{w}_1^\top \left(\frac{1}{n} \sum_{s=1}^{n} \mathbf{x}_s \right) \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{v}_1^\top \mathbf{x}'_t - \mathbf{v}_1^\top \left(\frac{1}{n} \sum_{s=1}^{n} \mathbf{x}'_s \right) \right)^2 = 1$$

Let $\mu = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t$ and $\mu' = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}'_t$. We can rewrite the above as

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top (\mathbf{x}_t - \mu) \right) \left(\mathbf{v}_1^\top (\mathbf{x}'_t - \mu') \right)$$
$$\text{s.t. } \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_1^\top (\mathbf{x}_t - \mu) \right)^2 = \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{v}_1^\top (\mathbf{x}'_t - \mu') \right)^2 = 1$$

Using the fact that $(a^\top b)(c^\top d) = a^\top (bc^\top) d$ and also noting that since $c^\top d = d^\top c$, realizing that $(a^\top b)(c^\top d) = a^\top (bd^\top) c$ we can rewrite the above as:

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}_1^\top \left((\mathbf{x}_t - \mu)(\mathbf{x}'_t - \mu')^\top \right) \mathbf{v}_1$$
$$\text{s.t. } \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}_1^\top \left((\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \right) \mathbf{w}_1 = \frac{1}{n} \sum_{t=1}^{n} \mathbf{v}_1^\top \left((\mathbf{x}'_t - \mu')(\mathbf{x}'_t - \mu')^\top \right) \mathbf{v}_1 = 1$$

hence we have

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \mathbf{w}_1^\top \left(\frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}'_t - \mu')^\top \right) \mathbf{v}_1$$
$$\text{s.t. } \mathbf{w}_1^\top \left(\frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \right) \mathbf{w}_1 = \mathbf{v}_1^\top \left(\frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}'_t - \mu')(\mathbf{x}'_t - \mu')^\top \right) \mathbf{v}_1 = 1$$

Now note that $\Sigma_{2,2} = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}'_t - \mu')(\mathbf{x}'_t - \mu')^\top$ is covariance matrix in view 2,
$\Sigma_{1,1} = \left(\frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \right)$ is covariance matrix for view 1
and $\Sigma_{1,2} = \left(\frac{1}{n} \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}'_t - \mu')^\top \right)$ is the covariance matrix between views 1 and 2.

Hence we have

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \mathbf{w}_1^\top \Sigma_{1,2} \mathbf{v}_1$$
$$\text{s.t. } \mathbf{w}_1^\top \Sigma_{1,1} \mathbf{w}_1 = \mathbf{v}_1^\top \Sigma_{2,2} \mathbf{v}_1 = 1$$

Now to solve the optimization problem we use Lagrange multipliers. We want to optimize

$$\mathbf{w}_1, \mathbf{v}_1 = \underset{\mathbf{w}_1, \mathbf{v}_1}{\operatorname{argmax}} \mathbf{w}_1^\top \Sigma_{1,2} \mathbf{v}_1 + \lambda_1^*(1 - \mathbf{w}_1^\top \Sigma_{1,1} \mathbf{w}_1) + \lambda_2^*(1 - \mathbf{v}_1^\top \Sigma_{2,2} \mathbf{v}_1)$$

Taking derivative and equating to 0 for each of $\mathbf{w}_1$ and $\mathbf{v}_1$ we find,

$$\Sigma_{1,2}\mathbf{v}_1 = \lambda_1^* \Sigma_{1,1}\mathbf{w}_1 \quad \& \quad \Sigma_{2,1}\mathbf{w}_1 = \lambda_1^* \Sigma_{2,2}\mathbf{v}_1 \tag{1}$$

However we know that $\mathbf{w}_1^\top \Sigma_{1,1}\mathbf{w}_1 = 1$ and so multiplying the first equation above by $\mathbf{w}_1^\top$ from the right we get,

$$\mathbf{w}_1^\top \Sigma_{1,2}\mathbf{v}_1 = \lambda_1^*$$

Similarly, since $\mathbf{v}_1^\top \Sigma_{2,2}\mathbf{v}_1 = 1$ and so multiplying the second equation above by $\mathbf{v}_1^\top$ from the right we get,

$$\mathbf{v}_1^\top \Sigma_{2,1}\mathbf{w}_1 = \lambda_2^*$$

Hence we conclude that $\lambda_2^* = \mathbf{v}_1^\top \Sigma_{2,1}\mathbf{w}_1 = \mathbf{w}_1^\top \Sigma_{1,2}\mathbf{v}_1 = \lambda_1^* = \lambda^*$ Finally, note again that from Eq. 1 (the second one), multiplying both size by $\Sigma_{2,2}^{-1}$, we have $\Sigma_{2,2}^{-1}\Sigma_{2,1}\mathbf{w}_1 = \lambda^*\mathbf{v}_1$. Using this in the first equation in (1) we find that

$$\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\mathbf{w}_1 = (\lambda^*)^2 \Sigma_{1,1}\mathbf{w}_1$$

Multiplying both side above by $\Sigma_{1,1,}^{-1}$ we finally conclude that for the solution $\mathbf{w}_1$,

$$\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\mathbf{w}_1 = (\lambda^*)^2 \mathbf{w}_1$$

That is $\mathbf{w}_1$ is an eigenvector of $\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$. Further, note that the objective

$$\mathbf{w}_1^\top \Sigma_{1,2}\mathbf{v}_1 = \lambda^*$$

and so maximizing the objective corresponds to maximizing the eigenvalue of $\mathbf{w}_1$. Thus $\mathbf{w}_1$ is the Top eigenvector of matrix $\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$. Similarly we find that $\mathbf{v}_1$ is the top eigenvector of $\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$.

To find the remaining $K-1$ directions we simply look for subsequent $\mathbf{w}_i, \mathbf{v}_i$ that maximize the same objective above but are such that $\mathbf{v}_i$ is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$ and $\mathbf{w}_i$ is orthogonal to $\mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$. This solution turns out to be the top K eigen vectors of matrices $\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$ and $\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$ to find W and V respectively.

Thus the solution to CCA is

$$W = \text{eigs}\left(\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}, K\right)$$

and

$$V = \text{eigs}\left(\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}, K\right)$$