Lecture 4: Canonical Correlation Analysis (CCA)

Karl Stratos

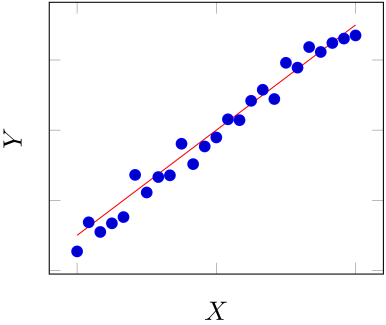
October 15, 2018

Correlation Coefficient

- **Correlation coefficient** between random variables $X, Y \in \mathbb{R}$:

$$\mathrm{cor}\left(X, Y\right) := \frac{\mathbf{E}\left[\left(X - \mathbf{E}\left[X\right]\right)\left(Y - \mathbf{E}\left[Y\right]\right)\right]}{\sqrt{\mathbf{E}\left[\left(X - \mathbf{E}\left[X\right]\right)^2\right]}\sqrt{\mathbf{E}\left[\left(Y - \mathbf{E}\left[Y\right]\right)^2\right]}}$$

Degree of linear relationship $[-1, 1]$



$$\mathrm{cor}\left(X, Y\right) \approx 1$$

Facts About Correlation Coefficient

- Cosine of the angle b/t *centered* X, Y (under $\langle X, Y \rangle := \mathbf{E}[XY]$)

$$\mathrm{cor}(X, Y) = \frac{\langle X - \mathbf{E}[X], Y - \mathbf{E}[Y] \rangle}{||X - \mathbf{E}[X]|| \, ||Y - \mathbf{E}[Y]||} = \cos\theta$$

- Invariant to scale/location: $X - \mathbf{E}[X] = (X + c) - \mathbf{E}[X + c]$

$$\mathrm{cor}(X, Y) = \mathrm{cor}(\alpha X + c, \beta Y + c') \qquad \forall \alpha, \beta, c, c' \in \mathbb{R}$$

- 0 when independent, ± 1 when parellel

$$\begin{aligned} \mathrm{cor}(X, Y) = 0 \quad &\iff \quad \mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y] \\ \mathrm{cor}(X, Y) = 1 \quad &\iff \quad X = \alpha Y \;\; \alpha > 0 \\ \mathrm{cor}(X, Y) = -1 \quad &\iff \quad X = \alpha Y \;\; \alpha < 0 \end{aligned}$$

Overview

- Views of CCA
  - Correlation Maximization
  - Subspace Optimization
- Deep CCA

Optimization Problem Underlying CCA

Input:

1. $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ // two "views" of an object
2. $m \leq \min(d, d')$ // number of projection vectors

Output: $(a_1, b_1) \ldots (a_m, b_m) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ such that

- (a_1, b_1) is a solution of

$$\arg\max_{a,b} \quad \mathrm{cor}\left(a^\top X, b^\top Y\right)$$

- For $i = 2 \ldots m$: (a_i, b_i) is a solution of the above subject to:

$$\mathrm{cor}\left(a^\top X, a_j^\top X\right) = 0 \qquad \forall j < i$$
$$\mathrm{cor}\left(b^\top Y, b_j^\top Y\right) = 0 \qquad \forall j < i$$

Definitions

Cross-covariance matrix given by

$$C_{XY} := \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])^\top\right] \in \mathbb{R}^{d \times d'}$$

Covariance matrices are assumed to be invertible

$$C_{XX} := \mathbf{E}\left[(X - \mathbf{E}[X])(X - \mathbf{E}[X])^\top\right] \in \mathbb{R}^{d \times d}$$
$$C_{YY} := \mathbf{E}\left[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^\top\right] \in \mathbb{R}^{d' \times d'}$$

Define **correlation matrix**

$$\Omega := C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2} \in \mathbb{R}^{d \times d'}$$

Exact Solution via SVD (Hotelling, 1936)

$$(a_i, b_i) \in \mathop{\arg\max}_{\substack{a \in \mathbb{R}^d,\ b \in \mathbb{R}^{d'}: \\ \operatorname{cor}\left(a^\top X, a_j^\top X\right) = 0\ \forall j < i \\ \operatorname{cor}\left(b^\top Y, b_j^\top Y\right) = 0\ \forall j < i}} \operatorname{cor}\left(a^\top X, b^\top Y\right)$$

Claim. If $U \Sigma V^\top$ is an SVD of Ω, then

$$\sigma_i = \max_{\substack{a \in \mathbb{R}^d,\ b \in \mathbb{R}^{d'}: \\ \operatorname{cor}\left(a^\top X, a_j^\top X\right) = 0\ \forall j < i \\ \operatorname{cor}\left(b^\top Y, b_j^\top Y\right) = 0\ \forall j < i}} \operatorname{cor}\left(a^\top X, b^\top Y\right)$$

with a solution

$$a_i = C_{XX}^{-1/2} u_i \qquad\qquad b_i = C_{YY}^{-1/2} v_i$$

Matrix Form

- Organize $A = [a_1 \ldots a_m] \in \mathbb{R}^{d \times m}$ and $B = [a_1 \ldots a_m] \in \mathbb{R}^{d \times m}$
- Solution given by $A = C_{XX}^{-1/2} U^*$ and $B = C_{YY}^{-1/2} V^*$

$$(U^*, V^*) \in \operatorname*{arg\,max}_{\substack{U \in \mathbb{R}^{d \times m},\, V \in \mathbb{R}^{d' \times m}: \\ U^\top U = V^\top V = I_m}} \left|\left| U^\top \Omega V \right|\right|_1$$

where $||M||_1 := \mathrm{tr}\left(\left(M^\top M\right)^{1/2}\right) = \sum_i \sigma_i(M)$ is the **nuclear norm**

- Optimal value $\sum_{i=1}^m \sigma_i(\Omega)$ at top m left/right singular vectors of Ω

Empirical Version

Input: N samples of (X, Y) organized as $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{d' \times N}$

1. Center the data (okay to skip if sparse and binary)

$$\overline{\boldsymbol{X}} = \boldsymbol{X} - \hat{\mu}_{\boldsymbol{X}} \qquad \overline{\boldsymbol{Y}} = \boldsymbol{Y} - \hat{\mu}_{\boldsymbol{Y}}$$

2. Calculate $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$, an SVD of

$$\widehat{\Omega} = \left(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top + \frac{\kappa}{N} I_d\right)^{-1/2} \overline{\boldsymbol{X}}\,\overline{\boldsymbol{Y}}^\top \left(\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^\top + \frac{\kappa}{N} I_{d'}\right)^{-1/2}$$

3. Given sample $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$, calculate their new m-dimensional representations $(\underline{x}, \underline{y}) \in \mathbb{R}^m \times \mathbb{R}^m$ by

$$\underline{x} = U_m^\top \left(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top + \frac{\kappa}{N} I_d\right)^{-1/2} (x - \hat{\mu}_{\boldsymbol{X}})$$

$$\underline{y} = V_m^\top \left(\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^\top + \frac{\kappa}{N} I_{d'}\right)^{-1/2} (y - \hat{\mu}_{\boldsymbol{Y}})$$

Overview

- Views of CCA
  - Correlation Maximization
  - Best-Match Subspaces
- Deep CCA

Best-Match Subspaces

Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N$ be subspaces with dimensions $d \leq d' \leq N$.

For $i = 1 \ldots d$, cosine of the **canonical angle** between $\mathcal{X}$ and $\mathcal{Y}$ is

$$\cos \angle_i(\mathcal{X}, \mathcal{Y}) := x_i^* \cdot y_i^* \qquad (x_i^*, y_i^*) = \operatorname*{arg\,max}_{\substack{x \in \mathcal{X}:\ ||x||=1 \\ y \in \mathcal{Y}:\ ||y||=1 \\ x \cdot x_j^* = y \cdot y_j^* = 0\ \forall j < i}} x \cdot y$$

Define "best-match" subspaces with dimension $m \leq d$ by

$$(\mathcal{S}^*, \mathcal{T}^*) = \operatorname*{arg\,max}_{\substack{\mathcal{S} \subseteq \mathcal{X}:\ \dim(\mathcal{S})=m \\ \mathcal{T} \subseteq \mathcal{Y}:\ \dim(\mathcal{T})=m}} \sum_{i=1}^{m} \cos_i \angle_i(\mathcal{S}, \mathcal{T})$$

Claim. $\{x_i^*\}_{i=1}^m$ is an orthonormal basis of $\mathcal{S}^*$. $\{y_i^*\}_{i=1}^m$ is an orthonormal basis of $\mathcal{T}^*$.

Best-Match Subspaces (Cont.)

Claim. Let $X \in \mathbb{R}^{N \times d}$ and $Y \in \mathbb{R}^{N \times d'}$ be orthonormal bases of $\mathcal{X}, \mathcal{Y}$. Consider an SVD of $X^\top Y \in \mathbb{R}^{d \times d'}$

$$X^\top Y = U \Sigma V^\top$$

Then $XU_m,\ YV_m \in \mathbb{R}^{N \times m}$ are orthonormal bases of $\mathcal{S}^*, \mathcal{T}^*$.

Back to CCA

- View (centered) data matrices $\overline{\boldsymbol{X}} \in \mathbb{R}^{d \times N}$ and $\overline{\boldsymbol{Y}} \in \mathbb{R}^{d' \times N}$ as subspaces of $\mathbb{R}^N$: namely $\mathrm{row}\left(\overline{\boldsymbol{X}}\right)$ and $\mathrm{row}\left(\overline{\boldsymbol{Y}}\right)$.

- Orthonormal bases given by $(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top)^{-1/2}\overline{\boldsymbol{X}}$ and $(\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^\top)^{-1/2}\overline{\boldsymbol{Y}}$.

- Hence considering an SVD of

$$(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top)^{-1/2}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Y}}^\top(\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^\top)^{-1/2} = U\Sigma V^\top$$

  orthonormal bases of the best-match subspaces of dimension m between $\mathrm{row}\left(\overline{\boldsymbol{X}}\right)$ and $\mathrm{row}\left(\overline{\boldsymbol{Y}}\right)$ given by

$$U_m^\top(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top)^{-1/2}\overline{\boldsymbol{X}} \qquad V_m^\top(\overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}^\top)^{-1/2}\overline{\boldsymbol{Y}}$$

A Bunch of Other Views

- See Golub and Zha (1992) for a compilation of different formulations.
- See Bach and Jordan (2006) for a latent-variable formulation.

Overview

- Views of CCA
  - Correlation Maximization
  - Subspace Optimization
- Deep CCA

Deep CCA

- Let $f_\phi : \mathbb{R}^{d \times N} \to \mathbb{R}^{m \times N}$ be some neural net parameterized by ϕ.
- Let $g_\psi : \mathbb{R}^{d' \times N} \to \mathbb{R}^{m \times N}$ be some neural net parameterized by ψ.
- Example: $\phi = \left\{W^1, W^2, b^1, b^2\right\}$ with

$$f_\phi(\boldsymbol{X}) = W^2 \tanh\left(W^1 X + b^1\right) + b^2$$

- Let $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}$ denote $f_\phi(\boldsymbol{X}), g_\psi(\boldsymbol{Y})$ after centering and division by N.
- Sum of the m canonical correlations between datasets under this transformation is

$$\left\|\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top\right)^{-1/2} \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{Y}}^\top \left(\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^\top\right)^{-1/2}\right\|_1 \in [0, m]$$

This is differentiable wrt. $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}$ and hence ϕ, ψ.

Questions

- When does dimensionality reduction happen?
- What if $\boldsymbol{Z} = f_\phi(\boldsymbol{X}) = g_\psi(\boldsymbol{Y})$ for some full-rank $\boldsymbol{Z} \in \mathbb{R}^{m \times N}$?
- What if $0 = f_\phi(\boldsymbol{X}) = g_\psi(\boldsymbol{Y})$?