Low Rank Approximations, PCA, and CCA

Abhinav Gopal, Shrey Vasavada, William McEachen, Grace Kull, Jai Bansal

Berkeley
UNIVERSITY OF CALIFORNIA

- What is a low rank approximation?
  - Large amounts of data is expensive
  - Furthermore, some *unimportant* aspects of data reflect noise
  - How do we create an approximation of the data with just the important features?
- Option: Reduce features
  - How? Choose the most “important” features of the matrix
  - Decompose to svd and only use the top “k” values
  - $\sum_{i \in (1 \ldots k)} u_i \sigma_i v^T_i$

Berkeley
UNIVERSITY OF CALIFORNIA

- Notice from the previous slide that our low rank approximation's column space was the k vectors v_i
- These vectors represent the directions of most variance in our data
- Idea: Use these vectors in Xw = y prediction tasks
  - Project X onto these important directions, and use this projection to predict y.

Berkeley
UNIVERSITY OF CALIFORNIA

- This brings us to PCA!
- Procedure
  - Compute SVD for matrix X
  - Get the top k eigenvectors V^T_i and choose these eigenvectors
  - Get the principal components by finding $Z_k = XV_k$
  - Get the approximation for X by doing ZV^T_k
  - Use this approximation and do least squares

- Great, but flawed, way to approximate a matrix and eliminate noise.
  - Why? Relies on the assumption that the largest variance directions are the most important ones.
- Very easy to modify data in order to render PCA useless
- What is the workaround?
  - We try to find the actual correlation between X and y.

- This brings us to CCA
  - Canonical Correlation Analysis
- CCA uses the pearson correlation coefficient in order to determine the relationship between X and y.

$$\hat{\rho}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{\tilde{x}^\top \tilde{y}}{\sqrt{\tilde{x}^\top \tilde{x} \cdot \tilde{y}^\top \tilde{y}}} \quad \text{where } \tilde{x} = x - \bar{x}, \tilde{y} = y - \bar{y}$$

Berkeley
UNIVERSITY OF CALIFORNIA

- Goal: To Solve

$$\max_{\mathbf{u},\mathbf{v}} \rho(\mathbf{X}_{\mathrm{rv}}{}^{\top}\mathbf{u}, \mathbf{Y}_{\mathrm{rv}}{}^{\top}\mathbf{v}) = \max_{\mathbf{u},\mathbf{v}} \frac{\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}{}^{\top}\mathbf{u}, \mathbf{Y}_{\mathrm{rv}}{}^{\top}\mathbf{v})}{\sqrt{\mathrm{Var}(\mathbf{X}_{\mathrm{rv}}{}^{\top}\mathbf{u})\,\mathrm{Var}(\mathbf{Y}_{\mathrm{rv}}{}^{\top}\mathbf{v})}}$$

- X_{rv} and Y_{rv} are both real vectors that correspond to the size of X and Y.
- With some algebra, we find that the problem becomes

$$= \max_{\mathbf{u},\mathbf{v}} \underbrace{\frac{\mathbf{u}^{\top}\mathbf{X}^{\top}\mathbf{Y}\mathbf{v}}{\sqrt{\mathbf{u}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{u} \cdot \mathbf{v}^{\top}\mathbf{Y}^{\top}\mathbf{Y}\mathbf{v}}}}_{\hat{\rho}(\mathbf{X}\mathbf{u},\mathbf{Y}\mathbf{v})}$$

Berkeley
UNIVERSITY OF CALIFORNIA

- Whiten the X and Y matrices (make their covariance 0) by
  - Use $W_x = U_x S^{1/2}{}_x U_x{}^T$
  - Use $W_y = U_y S^{1/2}{}_y U_y$
  - SVD of $X^T X = U_x S_x V_x{}^T$
  - SVD of $Y^T Y = U_y S_y V_y{}^T$
  - Use $X_w = XW_x$, $Y_w = YW_Y$

Berkeley
UNIVERSITY OF CALIFORNIA

- With the change of variables $u_w = W_x^{-1}u$ and $v_w = W_y^{-1}v$, the max expression becomes:
  - $$\max_{\mathbf{u}_w, \mathbf{v}_w} \underbrace{\frac{\mathbf{u}_w^\top \mathbf{X}_w^\top \mathbf{Y}_w \mathbf{v}_w}{\sqrt{\mathbf{u}_w^\top \mathbf{u}_w \cdot \mathbf{v}_w^\top \mathbf{v}_w}}}_{\hat{\rho}(\mathbf{X}_w \mathbf{u}_w, \mathbf{Y}_w \mathbf{v}_w)}$$
- Now, let's decorrelate X_w and Y_w, making $(X_w D_x)^T (Y_w D_y)$
- We decorrelate them with the choice of
  - U for D_x and V for D_y, where $USV^T = X_w^T Y_w$

Berkeley UNIVERSITY OF CALIFORNIA

- Changing variables with
    - $u_d = D_x^T u_w$ and $v_d = D_y^T v_w$, the maximization becomes the following:

$$\max_{\substack{||\mathbf{u}_d||=1\\ ||\mathbf{v}_d||=1}} \mathbf{u}_d^\top \mathbf{S} \mathbf{v}_d$$

- Finally, we get the associated eigenvectors of interest!
  - $U = W_x D_x U_d$, $V = W_y D_y V_d$, with

$$\mathbf{U}_d = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{p-k,k} \end{bmatrix} \in \mathbb{R}^{p \times k}$$

$$\mathbf{V}_d = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{q-k,k} \end{bmatrix} \in \mathbb{R}^{q \times k}$$