CS 189 Introduction to Machine Learning
Fall 2019 Jennifer Listgarten & Stella Yu

HW 05

Due: October 29th, 2019

1 Getting Started

Read through this page carefully. You may typeset your homework in latex or submit neatly handwritten/scanned solutions. Please start each question on a new page. Deliverables:

1. Submit a PDF of your writeup, **with an appendix for your code**, to assignment on Gradescope, "HW5 Write-Up". If there are graphs, include those graphs in the correct sections. Do not simply reference your appendix.
2. If there is code, submit all code needed to reproduce your results, "HW5 Code".
3. If there is a test set, submit your test set evaluation results, "HW5 Test Set".

2 Rayleigh Quotients

(a) Given an $n \times n$ symmetric matrix M, its Rayleigh quotient is defined as

$$R(M, x) = \frac{x'Mx}{x'x}.$$

What are the minimum and maximum values of $R(M, x)$? What values of x achieve these lower and upper bounds?

(b) How does the Rayleigh quotient relate to the following optimization problems?

$$\max_{w:\|w\|_2=1} \|Xw\|_2^2.$$

Note: You may recognize this optimization problem as exactly PCA. The largest eigenvalue represents the largest amount of variance in one direction. Using the largest eigenvalue direction captures $\frac{\lambda_1}{\sum_{i=1}^n \lambda_i}$ proportion of the total variance. Using the k largest eigenvalue directions captures $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}$ proportion of the total variance. Figure 1 shows an example where 11 PCA components already capture 76% of the total variance of the data. Such an approach, measuring the percentage of achieved variation in a reduced data representation, is sometimes used in practice to pick the number of PCAs.

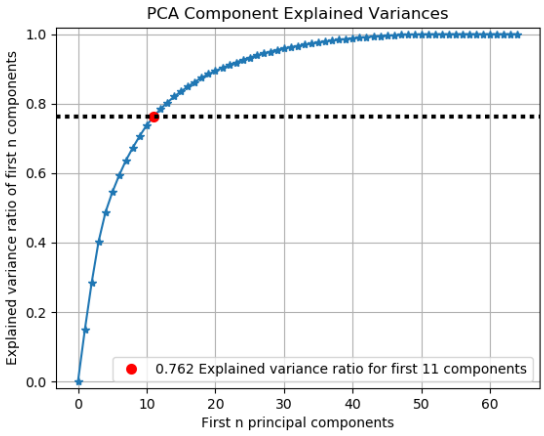


Figure 1: Image from https://scikit-plot.readthedocs.io/en/stable/decomposition.html

(c) We may consider Rayleigh quotients from an alternate perspective. Consider $R(A, x)$ for an arbitrary x, not necessarily an eigenvector. Show that

$$\arg\min_{\lambda} \|Ax - \lambda x\|_2^2 = R(A, x).$$

What happens when x is an eigenvector?

3 Correlation Coefficient

The correlation between random variables X and Y is captured by the so-called Pearson correlation coefficient:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V[X] \cdot V[Y]}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X) \cdot \text{cov}(Y, Y)}} \tag{1}$$

$$\text{cov}(X, Y) = E\left[(X - E[X]) \cdot (Y - E[Y]) \right] \tag{2}$$

Note that $\rho(X, Y)$ is *not* defined when the denominator is 0, i.e. either X or Y has zero variation.

(a) Show that $\rho(X, Y)$ is affine invariant in the following sense.

$$|\rho(aX + c, bY + d)| = |\rho(X, Y)|, \qquad a \cdot b \neq 0 \tag{3}$$

That is, no matter how you scale and shift X and Y, the size of correlation remains the same.

(b) Write your code to estimate ρ from a set of 2D points $(x_1, y_1), \ldots (x_n, y_n)$. The so-called sample correlation coefficient $r(x, y)$ is defined as:

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \ \sum_{i=1}^n (y_i - \bar{y})^2}}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i. \tag{4}$$

Please make your code general enough to handle degenerate cases: When there is no variation in either x or y, it should return "N/A" – i.e., r is not defined.

The code to generate the datasets shown below is provided to you (*pointsGenerator.py*). Report your results on these datasets (including the plots). Play with the data generation parameters to get a sense on how the correlation coefficient would change.

4 Canonical Correlation Analysis

The goal of canonical correlation analysis (CCA) is to find corresponding projected spaces where the correlation of two random vectors is maximized.

Given paired data (X, Y), each consisting of points in 2D, CCA seeks projection axis u for X and projection axis v for Y such that their projections achieve the maximum cross correlation:

$$\max_{\|u\|=1,\|v\|=1} \rho(Xu, Yv). \tag{5}$$

(a) Show that the optimal (u, v) must satisfy:

$$\begin{bmatrix} & C_{xy} \\ C_{yx} & \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}. \tag{6}$$

The optimum correlation value is the largest eigenvalue achieved when (u, v) is the corresponding eigenvector:

$$\lambda_1 = \max_{u,v} \rho(Xu, Yv) \tag{7}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \text{eig}_1 \left(\begin{bmatrix} & C_{xy} \\ C_{yx} & \end{bmatrix}, \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix} \right) \tag{8}$$

(b) Show that the generalized eigenvector can be solved in three steps.

1) Step 1: Transform the projection axis representation.

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} u \\ v \end{bmatrix} \tag{9}$$

2) Step 2: Solve the SVD of a normalized covariance matrix.

$$(\tilde{u}, \tilde{v}, \lambda) = \text{svd}(C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}}) \tag{10}$$

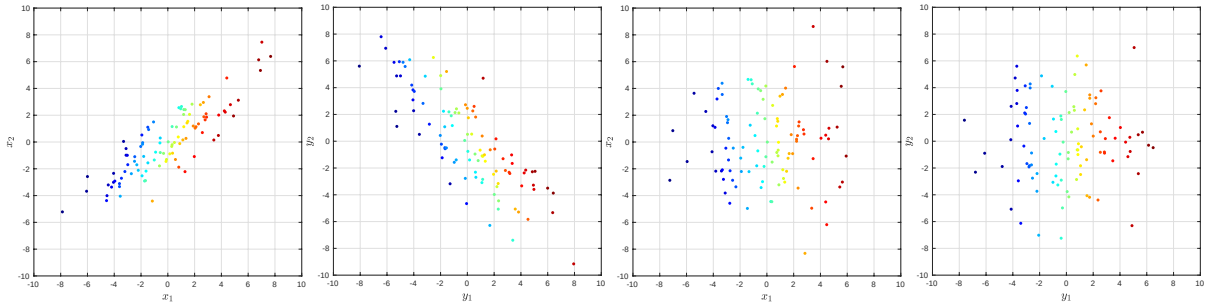
3) Step 3: Transform the axes back into their original coordinate representations.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} \tag{11}$$

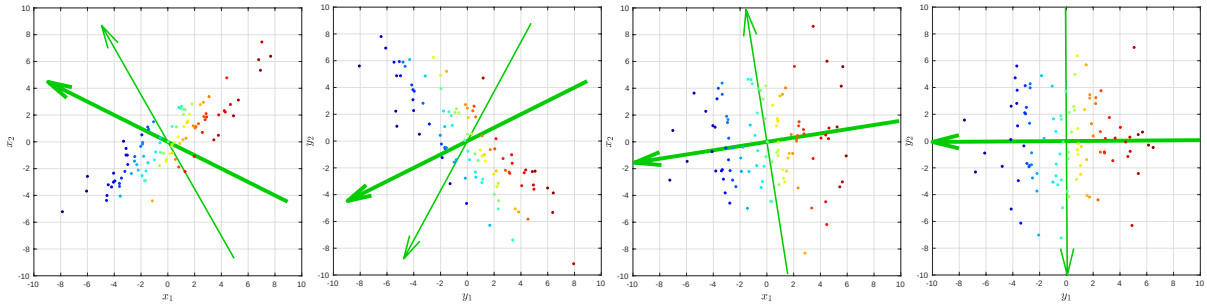
(c) Show that the first step of changing the basis for the projection axes is equivalent to whiten the data X and Y in their individual spaces.

(d) Show that the second step of solving the SVD is equivalent to maximally align the two data sets, i.e. to maximize the correlation coefficient between whitened data points.

(e) Code up these three steps, and use your code to compute the CCA of 2D data points. Plot the original data points, their whitened versions, the CCA in the whitened spaces and in the original spaces, in the same fashion shown below.



X Y Step 1. whitened X whitened Y



Step 3. u v Step 2. $\tilde{u}$ $\tilde{v}$

The code to generate the dataset is provided to you (*CCA.py*). Play with the data generation parameters to get a sense on how the CCA would change.