Syllabus for the Datamining Class

Gener Avilés R

2017-04-12

Contents

Chapter 1

Introduction

This course is taught in the ***Maestría y Doctorado en Ciencias e Ingeniería*** (MyDCI) programm of Facultad de Ingeniería, Arquitectura y Diseño of Universidad Autónoma de Baja California in the Ensenada Campus.

The course is taught by Dr. María de los Ángeles Cosío León.

Chapter 2

Principal Components Analysis

2.1 What does PCA do?

This methods tries to explain the correlation structure of a set of predictor variables using a smaller set o linear combinations of these variables called ***components***, note that components are not variables, rather indicators of linear combinations between variables. Given a dataset with m variables a set of k linear combinations can be used to represent it (meaning that k contains almost as much information as the m variables), also $k << m$.

2.2 PCA Step by Step

2.2.1 1. Getting the dataset and things ready.

Before starting the process of dimensionality reduction one should make sure the data is standardized, this is done to avoid bised in the results by values to large or to small when compared to each other.

2.2.2 2. Centering the points

- The **standardization process** is acomplished when the mean for each variable = 0 and the standard deviation = 1. The following formula can be followed to acomplish this process:

$$Z_i = \frac{(X_i - \mu_i)}{\sigma_{ii}}$$

Where: μ_i equals the mean of X_i and σ_{ii} equals the standard deviation of X_i.

- If the values are given as a set of points the process can be acomplished with the following formula:

$$x_{i,a} = x_{i,a} - \mu_a$$

This move will facilitate the calculations down the road.

2.2.3 3. Compute covariance ($\sigma_{X,Y}$) matrix

The **covariance** is a measure of the degree to which two variables vary together. Positive covariance indicates that when one variable increases, the other tends to increase. Negative covariance indicates that when one variable increases, the other tends to decrease. The covariance measure **is not scaled**.

In a $2x2$ matrix:

$$\begin{vmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{vmatrix}$$

Since the mean (μ) is equal to $\emptyset$ thanks to centering the values in the previous step, the formula to calculate the covariance of the values in the matrix is:

$$cov(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}x_{i,2}$$

The way to interpret *covariance* is to understand it's results as information about how one attribute changes as the other one changes.

It is important to remember that, if we multiply a vector by the covariance matrix or $\sum$ the resulting vector will turn towards the direction of the variance.

Changing the units of measure would change the results, this is an inconvenience and is addressed by calculating the ***correlation coefficient*** r_{ij}:

r_{ij} scales the covariance by each of the standard deviations:

$$r_{ij} = \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}}$$

The r_{ij} gives us a value with reference to know how much of a correlation exists between two variables.

2.2.4 4. Eigenvectors + Eigenvalues

Define a **new set of dimentions** by:

1. Taking the dataset and looking for the direction of the data, looking to draw a line in which, along it, there is the **greatest amount of variance** σ^2 in the data, this line will be called the **principal component 1 (PC1)**.

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \text{or} \sigma^2 = \frac{\sum X^2}{N} - \mu^2$$

In the previous formula σ^2 is defined as the sum of the squared distances of each term in the distribution from the mean (μ^2) divided by the number of terms in the distribution (N). In simple words: σ^2 measures **how far a set of random numbers are spread out from their mean**.

2. Once PC1 is determined, it will established the next dimension by drawing an ***orthogonal*** (perpendicular) line in relation to PC1, the exact area where the line will be drawn is determined by the same process of finding the gratest σ^2 of the remaining data, once this is done PC2 is ready.

3. This will be done iteratively until all the dimensions (d) of the dataset are covered and components (m) are generated for every single d.

- The first $m << d$ components become m new dimensions.
  - Coordinates from every datapoint will be changed to these "new" dimensions.
- **Greatest variability** is pursued to maintain the smoothness assumption of dimensions.

Eigenvectors and eigenvalues are mathematically expressed as:

$$A\overrightarrow{v} = \lambda\overrightarrow{v}$$

Where A represents transformation, $\overrightarrow{v}$, a vector, also known as **eigenvector**, that comes out of the matrix being analysied and λ, a scalar value also known as **eigenvalue**.

Principal components = eigenvectors with largest eigenvalues.

2.2.4.1 Finding Eigenvalues and Eigenvectors

In order to exemplify the process of finding these values and vector steps are presented for a $2x2$ matrix, but this can be done with any matrix of nxn dimensions following the rules of matrix algebra.

To begin with the example we will declare a matrix:

$$A = \begin{bmatrix} 7 & 3 \\ 3 & -1 \end{bmatrix}$$

Now the steps:

1. **Multiply an nxn identity matrix by the scalar λ:** $IA\lambda$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \lambda = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

2. **Substract the identity matrix multiple from matrix A:** $A - \lambda I$

$$\begin{bmatrix} 7 & 3 \\ 3 & -1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7-\lambda & 3 \\ 3 & -1-\lambda \end{bmatrix}$$

3. **Find the determinant of the matrix obtained in previous step:** $det(A - \lambda I)$

$$det \begin{bmatrix} 7-\lambda & 3 \\ 3 & -1-\lambda \end{bmatrix} = (7-\lambda)(-1-\lambda) - (3*3)$$

$$= -7 - 7\lambda + \lambda + \lambda^2 = -16 - 6\lambda + \lambda^2$$
$$= \lambda^2 - 6\lambda - 16$$

4. **Solve for the values of λ that satisfy the equation** $det(A - \lambda I) = 0$ Solving for $\lambda^2 - 6\lambda - 16 = 0$ will result in:

$$(\lambda - 8)(\lambda + 2) = 0$$

   Therfore $\lambda_1 = 8$ and $\lambda_2 = -2$ **these are the eigenvalues for matrix A.**

5. **Solve for the corresponding vector to each λ**

Solving for $\lambda = 8$, in this process we will call the matrix with substituted values: B.

$$\begin{bmatrix} 7-(8) & 3 \\ 3 & -1-(8) \end{bmatrix} = \begin{bmatrix} -1 & 3 \\ 3 & -9 \end{bmatrix}$$

We will assume the following $B\overline{X} = 0 \therefore$

$$\begin{bmatrix} -1 & 3 \\ 3 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Applying row reduction $3R_1 + R_2 = R_2$ to:

$$\left[\begin{array}{cc|c} -1 & 3 & 0 \\ 3 & -9 & 0 \end{array}\right] = \left[\begin{array}{cc|c} -1 & 3 & 0 \\ 0 & 0 & 0 \end{array}\right] \therefore -x_1 + 3x_2 = 0$$

From the previous operation we obtain $3x_2 = x_1$, at this point we can choose a value for any x, we will go for $x_2 = 1$ to keep it simple.

$$3x_2 = x_1 \therefore 3(1) = x_1 \therefore x_1 = 3$$

Now we know that the eigenvalue $\lambda = 8$ $ corresponds to the eigenvector $\overline{X} = (3, 1)$.

Solving for $\lambda - 2$, generating matrix C.

$$C = \begin{bmatrix} 7 - (-2) & 3 \\ 3 & -1 - (-2) \end{bmatrix}$$

$C\overline{X} = 0 \therefore$

$$\begin{bmatrix} 9 & 3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Applying row reduction $3R_2 - R_1 = R_1$:

$$\left[\begin{array}{cc|c} 0 & 0 & 0 \\ 3 & 1 & 0 \end{array}\right] \therefore 3x_1 + x_2 = 0$$

Assigning $x_1 = 1$:

$$x_2 = -3x_1 \therefore x_2 = -3(1)$$

The eigenvalue $\lambda = 8$ corresponds to the eigenvector $\overline{X} = (1, -3)$

2.2.5 5. Pick $m < d$ eigenvectors with highest eigenvalues.

In other words, usually the **2** eigenvectors with the highest scalars, or λ, will be selected to represent the whole dataset as Principal Component 1 and Principal Component 2.

2.2.6 6. Project datapoints to those eigenvectors.

One or the algoritm has to project the datapoints to these new set of dimensions so they can be analyized.

2.2.7 7. Perform analysis as needed according to study.

2.3 Pros and Cons of PCA

This algorithm, as all, is better suited for specific circumstances and performs poorly in others. The following list trys to summarize this idea:

2.3.0.1 Pros

- Reduction in size of data.
- Allows estimation of probabilites in high-dimensional data.
- It renders a set of components that are uncorrelated.

2.3.0.2 Cons

- It has a high computational cost, therefore it cannot be applied to very large datasets.
- Not good when working with fine-grained classes.

Chapter 3

Canonical Correlation Analysis (CCA)

3.1 What is CCA?

- Seeks the weighted linear composit for each variate (sets of D.V. or I.V.) to maximize the overlap in their distributions.
- Labeling of DV and IV is arbitrary. The procedure looks for relationships and not causation.
- Goal is to **maximize the correlation** (not the variance extracted as in most other techniques).
- Lacks specificity in interpreting results, that may limit its usefulness in many situations.

CCA helps us answer the questions:

- ***What is the best way to understand how the variables in two sets are related?*** , mathematically speaking:
  - ***what linear combinations of the set X variables (u) and the set Y variables (t) will maximize their correlation?***

3.1.1 Canonical R (R_c)

It represents the overlapping variance between two variates which are linear composites of each set of variables.

3.2 Assumptions for CCA

- Multiple continuous variables for DVs and IVs or categorical with dummy coding.
- Assumes **linear relationship** between any two variables and between variates.
- Multivariate normality is necessary to perform CCA.
- Multicollinearity in either variate **confounds** interpretation of canonical results.

3.3 Objectives of CCA

- Determine the magnitude of the relationships that may existe between two sets of variables.

- Derive a variate(s) for each set of criterion and predictor variables such that the variate(s) of each set is maximally correlated.
- Explain the nature of whatever relationships exist between the sets of criterion and predictor variables.
- Seek the max correlation of shared variance between the two sides of the equation.

3.4 Terms used in the context of a CCA analysis

- **Canonical correlation:** Correlation between two sets; the largest possible correlation that can be found between linear combinations.
- **Canonical variate:** The linear combinations created from the IV set and DV set.
- **Canonical weights:** weights used to create the liniear combinations; interpreted like regression coefficients.
- **Canonical loadings:** correlations between each variable and its variate; interpreted like loadings in PCA.
- **Canonical cross-loadings:** Correlation of each observed independent or dependent variable with opposite canonical variate.

3.5 Interpreting canonical variates

- Canonical weights
  - larger wight contributes more to the function.
  - negative weight indicates an inverse relationship with other variables.
  - always look out for multicollinearity, it can skew the whole analysis.
- Canonical Loadings.
  - A direct assessment of each variable´s contribution to its respective canonical variate.
  - Larger loadings are interpreted as more important to deriving the canonical variate.
  - Correlation between the original variable and its canonoical variate.
- Canonical Cross-Loadings
  - Measure of correlation of each original D.V. with the independent canonical variate.
  - Direct assessment of the relationship between each D.V. and the independent variate.
  - Provides a more pure measure of the dependent and independent variable relationship.
  - Preferred approach to interpretation.

3.6 Considerations when working with CCA

- Small samples sizes may have an adverse effect.
- Suggested minimun sample size = 10 * # of values.
- Selection of variables to be included:
- Select them with domain knowledge or theoretical basis.
- Inclusion of irrelevant or deletion of relevant variables may adversely affect the entire canonical solution.
- All I.V.s must be interrelated and all D.V.s must be interrelated.
- Composition of D.V. and I.V. variates is critical to producing practical results.

3.7 Limitations of CCA

- R_c (canonical R) reflects only the variance shared by the linear composites, not the variances extracted from the variables.
- Canonical weights are subject to a great deal of instability, particularly when there is multicollinearity.

- Interpretation difficult because rotation is not possible.
- Precise statistics have not been developed to interpret canonical analysis.

This chapter is under construction.

Figure 3.1:

Chapter 4

Self Organizing Maps (SOM)

4.1 Other names:

- Self-Organizing Feature Map (SOFM).
- Kohonen Map.
- Kohonen Networks.

4.2 Generalities

- **S**elf **O**rganizing **M**aps belong to the family of **Artificial Neural Networks**.
- In the subgroup of **Unsupervised Learning**,
- To function they use a **competitive learning strategy** (winer takes all).
- They are considered to be a **non-linear** implementation of the Principal Components Analysis (PCA) algorithm.

Self Organizing Maps (SOM) where first described by Teuvo Kohonen (Kohonen, 1995), others have extended his work and modified SOMs to tackle speficif problems.

"The Self-Organizing Map is inspired by postulated feature maps of neurons in the brain comprised of feature-sensitive cells that provide ordered projections between neuronal layers, such as those that may exist in the retina and cochlea. For example, there are acoustic feature maps that respond to sounds to which an animal is most frequently exposed, and tonotopic maps that may be responsible for the order preservation of acoustic resonances." (Brownlee, 2011) Different sensory inputs are maped into corresponding areas of the cerebral cortex in an orderly way. The map generated in the cerebral cortex is called a **topographic map** and it has two very important properties, (Bullinaria, 2004):

1. At each stage of representation, or processing, each piece of incoming information is kept in its proper context/neighbourhood.
2. Neurons dealing with closely related pieces of information are kept close together so that they can interact via short synaptic connections.

Following the principles observed in the sensory input processing by neurological structures , the previous two properties should be kept in an artificial intelligence algorithm looking to reproduce this phenomenon. In shorter words: the principle of topographic map formation is the escence of this process, where:

> "The spatial location of an output neuron in a topographic map corresponds to a particular domain or feature drawn from the input space." (Bullinaria, 2004)

Bibliography

Brownlee, J. (2011). Clever algorithms: nature-inspired programming recipes. Jason Brownlee.

Bullinaria, J. A. (2004). Self organizing maps: Fundamentals.

Kohonen, T. (1995). Self-organizing maps, volume 30 of springer series in information sciences.