Week 7b

Thursday, October 22, 2020 10:50 AM

document

Chapter 19

Week 7b: Principal Components Analysis

Schedule

19.1 Principal Eigenvectors as Directions with the Largest Variance [30 mins]

PCA at a high level

Recall the following properties related to PCA that we saw at the end of last class:

- In many applications, being able to reduce dimensionality of data is extremely helpful as it can lead to efficient representations and reduced computational complexity.
- Given some data, PCA enables us to express multidimensional data as a linear combination of orthonormal vectors, starting with the vector in the direction with most variation in the data. The next vector will be in the direction with most variation of all directions orthogonal to the first, and so on.
- So, if we want to work with a lower-dimensional representation of our data, we can focus on those directions that contain the most variation.
- The eigenvectors of the covariance matrix of the data are the principal component vectors. The eigenvector corresponding to the largest eigenvalue lies in the direction with most variation in the data set. The eigenvector corresponding to the second largest eigenvalue lies in the direction with the next most variation, of all directions orthogonal to the first eigenvector, etc.

The Principal Eigenvector as the Direction of Maximum Variance

The graphs of the daily temperature data show, graphically, that the principal eigenvector of the covariance matrix corresponds to the direction of maximum variation in the data. In this section we'll be formalizing this result. We've decided to structure this part of the day assignment as an extended exercise where you will be working through the proof of this fact step-by-step. While there are many ways to do this proof, we'll be walking you through one way that will connect well with the ideas we've been exploring in the last week or so of the course. We recommend that you do a part of the proof, check it against the solutions and then move onto the next piece.

Before getting started, let's look at some material from night 6 that shows that the covariance matrix can be computed using matrix multiplication.

Suppose that we have two different data variables x and y (e.g. corresponding to temperatures in Boston and Sao Paolo), with x_i and y_i being different values in the data set we can define a a matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \frac{1}{\sqrt{N-1}} \begin{pmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \\ \vdots & \vdots \\ x_N - \mu_x & y_N - \mu_y \end{pmatrix} \tag{19.1}$$

where μ_x is the mean of the first column, and N is the number of samples (rows). The covariance matrix of x and y is $\mathbf{R} = \mathbf{A}^T\mathbf{A}$. You can think of the entries of this matrix as storing the un-normalized correlations between the temperatures. Because $\mathbf{R}^T = \mathbf{R}$, this matrix is symmetric, and hence has orthogonal eigenvectors.

Let's assume that we are given a dataset with n samples and d dimensions (instead of just 2 dimensions as shown above). We can transform it into the form given in Equation 19.1 by subtracting the mean from each column and dividing the entire matrix by $\sqrt{N-1}$. We now have a mean-centered data matrix $\mathbf{A}$ with n rows and d columns and the covariance matrix of our data is given by $\mathbf{A}^\top\mathbf{A}$.

Exercise 19.1

Our overall goal is to show that if we take a unit vector $\mathbf{u}$, project our mean-centered data onto it (as $\mathbf{Au}$), and examine the variance of the projected data, that this variance is largest when $\mathbf{u}$ is the principal eigenvector of the covariance matrix $\mathbf{A}^\top\mathbf{A}$.

1. First we'll write down an expression for the variance of $\mathbf{Au}$ (we'll write this as $Var[\mathbf{Au}]$) as a matrix multiplication. We'll do this step together (i.e., we'll show you how to do it). For this part of the exercise you should make sure you understand the steps we performed.

   If $\mathbf{A}$ is in the form given in Equation 19.1, then $\mathbf{Au}$ will have 0 mean (since $\mathbf{Au}$ is a linear combination of columns with 0 mean). Using the same logic that led us to conclude that $\mathbf{A}^\top\mathbf{A}$ is the covariance matrix of the data, $(\mathbf{Au})^\top(\mathbf{Au})$ will give us the variance of the data projected onto $\mathbf{u}$ (remember that variance is just a special case of covariance where we are comparing a quantity to itself). It's worth noting that since $\mathbf{Au}$ is a vector, the expression $(\mathbf{Au})^\top(\mathbf{Au})$ is known as the inner product, which is really the same as the dot product (that is, $(\mathbf{Au})^\top(\mathbf{Au}) = \mathbf{Au} \cdot \mathbf{Au}$). Thus, the variance is given by the following equation.

   $$\begin{aligned} Var[\mathbf{Au}] &= (\mathbf{Au})^\top(\mathbf{Au}) \\ &= \mathbf{u}^\top\mathbf{A}^\top\mathbf{Au} \quad \text{note: we are applying the rule that } (\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top \end{aligned}$$

2. Substitute the eigenvalue decomposition, $\mathbf{VDV}^\top$, for the covariance matrix $\mathbf{A}^\top\mathbf{A}$ (since $\mathbf{A}^\top\mathbf{A}$ is symmetric and real, we can substitute $\mathbf{V}^\top$ for the inverse of $\mathbf{V}$ in the eigenvalue decomposition).
3. Define the vector $\mathbf{y} = \mathbf{V}^\top\mathbf{u}$ and substitute it into the expression from part 2.
4. Expand out the expression in part 3 so that it is in terms of the squares of the elements of $\mathbf{y}$ and the diagonal entries of $\mathbf{D}$ in order of largest to smallest.

5. Show that $\mathbf{y}$ is a unit vector by taking the inner product with itself and showing that it is equal to 1 (recall that the inner product is the same as the dot product). *Hint:* $\mathbf{VV}^\top = \mathbf{I}$ *since* $\mathbf{V}$ *is orthonormal and has* d *linearly independent columns.*
6. Argue that since $\mathbf{y}$ is a unit vector (which implies $\sum_{i=1}^{d} y_i^2 = 1$), that the expression in part 4 is maximized when $y_i = 1$ when i is the index of the principal eigenvector and $y_i = 0$ when i is any other index. To get a feel for why this is true, try writing out a specific case where, perhaps, $\mathbf{y}$ has two or three dimensions.
7. Show that we achieve the value of $\mathbf{y}$ in part 6 (that is where $y_i = 1$ when i is the index of the principal eigenvector and $y_i = 0$ when i is any other index) when $\mathbf{u}$ is the principal eigenvector of $\mathbf{A}^\top\mathbf{A}$.
8. What have you just shown?!? Make sure you have a sense of what you just did (don't get lost in the mathematical symbols).

Beyond the first principal component

We've now gone into depth in understanding the first principal component and its amazing property of maximizing variance. The second principal component is simply going to be the direction that maximizes variance subject to the requirement that it is orthogonal to the first principal component. With a slight modification to your proof you can show that the second principal component will be in the direction of the eigenvector with the second largest eigenvalue. The trend continues for other principal components (i.e., the ith principal component is the eigenvector with the ith largest eigenvalue).

19.2 Applications of PCA (thinking it through conceptually) [45 mins]

In this section you're going to be thinking about what the PCA algorithm might do when applied in different domains. The focus of this section will be on trying to understand at a conceptual level what might happen when we apply PCA. In the next section, you'll be reading through an example of applying PCA to some actual data.

Exercise 19.2

For each application, hypothesize what the first principal component might be. That is, for each particular scenario what would the direction be that maximizes the variance of the data projected

Exercise 19.2

For each application, hypothesize what the first principal component might be. That is, for each particular scenario what would the direction be that maximizes the variance of the data projected onto that direction? What might the second principal component be (that is a vector orthogonal to the first that maximizes the variance of the data)?

1. Consider a dataset consisting of ratings from n users of m movies. Let's assume that the ratings are numerical and are on a scale of 1 to 5 (5 being the best). Consider some collection of movies (they could be some specific movies or you could just think of movie genres) and a particular population of users (could be college students, QEA professors, or just the general population). Draw the data matrix $\mathbf{A}$ and label the rows and columns (e.g., with movies or users). In a qualitative sense, make a prediction as to what the first principal component would look like for this dataset. What might the second principal component look like? No numbers... just guess at which dimensions would be positive, negative, or close to 0 for your principal components.

2. Consider a dataset consisting of the prevalence of the flu in various parts of the US. The CDC maintains an animated map of the flu activity over time, which you can (and should) access at `https://www.cdc.gov/flu/weekly/usmap.htm`. To simplify this data, let's think about the number of flu cases in each of the six major major geographical regions of the US.

If we think about our data matrix as consisting of a row for each week of measured flu activity and each column as a region of the US, in a qualitative sense, make a prediction as to what the first principal component would look like for this dataset. What might the second principal component look like? No numbers... just guess at which dimensions would be positive, negative, or close to 0 for your principal components.

Exercise 19.3

With your table-mates, read through this post that shows the application of PCA to understanding the US political leanings (if you are viewing this in DropBox preview and can't click the link, go to `http://bit.ly/37n9qwe`). Before, starting here are some process suggestions.

- Checkin with folks at your table as to how they'd like to go through this document (e.g., read the entire thing individually and come together and ask questions, read it individually but stop after each major section to ask questions, read it aloud as a table).
- If you don't understand something, you can either call over an instructor or note your confusion on the whiteboard and keep going (e.g., if its something that doesn't impede your understanding of the main points in the article).

Solution 19.1

1. Solution is already given in the problem

2.

$$Var[\mathbf{Au}] = \mathbf{u}^\top \mathbf{VDV}^\top \mathbf{u}$$

3.

$$\begin{aligned} Var[\mathbf{Au}] &= (\mathbf{V}^\top \mathbf{u})^\top \mathbf{D}(\mathbf{V}^\top \mathbf{u}) \\ &= \mathbf{y}^\top \mathbf{Dy} \end{aligned}$$

4.

$$\begin{aligned} Var[\mathbf{Au}] &= \mathbf{y}^\top \mathbf{Dy} \\ &= \mathbf{y}^\top \begin{bmatrix} y_1 D_{1,1} \\ y_2 D_{2,2} \\ \vdots \\ y_d D_{d,d} \end{bmatrix} \\ &= \sum_{i=1}^{d} y_i^2 D_{i,i} \end{aligned}$$

5.

$$\begin{aligned} \mathbf{y}^\top \mathbf{y} &= (\mathbf{V}^\top \mathbf{u})^\top (\mathbf{V}^\top \mathbf{u}) \\ &= \mathbf{u}^\top \mathbf{VV}^\top \mathbf{u} \\ &= \mathbf{u}^\top \mathbf{u} \\ &= 1 \end{aligned}$$

6. If we choose $y_i = 1$ where i is the index of the principal eigenvector, then the expression in part 4 will give us $D_{i,i}$. Any other choice of $\mathbf{y}$ will result in some weighted combination of the eigenvalues (the diagonal elements of $\mathbf{D}$) where the weights are all positive and add up to 1. It is easy to see that putting any weight on a non-maximal eigenvalue will result in a lower variance as computed by the expression in part 4.

7. Since $\mathbf{y} = \mathbf{V}^\top \mathbf{u}$, y_i is the dot product of $\mathbf{u}$ and the ith eigenvector, $\mathbf{v}_i$, with $\mathbf{u}$. Since we assume all of the eigenvectors are unit vectors and mutually orthogonal, if we set $\mathbf{u}$ to be the principal eigenvector of $\mathbf{A}^\top \mathbf{A}$, then the dot product of $\mathbf{u}$ and $\mathbf{v}_i$ will be 1 for i corresponding to the principal eigenvector and 0 for all other indices.

8. You just showed that the direction along the principal eigenvector of the covariance matrix maximizes the variance of the projected data. That's pretty cool!