Discussion 11A

Covariances Matrices and Principal Component Analysis

Guide to the advance lecture notes (posted on Piazza thread @819):

(a) First of all, realize that numerical "data" can be organized as rectangular matrices; *e.g.*, by having each row represent a data "point", and having a large number of rows. [Page 2 of the notes, top half - (52.5)]. Call the data matrix A, of size n rows and m columns. Usually, $n > m$. We will refer to each column of A as a "dimension" of the data.

(b) The next concept is that of forming a *covariance matrix* from A. This takes two steps:

   i. First form a *mean-centered version* of the data matrix, *i.e.*, take each column of A, find the mean of all the entries in the column, and subtract this column mean from each entry of the column. Call this mean-centered matrix $\tilde{A}$. [Pages 2–3, (53)–(55.2)].

   ii. Then form the covariance matrix: $S \triangleq \frac{1}{n}\tilde{A}^T\tilde{A}$. [Page 3, (56)–(57.3)].

   The covariance matrix S has some nice properties [Pages 4–5, (60)–(62.81)]: it is symmetric, the diagonal entries are all ≥ 0 and are, in fact, the *variances* of the columns of A, *etc.*.

(c) Why is the covariance matrix interesting or important? Well, the diagonal entries are the column variances of the data, which might be somewhat useful. But if we develop the notion a bit further, *i.e.*, use it to define something called the *correlation matrix*, then it becomes much more interesting. The correlation matrix R is defined in [Page 6, (63)] – it is essentially the covariance matrix S, but with the entries within divided by the diagonal entries in a particular way. R is also symmetric, and its $(i, j)^{\text{th}}$ entry is called the *correlation between data dimensions i and j*. Correlations are always between 1 and -1 [(64)].

(d) It turns out that correlations indicate how well data fits on a straight line. For example, if the data is exactly on a straight line [Page 7, (70)], then the correlation is exactly +1 (if the slope is positive) or -1 (if negative). If the data is roughly a circular blob, then the correlation is 0, or close; if it is a blob around a straight line, then it is some number in between [Page 7, (70)]. Correlation is, in fact, widely used in many fields to get a quick sense of the relationship between different dimensions of data.

(e) But correlation has its limitations. For example, if the data is exactly on a *horizontal* line, then the correlation becomes undefined (sometimes taken to be 0 – the same as for a circular blob). It can switch abruptly from +1 to -1 if the slope of the data changes slightly [Page 7, (73)].

(f) This is where PCA can be much more useful. PCA consists simply of eigendecomposing the covariance matrix S. The resulting eigenvectors are called *principal components*, and they are ordered by the values of the corresponding eigenvalues (which are all real and ≥ 0), in decreasing order [Page 11, (86)–(91)]. It turns out that the principal components (which are all orthogonal to each other) capture the directions of maximum spread of the data [Page 8, (74)–(75)]. More precisely, the first principal component (the eigenvector corresponding to the biggest eigenvalue) is the direction along which the data is most spread; and the data's variance along that direction is simply the corresponding eigenvalue [Page 12, (94)–(94.3)]. The second principal component captures the direction in the space *orthogonal to the first PC* that maximizes the spread (variance) along any direction in the space. And so on with the third, fourth, *etc.*, principal components [Page 14, (97)]. Thus, PCA gives us a much more precise way to understand how the data is distributed – in many dimensions – than simply correlations.

(g) In the above, we developed PCA without thinking about the SVD at all. But it turns out that PCA and the SVD are intimately related; if you do one, you have effectively done the other [Page 15, (98)–(102)]. For example, the PCA eigenvalues are the singular values (of A) squared and divided by the number of data points n [(102)].

(h) And it turns out that if you want to compute the SVD of A, you can adapt the above PCA connection to devise a procedure based on eigendecomposing A^TA (or AA^T) and using Gram-Schmidt orthonormalization [Pages 16–18, (103)–(123)].

Questions

1. PCA and Clustering Artificial Data - IPython Notebook

2. Calculating SVD from PCA

Say we have the $n \times m$ matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \\ 5 & 10 \end{bmatrix}$$

which we want to analyze. Let's take a look at different ways to analyze this matrix, leading up to calculating the singular value decomposition via principal component analysis.

(a) **Find the mean-centered data matrix $\widetilde{A}$.** Recall that $\widetilde{A}$ is found by subtracting off the mean of each k^{th} column (denoted μ_k) from the corresponding k^{th} column.

(b) **Calculate the covariance matrix S of A.** Recall that that the covariance matrix is given by

$$S = \frac{1}{n}\widetilde{A}^T\widetilde{A}.$$

(c) **Calculate the correlation matrix R of A.** Recall that the correlation matrix of A is calculated from the covariance matrix, S, where the diagonal is all 1 and the off-diagonal i^{th} row and j^{th} column entry of R is calculated as

$$R_{ij} = \frac{S_{ij}}{S_i S_j}.$$

The correlation is a normalized version of covariance, which makes it possible to compare relative correlation among different variables. Here

$$S_i = \sqrt{S_{ii}},$$

and is the standard deviation of each column of A. That is, the diagonal terms of S are denoted $S_{ii} = S_i^2$, which is the variance of each column of A.

(d) **Calculate AA^T and A^TA.** What are some properties you observe about these matrices?

(e) Principal Component Analysis (PCA) is basically nothing more than the eigendecomposition of the covariance matrix S. As we saw above, though, the matrix $T = A^TA$ contains the same information as S, so we can also apply PCA to a data matrix that is not mean-centered. We can write the eigendecomposition of T as

$$T = P\Omega P^T.$$

Find the eigenvalues and eigenvectors of T. Compose the matrices P and Ω. Remember to orthonormalize P.

(f) In the SVD, we are decomposing $A = U\Sigma V^T$. We choose $V = P$ and the singular values $\sigma_i = \sqrt{\lambda_i}$. Therefore

$$V = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \tag{1}$$

$$\Sigma = \begin{bmatrix} \sqrt{175} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{2}$$

We're going to find U which completes the SVD decomposition. **Calculate the first column of U, $\vec{u}_1$, in terms of A, the first singular value σ_1, and the first column of V, $\vec{v}_1$.**

(g) **Find the other columns of U such that the columns are all mutually orthogonal and normalized.**

Contributors:

- Jaijeet Roychowdhury.
- Saavan Patel.
- Regina Eckert.