WEEK 8 ASSIGNMENT

Data Systems in the Life Sciences – BIOL 51000 | Fall 2 2020

Christina Morgenstern

DATA DISCRIMINATION METHODS

Data discrimination refers to the process of finding most discriminating features of a data set that yields the best separation of the data objects. In subsequent sections, two data discrimination methods will be discussed: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Both techniques help to reduce the dimensionality of the data which reduces the number of variables by obtaining a set of principal variables. This leads to a transformation of the data from a high-dimensional space into a low-dimensional space with the goal of retaining the properties that explain most of the variation [1]. Reducing the number of dimensions in a data set is often desirable because high-dimensional data is often computationally intractable and the curse of dimensionality which refers to properties of the data that are present at high-dimensional spaces but not in low-dimensional settings. Dimensionality reduction is often performed as a data pre-processing step for pattern-classification and machine learning applications [2]. In exploratory data analysis it is helpful in reducing noise, for data visualization and for clustering.

1. Introduction to Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used and easy to implement data reduction technique applied in unsupervised data compression. It allows us to extract the features from the data which explain most of the variability and helps to identify patterns within the data. These features are vectors that separate the data objects best with regard to their covariances. We call them principal components. Mathematically, they are eigenvectors of the covariance matrix which are arranged in size with the first ones explaining most of the variability within the data [3]. Principal components are vectors that refer to the directions within the data and are perpendicular to one another with each component vector representing an independent axis. Theoretically, there can be as many principal components as there are dimensions within the data [3]. In Figure 1 an interpretation of the principal components PC1 and PC2 is shown in a new subspace. λ 1 and λ2 refer to the principal components PC1 and PC2 which point in the direction of maximum variance.

In order to perform a PCA, the following steps are taken [4]:

1. Standardize the data.
2. Calculate the covariance matrix for the features in the dataset.
3. Calculate the eigenvalues and eigenvectors for the covariance matrix.
4. Sort the eigenvalues and corresponding eigenvectors in descending order.
5. Pick *k* eigenvalues and form a matrix of eigenvectors.
6. Transform the original matrix.

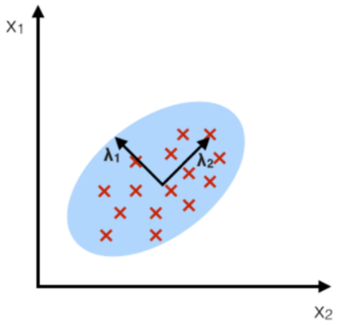

Figure 1. PCA component axes that maximize the variance (taken from [2]).

2. Extracting principal components

This section explains step-by-step the process of principal component analysis as listed above in section 1. Since PCA is highly sensitive to data scaling, the data needs to be standardized before applying PCA. This means if there are initial variables in the data set that have large differences between the ranges these might dominate in the analysis. Thus, the data need to be transformed to a comparable scale. By subtracting the mean and dividing by the standard deviation for each value of each variable the data is standardized and are now on the same scale. Figure 2 highlights the formula used for data standardization.

$$z = \frac{value - mean}{standard\ deviation}$$

Figure 2. Formula used for data standardization (taken from [5]).

In step 2, the covariance matrix is computed with the goal to understand if the variables in the dataset are varying from the mean or if there is any relationship between them. With the help of a covariance matrix such correlations can be identified. Figure 3 shows a covariance matrix for 3-dimensional data.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Figure 3. Covariance matrix for 3 variables, *x, y* and *z* (taken from [5]).

The value of the covariance matrix already suggest the type of correlation: if positive, then the variables are correlated to each other, if negative they are inversely correlated to each other [5]. In step 3 the eigenvectors and eigenvalues of the covariance matrix are computed, and the principal components identified. See Figure 4 for an example of eigenvectors and eigenvalues. As pointed out earlier, the principal components represent the directions of the data that explain a maximal amount of variance and thus hold most information of the data.

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

Figure 4. Example eigenvectors and associated eigenvalues (taken from [5]).

Ranking the eigenvectors in order of their eigenvalues (step 4), from highest to lowest, yields the principal components in order of significance. In step 5, the most important components i.e. the ones that are able to explain most of the variability in the data are chosen and the eigenvalues ranked lower are discarded. The remaining chosen eigenvalues make up a matrix that corresponds to the feature vector (see Figure 5).

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Figure 5. Example of a feature vector (taken from [5]).

In the last step, the feature vector is used to transform the data from the original axes to the axes represented by the principal components. This can be achieved by multiplying the transpose of the original data by the transpose of the feature vector [5].

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Figure 6. Transformation of the data taking into account the feature vector (taken from [5]).

3. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) was formulated by Ronald A. Fisher in 1936 [6] and it is most commonly used as a dimensionality reduction method but has also some practical applications as classifier [2]. Originally, it was formulated for a 2-class problem and only later in 1948 generalized as multi-class LDA by C. R. Rao [7]. LDA is similar to PCA in the way that both are linear transformation techniques used for dimensionality reduction. While PCA can be seen as an unsupervised algorithm which aims at finding the principal components with maximum variance within the dataset, LDA is a supervised technique which computes the linear discriminants that will separate the classes [2]. Thus, PCA tries to find the axis that explain most of the variability within the data and LDA finds the axes that separate different classes best (compare Figure 1 and Figure 7).

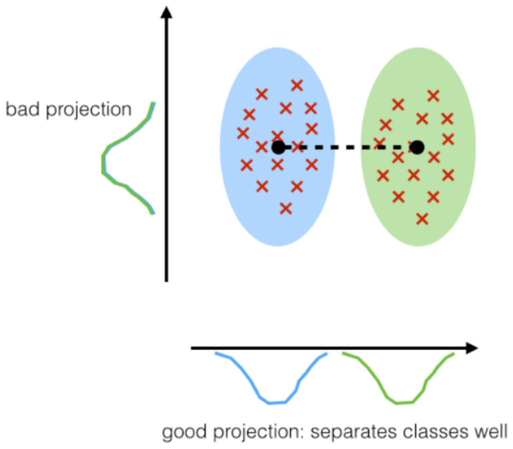

Figure 7. LDA maximizes the component axes for class-separation (taken from [2]).

Assumptions that need to be made before applying LDA are that the data is normally distributed, with identical covariance matrices of the classes and that the training examples are statistically independent of each other. Although researchers have shown that if some of these assumptions are slightly violated, LDA still performs well [8]. Figure 7 demonstrates the concept behind LDA for a 2-class problem where the blue and green areas denote the two classes. A linear discriminant on the x-axis separates the two normally distributed classes well while a linear discriminant on the y-axis captures a lot variance in the dataset but fails to discriminate between the two classes [8].

The process of applying LDA as a data reduction technique is to compute the within-class and between-class scatter matrices. For these scatter matrices, the eigenvectors and corresponding eigenvalues are calculated. Subsequent sorting in descending order and selecting the top *k* eigenvectors corresponding to the *k* largest eigenvalues. With these eigenvectors and eigenvalues, a *d x k*-dimensional transformation matrix is constructed and the data projected into a new feature subspace [8].

PCA and LDA analysis are both linear dimensionality techniques but LDA also considers the class labels as a supervised algorithm. Implementation of PCA and LDA can be performed in Python using the scikit-learn library [9].

Bibliography

[1] 'Dimensionality reduction', *Wikipedia*. Dec. 02, 2020, Accessed: Dec. 18, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Dimensionality_reduction&oldid=991883514.

[2] 'Linear Discriminant Analysis', *Dr. Sebastian Raschka*, Aug. 03, 2014. https://sebastianraschka.com/Articles/2014_python_lda.html (accessed Dec. 18, 2020).

[3] T. J. Stevens and W. Boucher, 'Python Programming for Biology: Bioinformatics and Beyond', *Cambridge Core*, Feb. 2015. /core/books/python-programming-for-biology/61762A9F672FDD8B2DD3FFF8773027B2 (accessed Nov. 07, 2020).

[4] T. Nobles, 'Understanding Principle Component Analysis(PCA) step by step.', *Medium*, Jan. 16, 2020. https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9 (accessed Dec. 18, 2020).

[5] 'A Step-by-Step Explanation of Principal Component Analysis', *Built In*. https://builtin.com/data-science/step-step-explanation-principal-component-analysis (accessed Dec. 18, 2020).

[6] R. A. Fisher, 'The Use of Multiple Measurements in Taxonomic Problems', *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936, doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[7] C. R. Rao, 'The Utilization of Multiple Measurements in Problems of Biological Classification', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, pp. 159–203, 1948.

[8] 'Data Compression via Dimensionality Reduction: 3 Main Methods', *KDnuggets*. https://www.kdnuggets.com/data-compression-via-dimensionality-reduction-3-main-methods.html/ (accessed Dec. 19, 2020).

[9] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.