Ministry of Education, Culture, and Research of the Republic of Moldova

Technical University of Moldova

Department of Software Engineering and Automatics

Study Program: Software Engineering

Report

Data analysis and visualisation

Done by: Ion Dodon, IS211-M

Verified by: Grozavu Nistor

Chisinau, 2021

Laboratory work no. 1

Theoretical material

Feature scaling - Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.[1]

It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately).

Partial component analysis - The principal components of a collection of points in a real coordinate space are a sequence of unit vectors, where the -th vector is the direction of a line that best fits the data while being orthogonal to the first vectors. Here, a best-fitting line is defined as one that minimizes the average squared distance from the points to the line. These directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated. Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

Linear discriminant analysis - Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable (*i.e.* the class label). Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable

in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

Conclusions

Working on this laboratory work I used different dimensionality reduction techniques and analyzed the differences between them. I have understood how to print useful information about a dataset like feature names, the classes, the number of objects, etc., and how to plot the clusters on subplots.

I compared the correlations between all possible combinations of variables from the iris dataset and chose the best pair of features by analyzing the correlation given by *corrcoef* function and by analyzing the clusters on the plots. Also, I practiced how to display the number of images from the MNIST dataset.

Bibliography

- https://stats.stackexchange.com/questions/109071/standardizing-features-when-using-lda-as-a-pre-processing-step
- https://en.wikipedia.org/wiki/Feature_scaling#:~:text=Feature%20scaling%20is%20a%20method,during%20the%20data%20preprocessing%20step.
- https://en.wikipedia.org/wiki/Principal_component_analysis#:~:text=Principal%20component%20analysis%20(PCA)%20is,components%20and%20ignoring%20the%20rest.
- http://yann.lecun.com/exdb/mnist/