Tutorial Business Analytics

Tutorial 10 - Tutorial

Exercise 10.1 Compute the Principal Components

Given the following dataset D = {(-3, -1, -1), (0, -1, 0), (-2, -1, 2), (1, -1, 3)}, compute its principal components by following the PCA algorithm introduced in class and generate the transformed data.
Each tuple of the set D stands for an observation or row vector.

a) Calculate the zero-mean dataset X from the given dataset D. Note down the means.

b) Calculate the 3 x 3 covariance matrix ∑ using the following formulas. What can you infer from it?

$$var(x_j) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)^2$$

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij_1} - \bar{x}_{j_1}\right) \cdot \left(x_{ij_2} - \bar{x}_{j_2}\right)$$

Reminder: Since the matrix X is centered you can use the following formulas:

$$var(x_j) = \frac{1}{N-1}\sum_{i=1}^{N} {x_{ij}}^2$$

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij_1}\, x_{ij_2}$$

c) Find the eigenvalues for the covariance matrix by solving the equation: $|\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3}| = 0.$

d) Find the corresponding eigenvectors and order them by significance. How is the variance distributed among them?

Hint: Solving the equation $(\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3})v = 0$ gives you the corresponding eigenvectors.

e) Compute a one-dimensional PCA projection of the dataset.

f) Compute a two-dimensional PCA projection of the dataset.

Hint for e) and f): The general formula for projections is: $Z = X\Phi$

Exercise 10.2 Reconstruction of the Original Data

Making use of the PCA projections computed in Exercise 10.1, restore the original dataset using the formula: $D \approx Z\Phi^T + means$

a) Reverse the one-dimensional PCA projection to restore the original data. How would the data look when plotted into the original coordinate system?

b) What result do you expect when reconstructing the original data from the two-dimensional PCA projection? What is the information loss?

Exercise 10.3 Principal Component Regression vs Linear Regression

Install/open the “AER” (Applied Econometrics with R) package and open the “HousePrices” data set, which holds information about the prices of houses sold in Canada during three months in 1987.

a) Check the structure of the dataset. Filter the numerical attributes and discard the rest.

```
HousePrices <- HousePrices[,unlist(lapply(HousePrices, is.numeric))]
```

b) Build a model to predict the price of a house given the other independent variables using principal component regression with one component. How much of the dependent variable is explained by the model?

```
pcr_auto <- pcr(price~., data=HousePrices, scale=TRUE, ncomp = 1)
```

c) Build a model to predict the price of a house using simple OLS regression for each independent variable separately. Which OLS model explains best the price? What percentage of variation is explained in this case?

d) Compare the models derived from b) and c). Which one would you choose in this scenario? Give reasons.

Note: Use R to solve this exercise (Exercise10.3_R_template.R).