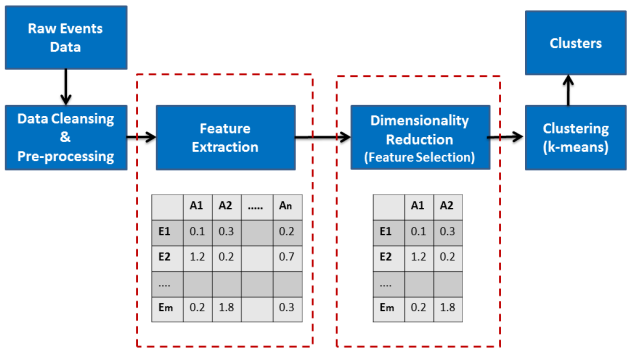
Department of Computer Engineering
University of Peradeniya

Data Mining and Machine Learning
Lab 09

September 19, 2017



1 Transformation

In computing, data transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.

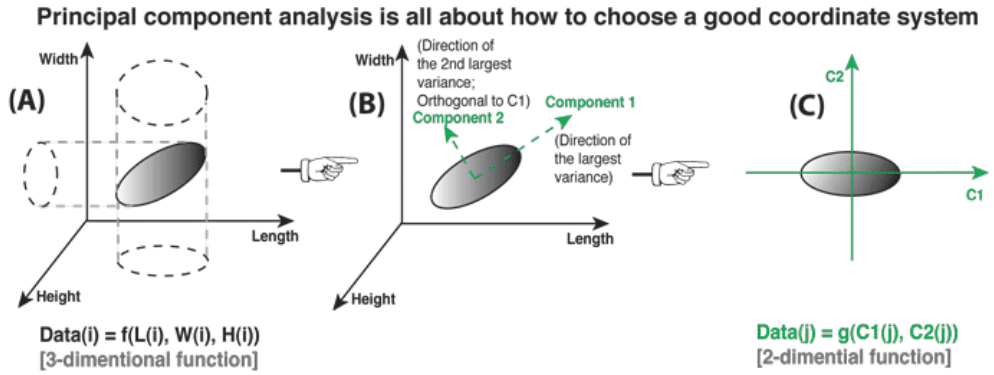
2 Feature selection

Feature Selection methods in Data Mining and Data Analysis problems aim at selecting a subset of the variables, or features, that describe the data in order to obtain a more essential and compact representation of the available information. The selected subset has to be small in size and must retain the information that is most useful for the specific application.

3 Try Out

3.1 Principal Component Analysis

Principal Component Analysis (PCA) could be used to reduce a large number of features into a much smaller number of features called Principal Components (PCs) while maximizing the variance of the original features set. PCA might also be useful to visualize the data by considering only the first two or three PCs.



1. First, import the necessary libraries that you require for the analysis.

```
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA
```

2. Load iris dataset by loading the data set as given..

```
from sklearn.datasets import load_iris
iris = load_iris()
X = iris.data
X.shape
df1 = pd.DataFrame(X)
```

3. Decide number of components to keep. if n-components is not set then all components are kept. Here we consider the first two principal components.

```
pca = PCA(n_components=2)
pca.fit(X)
```

4. Lets reduce the number of features in the original data set. Here X is projected on the first two principal components previously extracted from the training set.

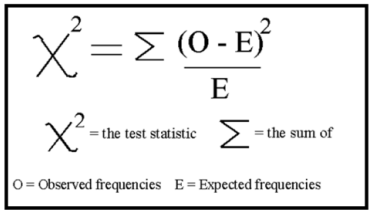
```
pca.transform(X)
df2 = pd.DataFrame(pca.transform(X))
```

5. The amount of variance explained by each of the selected components can be taken from the following.

```
print(pca.explained_variance_ratio_)
```

3.2 Chi-squre measure

The classes in the sklearn.feature-selection module can be used for feature selection/dimensionality reduction on sample sets. One of the method of Univariate feature selection works by selecting the best features based on univariate statistical tests. $SelectKBest$ removes all but the k highest scoring features.This score can be used to select the $n - features$ with the highest values for the test chi-squared statistic from X.



1. First, import the necessary libraries that you require for the analysis.

```
import numpy as np
import pandas as pd
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
```

2. Load iris dataset by loading the data set as given..

```
from sklearn.datasets import load_iris
iris = load_iris()
X, Y = iris.data, iris.target
X.shape
df1 = pd.DataFrame(X)
```

3. We can perform a chi-square test to the samples to retrieve only the two best features as follows.

```
X_2 = SelectKBest(chi2, k=2).fit_transform(X, Y)
X_2.shape
```

4 Lab Exercise

1. Load the colonTumor.csv data set.
2. Record the nearest neighbor classification accuracy of the data set using 10-fold cross validation.
3. Select the first 10 principal components. Record the 10-fold cross validation. Change the number of selected components as 1, 2, 3, . . . , 10, 50, 100 and record the accuracies. Comment on the accuracies when changing the number of components. Please note that this it may take considerable time and might not be possible in low performance computer.
4. Visualize the first two components. Comment on the output.
5. Now lets consider selecting the best set of features based on some measure. Consider chi-squre measure to keep the best set of features. Change the number of selected features as 1,2,3, ..., 10, 50, 100 and record the accuracies. Comment on the accuracies when changing the number of components.
6. Compare the accuracies obtained using Principal Components and feature selection using chi-square measure. Which method is best suitable for the given data set? Justify your answer.

5 Submission

Submit a single text file as [12|13]xxxlab09.py where xxx is your registration number. Add answers for questions in the same file as comments.

6 Important

Make sure that you have the basic understanding of dimensionality reduction/feature selection. If you do not understand any concepts, make sure you get some help from instructors.

7 Deadline

September 26, 23:59:59 GMT+5:30.