Partitioning the Parts of a Pizza: Running

Amar Lakshya (Student ID: 2096845)

School of Computer Science, University of Birmingham

Abstract

Pizza is one of the most popular fast food in the world. In this paper, experiments are conducted over a sample pizza dataset to develop a better understanding of the various attributes of pizza and their relationships. A number of data-analysis techniques like Principal Component Analysis and co-ordinate projections are employed to understand the characteristics of 10 anonymized brand names in the dataset.

Keywords: Food, Pizza, PCA

1. Introduction

This is the introduction. Here is a reference to [1] and [2].

2. Assignment

2.1. Submission deadline

The submission deadline for this assignment is 5pm on Friday 8th March (week 8).

2.2. Assignment details

You can chose any of the data set(s) from the list bellow, but finding your own dataset to investigate is much preferable! For example, there are many potential text data sets (Twitter, online reviews, etc) and these can be converted to vector format using methods from the course.

If you decide to go the easier route and use some of the data sets provided, un-tar the relevant file and go to the corresponding folder. The folder contains the data set, as well as additional information about the data. Read the available information, especially the description of the features (data dimensions).

Depending on the software that you use you will need to clean the data, for example so that it contains only numerical features (dimensions) and the features are space-separated (not comma-separated).

To make the plots informative, you should come up with a labeling scheme for data points. If the data can be classified into several classes (find out in the data and feature description!), use that information as the basis for your labeling scheme. In that case exclude the class information from the data dimensions that you analyse. Alternatively, you can make labels out of any dimension, e.g. by quantising it into several intervals. For example, if the data dimension represents age of a person, you can quantise it into 5 labels (classes) [child, teenager, young adult, middle age, old]. Associate the data labels with different markers and use the markers to show what kind of data points get projected to different regions of the visualization plot.

Learn as much as you can about an assigned data set(s) using the visualization and clustering methods developed in the module. Use various data labeling schemes. In the case of visualization, compare PCA with straightforward co-ordinate projections Data sets:

- abalone
- boston
- comp.activ
- image
- iris
- letter

Before starting to work on the assignment, please carefully study the example that Dr Tino prepared using the boston database. Un-tar the file boston.ex.tar.gz and go to the folder “BOSTON.EX”. The sub-folder “FIGURES” contains all the relevant figures as eps or gif files. Please consult the “boston.read.me” file in BOSTON.EX.

The report should describe experiments with a chosen data set(s) along the lines of the “boston” example. In the labeling scheme, concentrate on more than one coordinate (dimension), e.g. in the “boston” example don’t just consider the ‘price’ feature, but run separate experiments with ‘per capita crime rate in the town’, or ’pupil-teacher ratio in the town’ instead of the ‘price’ coordinate.

2.3. Marking scheme

In the report concentrate on the following questions:

- How did you pre-process the data? (worth 20
- What features (coordinates) did you use for labeling the projected points with different markers? (worth 10
- How did you design the labeling schemes? (worth 20
- What interesting aspects of the data did you detect based on the data visualisations? (worth 30
- What interesting aspects of the data did you detect based on eigenvector and eigenvalue analysis of the data covariance matrix? (worth 20

You should demonstrate that you

- Understand the visualisation techniques used
- Are able to extract useful information about otherwise inconceivable high-dimensional data using dimensionality-reducing visualisation techniques.
- The overall mark m (in the range 0-100%) will be linearly scaled to the range 0-20% by 0.2*m.

For examples of nice past reports developed on the ”wine dataset” (do not use this data in your report!), please see reports by Christoph Stich and Josephf Preece.

3. Acknowledgement

I would like to thank Dr Peter Tino for permission to use this assignment, and Christoph Stich and Joseph Preece for their example reports.

4. Including figures

Here is a figure:

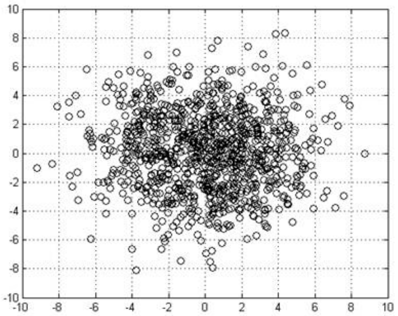

Figure 1: An example figure (JPG)

References

[1] Baber, C., Russell, M., Wing, A., Hermsdorfer, J. & Khattab, A, *Creating Affording Situations: Coaching through Animate Objects*, Sensors, **17**, 2308, 2017

[2] Dolan S. *mov is Turing-complete*. Cl. Cam. Ac. Uk. 2013:1-4