Project 1: Applied k-means clustering
Data Mining in Bioinformatics

Dr. Oliver Stegle
oliver.stegle@tuebingen.mpg.de

Dr. Karsten Borgwardt
karsten.borgwardt@tuebingen.mpg.de

http://agbs.kyb.tuebingen.mpg.de/wikis/bg/DataMiningInBioinformatics0910

WS 2009/2010

1 Remarks

The purpose of the tutorials is to familiarise yourself with one of the key algorithms that are widely applied in data mining and bioinformatics. Each exercise consists of 3 work packages. At the end of the course you will present your results for each of the work package in a short seminar presentation (15 min talk, 5 min questions). The presentation will count towards the overall course grade with the weight of 1/3, the remaining 2/3 are based on the oral examination.

1.1 Work packages

1.1.1 Theory/book work

- Thoroughly study your assigned algorithm. How does it relate to other methods?
- Explain the strengths and weaknesses in relation to other algorithms presented in the lecture.
- Conduct a short literature review listing the most important applications of the algorithm.
- (Optional: current research directions and extensions)

1.1.2 Problem

Implementation

- Implement the assigned algorithm yourself in a programming language of your choice.

Application

- Apply your implementation to a small research problem.
- What are the limitations of your algorithm for the particular problem?

1.2 Important dates

- **Progress meeting**
  At the end of week one we meet in person, either on Thursday 4 March or Friday 5 March (signup sheet) to discuss the project progress and potential problems. **It is essential that you start your project well before that day, such that this meeting is helpful for you.**
- **Seminar presentations**
  Thursday & Friday 11-12 March, 12:00 – 14:15. If you are facing problems with your project please bring them up at the progress meeting or email.

2 Problem

This project is concerned with the application k-means clustering. Investigate this algorithm and apply it to the Stockori floweringtime dataset (Appendix B).

References There exist a number of good references and tutorial for k-means clustering. A good place to start is MacKay (2003), chapter 20. Also the tutorial in Wu et al. (2008) provides a broader overview of k-means.

Implementation First, implement the standard k-means clustering algorithm. You can for example follow the description in MacKay (2003).

Application to Stockori Apply your k-means implementation to the genotypes from the stockori dataset (*genotypes.csv*). To measure the distance between datapoints and the centroids use the standard euclidean distance:

$$D(\mathbf{x}, \mathbf{c}) = \sqrt{\sum_{i=1}^{D} (x_i - c_i)^2}. \tag{1}$$

Initialization Compare alternative initializations

- Set cluster centers randomly.
- Pick existing data point as cluster centers.

Number of clusters Try out what happens if you vary the number of clusters. Propose a heuristics to choose a suitable number?

Relationship to countries

- Investigate the relationship of your clustering results with the country of origin of each plant (*countries.csv*).
- What relationship would you assume a priori?
- Visualize your clustering results in relation to the country labels (you can either do this directly in java or export your clustering results as CSV file and visualize in Excel or similar).
- Give a short biological interpretation of your finding.

A Useful resources

Depending on your chosen programming language you may need some basic tools to make your life easier. Below a list with some pointers for python and java. It is encouraged that you implement your project in Java.

A.1 Java

- There are a numer of packages to read CSV files, for example `http://opencsv.sourceforge.net/`.
- To calculate distances and perform matrix-vector operations you should use an established package. Jama (`http://math.nist.gov/javanumerics/jama/`) is a simple and practical solution which allows you to work with vectors and matrices.

A.2 Python

- Python comes with an integrated reader for CSV files.
- For matrix operations and rapid development you may want to use scipy/numpy. PythonXY(`http://www.pythonxy.com/`) is a good starting point for windows. For Mac there exist similar packages (`http://macinscience.org/?page_id=6`).

B Stockori dataset

The dataset, downloadable from (Stockori) contains information for 697 plants.

- *genotype.csv*
  CSV file with 149 genotypes for each plant ([697 x 149]).
- *floweringtime.csv*
  CSV file with floweringtimes (in days) for each plant ([697 x 1]).
- *floweringtime_binary.csv*
  CSV file with floweringtimes (binary) for each plant ([697 x 1]). A value of 0 corresponds to fast flowering plants, a value of 1 to slow flowering plants.
- *country.csv*
  CSV file with the country origin of all plants ([697 x 1]).

References

D. MacKay. *Information theory, inference, and learning algorithms*. Cambridge Univ Press, 2003. URL http://www.inference.phy.cam.ac.uk/itprnn/book.html.

Stockori. Stockori qtl dataset. URL http://agbs.kyb.tuebingen.mpg.de/wikis/bg/stockori.zip.

X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008. URL http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf.