Python exercises for the lecture "Models and Algorithms in Bioinformatics"

Organizers: Peter Meinicke, Sina Garazhian
Submission date: 18.01.2025, 12 p.m.

Sheet 4: Mixture Models

Task 1: EM-Algorithm

a) Implement the "soft clustering" EM algorithm as described in the lecture with multivariate Gaussian densities for the K mixture components. When calculating the assignment probabilities, note which technical (numerical) problems can arise and how can they be avoided? How does K-means clustering arise as a special case of the EM-algorithm?
First, test the method with $K = 4$ on the synthetic data from Sheet 3, Task 1a) using a maximum number of 100 iterations. Visualize the natural logarithm of the likelihood $\prod_{i=1\ldots N} p(x_i)$ as depending on the number of iterations in a function plot. Use the K-means algorithm from Sheet 3 for a pre-clustering to initialize the assignment variables h_{ij} .

b) Test the procedure with the gene expression data for different prototype numbers $K = 2, \ldots, 10$. Again use the K-means algorithm for initialization of the h_{ij}.

c) Test the procedure with $K = 10$ prototypes on the "Word Count" data from a Metagenome Binning evaluation [1]. The corresponding data matrix (data vectors are rows) can be found as `CountData.csv` in the StudIP Data folder.

The data entries actually do not directly correspond to (integer) counts, what kind of processing has possibly been applied to obtain these numerical values? There are 136 dimensions to represent tetramer-frequencies, why not 256 to cover all possible 4-length words? With this data, an efficient implementation of the distance calculation using numpy is highly recommended! How can you use numpy matrix multiplication (@ operator) to realize this?

Task 2: Bayes Information Criterion

Now calculate the Bayes Information Criterion:

$$BIC = 2\, logL(\theta) - M\, logN$$

where M is the number of free parameters and $L(\theta)$ is the likelihood of the estimated model. Compare and visualize the BIC score in a function plot on the synthetic data and the gene expression data (`GexprData.csv`). Vary the number of prototypes again according to $K = 2, \ldots, 10$. Visualize the BIC score also for the "Word Count" data with prototype numbers $K = 2, \ldots, 15$. Besides the BIC plot, for $K = 2, \ldots, 15$ also visualize the corresponding elbow criterion for the K-means results from the initialization (analogous to Sheet 3, Task 2).
Binning of metagenomic contigs is an important application for clustering but the BIC may not necessarily indicate the actual number of genomes in the data. What could be used as additional information to identify a meaningful number of clusters/bins?

[1] Lin and Liao: "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes". Nature Scientific Reports, 2016.

Submission instructions

Please submit your solution by **18th January** using the corresponding homework folder in StudIP. Upload a **zip** compressed file containing all documents and data that you use in your solution. The file name should indicate the author, for example *Peter_Meinicke_Sheet4.zip*.
With your solution you should provide the Python source code together with the documentation within a **Jupyter Notebook** answering all the questions above. Use Markdown cells for your answers and explanations well separated from the Python code.
For the programming in Python/numpy do not use any code or functions from special bioinformatics libraries or toolboxes! Make sure that your code is readable and understandable for others, i.e. use descriptive names for the variable definitions according to the conventions on **Sheet 0**, do not use implicit variables (hard-coded values without a variable) and make use of comments!