Clustering Data with Gaussian Mixture Model

Homework 6 for Deep Learning, Spring 2020

Deadline: 2020.05.24 12:00:00

1 Introduction

The UCI Iris dataset contains features of 150 iris plants from 3 species. In this homework, you are to apply Gaussian mixture models to cluster the data points.

2 Clustering Data with Gaussian Mixture Model

The **homework6** folder contains the dataset file and preprocessing code (**homework6.ipynb**).

You need to implement the EM algorithm for Gaussian mixture models as described in class, and visualize the results using scatter plots. You should experiment with varying number of clusters (2 ~10) and number of iterations. When the number of clusters match the number of species, you should also check if the algorithm assigns plants from the same species to the same cluster.

Note. when the number of clusters is large, occasionally you may have singular covariance estimates. In this case you can add a small diagonal matrix to the covariance estimate, so that the E-step can proceed.

3 Attention

- You need to submit your code and a short report (in **in PDF format**).
- You are required to implement the algorithm from scratch; machine learning toolkits such as sklearn is not allowed.
- **Plagiarism is not permitted**.