Homework 6

Clustering data with Gaussian mixture model

2022403086 Bendik H. Haugen

05 24, 2022

80240743 Deep Learning

Introduction

For this homework, I have been tasked with clustering 3 species of plants. The platns are represented by the iris data set, containing 150 samples.

Theory

A mixture model is used to find subpopulations within you dataset. In this case, the different components are Iris-setosa, Iris-versicolor and iris-virginica. The gaussian mixture model (GMM), is a probibalistic model thats made for representing normally distributed subpopulations within a given dataset. The model does not need to know which subpopulations the datapoints belongs to, but it will find clusters within the set. During training, og fitting, of the model, we aim to learn the weights, means and varianses. With μ being the mean, and σ being the variance, we get the following formulas for a miulti-dimentional model: $p(x) = \sum_{i=1}^{K} \omega_i N(x|\mu, \Sigma_i)$
$N(x|\mu, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K|\Sigma|}} exp(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)$
$\sum_{i=1}^{K} \omega_i = 1$

The model uses expectation maximization. This means the model uses two steps, known as e-step and m-step. The e-step calculates the expectation of the component assignments for each data point. The next step, the m-step consists of maximizing the expectations calculated in the e-step to the model parameters. In the end, we use Bayes theorem and the learned parameters of the model to find the clusters. At the end, we hope each data-point belongs to a cluster that represent its original class.

Results

After some trial and error, I found that using 1000 iterations and 5-6 classes gives give results. The idea being that some of the classes could be used to represent edge cases. There is not a lot of improvement from 3 classes, but some.

As comes clear from the results, the model somewhat struggles to differentiate between the yellow and green data-point. I would still argue it does a good job.

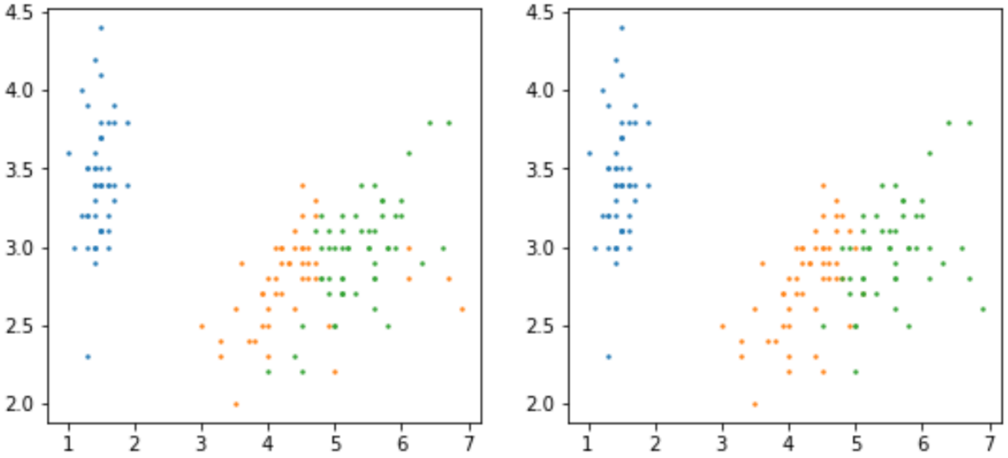

Figure 1: The result of running the algorithm, and the ground truth

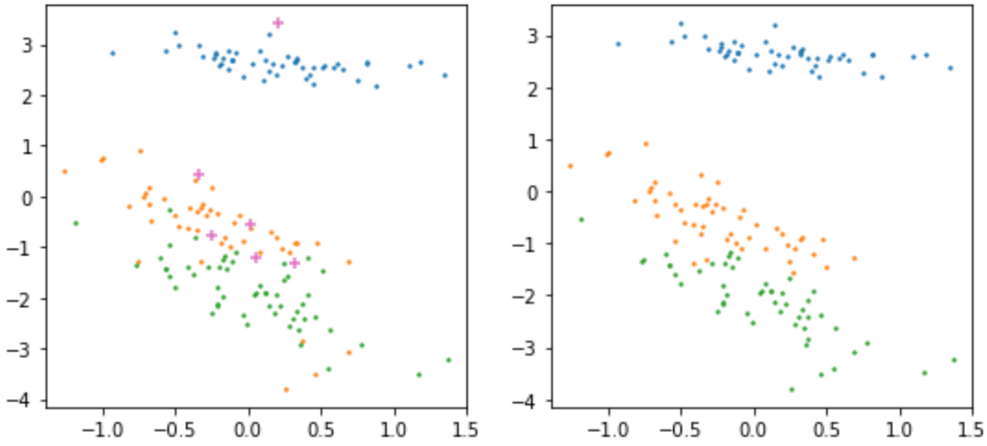

Figure 2: PCA of the model