Übungsblatt 6

Allgemeine Hinweise

- Präsentation: 03.05.2022
- Projektbericht: 17.05.2022
- Stimmabgabe: 26.04.2022, 20:00 Uhr

Exercise 1: Expectation-Maximization (Talk | Report)

The goal of this assignment is to perform model selection on mixture models using the Akaike Information Criterion.

For this, download the file `em.zip`, comprising the data files `iris.csv` and `bdp.csv` together with a notebook `em.ipynb` for loading both data sets and for generating a third synthetic data set. Implement the EM-algorithm for estimating the parameters of a mixture of k normal distributions $\mathcal{N}(x; \mu_j, \sigma_j)$ with means $\mu_j \in \mathbb{R}$ and standard deviations $\sigma_j \in \mathbb{R}$.

For each of the three data sets, conduct the following experiment: for each $k \in \{1, \ldots, 15\}$ fit a mixture model with k components to the data. Select the best model according to the Akaike Information Criterion

$$2|\hat{\theta}| - 2\mathfrak{L}_{\mathcal{D}}(\hat{\theta}),$$

where $\mathfrak{L}_{\mathcal{D}}(\hat{\theta})$ is the log-likelihood of estimate $\hat{\theta}$ and $|\hat{\theta}|$ is the number of *free* parameters to estimate (see remark below).

Model selection might not work as desired. Think about strategies to improve the results and test them on the data sets. The strategies are the contributions of your talk and report. Thus, the research question to be answered is how to improve the results of the vanilla EM algorithm for fitting mixture models to the data. Note that improving the results of an algorithm does not necessarily mean that you need to change the inner working of the algorithm.

Remark to AIC: A mixture of two normal distributions has 6 parameters

$$\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1 \sigma_2)$$

but only 5 free parameters. The reason is that we can infer π_2 from π_1 via $\pi_2 = 1 - \pi_1$. Thus, we only need to estimate $k-1$ instead of k mixing coefficients. The AIC uses this information when counting the number $|\hat{\theta}|$ of parameters to be estimated. Note that the definition of the Akaike Information Criterion slightly differs from the one presented in the lecture.

Exercise 2: Cross-Validation vs. Aikaike Information Criterion (Talk | Report)

The goal of this assignment is to compare cross validation and the Akaike Information Criterion for model selection.

For this, use the Gaussian mixture and cross validation implementation of `sklearn`. Note that the Gaussian Mixture implementation of `sklearn` provides a method for computing the Akaike Information Criterion. Conduct three experiments:

1. Use synthetic d-dimensional data generated by a Gaussian mixture model f with k components. Choose $d \geq 5$ and $k = 3$. Perform model selection over different types of covariance matrices. The type of covariance matrix can be specified by the parameter `covariance_type` of sklearn's Gaussian mixture implementation. Suppose that f is the true model that generated our data set and $\hat{f}$ is an estimated model selected by cross-validation or AIC. We can compute the average absolute error

   $$E(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left|f(x_i) - \hat{f}(x_i)\right|.$$

   Use the mean absolute error to compare the models selected by cross-validation and AIC as a function of the data set size n.

2. Use synthetic d-dimensional data generated by a Gaussian mixture model with k components. Choose $d = 2$, $k = 3$, and `covariance_type ='full'`. Perform model selection over different numbers of components. Compare the mean absolute errors of the estimated models selected by cross-validation and AIC as a function of the data set size n.

3. Select a real-world data set with features that are real-valued. Fit an appropriate Gaussian mixture to the data using model selection. Report the number of components. For each component, count the number of points for this component has maximum responsibility value.