Exercises Week 4

Machine Learning/Advanced Machine Learning
IT University of Copenhagen

Fall 2019

Theoretical Exercises 4.1: from the Book

As stated on learnit, solve the following exercises from the book:

- (7.11.11)
- (8.11.10.)

Theoretical Exercises 4.2. EM

Assume data points $\boldsymbol{x}_n \in \mathbb{R}^D$ are given and you know the data is separated into K classes. The goal is to (1) estimate the classes, each represented by one cluster center $\boldsymbol{m}_k \in \mathbb{R}^D$, and (2) to estimate to which of these cluster centers each data point $\boldsymbol{x}_n$ belongs to, hence minimize:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2, \tag{Ex. 4.2.1}$$

where r_{nk} is an indicator variable, which is one if $\boldsymbol{x}_n$ belongs to $\boldsymbol{m}_k$, else zero. In the EM-algorithm the estimation of the unknowns $\boldsymbol{m}_k$ and r_{nk} is alternated, by fixing one of them while estimating the other.

(a) Assume $\boldsymbol{m}_k$ are known, estimate the binary values r_{nk}. Keep in mind that for each n one k must be assigned.
Hint: Consider that this can be done independently for each single data point $\boldsymbol{x}_n$.

$$J_n(r_{nk}) = \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2 \tag{Ex. 4.2.2}$$

(b) Assume r_{nk} are known, estimate $\boldsymbol{m}_k$. Compute the derivative of J with respect to $\boldsymbol{m}_k$, set it to zero and then deduce the estimates for $\boldsymbol{m}_k$. By rewriting

$$J = \sum_{k=1}^{K} J_k(\boldsymbol{m}_k), \tag{Ex. 4.2.3}$$

we can focus on one cluster center k at a time by:

$$J_k(\boldsymbol{m}_k) = \sum_{n=1}^{N} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2 \tag{Ex. 4.2.4}$$

Hint: $\|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2 = (\boldsymbol{x}_n - \boldsymbol{m}_k)^{\mathrm{T}}(\boldsymbol{x}_n - \boldsymbol{m}_k)$

Programming Exercise 4.3: K-means

In this task the K-means algorithm shall be implemented and applied.

(a) Implement the K-means algorithm by yourself, see Fig. 7.3 in the book.

(b) The file `iris.txt` contains the Iris data set[1]. It consists of a total of 150 samples, with 50 for each of the three classes, while each sample is composed of 4 values, hence the data matrix $\boldsymbol{X} \in \mathbb{R}^{150\times 4}$. Choose $K = 3$ to estimate the three classes by your implementation of the K-means clustering, i.e. for each data point $\boldsymbol{x}_i \in \mathbb{R}^4$ assign one of three classes.
Hint: Consider that the absolute values of the true and estimated labels can differ. In the following example the three classes are 100% correctly estimated:
provided true labels: 1, 1, 2, 3, 3
estimated labels: 0, 0, 3, 2, 2

(c) Perform the clustering based on two of the four variables (columns), such that $\boldsymbol{x}_i \in \mathbb{R}^2$. Test all 6 variants: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4).
Remember that the initialization is based on random samples and hence the outcome will differ if you repeat the experiments. For a better comparison choose the same K random starting samples for each version.

(d) Using the ground truth labels provided in `iris_labels.txt` compute the percent of correctly classified samples, and fill the entries of the confusion matrix for results of (b) and (c).

		estimated class		
		class 1	class 2	class 3
true class	class 1			
	class 2			
	class 3			

- Which classes are most often confused? (Please refer to the original labels.)
- Which of the variants in (c) is same as good as (b)?

(e) (**NOT mandatory**) Apply your implementation to perform a color segmentation with different K of an image. You can use the gray and/or color image provided on learnit.

[1] `https://en.wikipedia.org/wiki/Iris_flower_data_set`