INF367: Spring 2020
Exercise 8

Instructions:

You can either return the solutions electronically via MittUiB by Monday 10.00 or show them on paper on Monday's meeting. Grades are awarded for effort so scanned notes are fine if you solve exercises by hand (no need to make fancy latex files).

Students are encouraged to write computer programs to derive solutions whenever appropriate.

Tasks

1. **EM algorithm for GMMs**

In this exercise, we try to find clusters of large cities. Download the data set `largest_cities.csv` which contains the coordinates (longitude and latitude) of the 500 largest cities in the world.

1. Implement the EM algorithm for GMMs using the formulas in the lecture slides
2. Compute log-likelihood

$$\log P(\mathbf{x}\,|\,\theta) = \sum_{i=1}^{n} \log \Big[\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i\,|\,\mu_k, \Sigma_k) \Big]$$

   after each M-step. Sanity check: log-likelihood should increase after each step
3. Use BIC to choose the number of components K.

   Note that representing one d-dimensional component requires $d+(d+1)\times d/2$ parameters (d parameters for the mean vector and $(d+1)\times d/2$ parameters for the covariance matrix). In addition, we need $K-1$ parameters for mixing coefficients ("-1" is there because mixing coefficients sum to one). Thus, the total number of parameters is $K \times (d + (d+1)d/2) + (K-1)$.

Hints: The EM algorithm is sensitive to initialization. Therefore, you may want to try several different initializations.

You may encounter `ValueError: array must not contain infs or NaNs` because some responsibilities are `nan`. This is typically due to some data points

lying so far away from all the cluster centers that Python interprets all probabilities as 0. In this case, you should try to initialize covariance matrices with larger variances.

2. **Extension of the simple example**

Suppose that we have n independent observations $\mathbf{x} = (x_1, \ldots, x_n)$ from a two-component mixture of univariate Gaussian distribution with unknown mixing coefficients and unknown mean of the second component:

$$P(x_i \,|\, \mu, p) = (1 - p) \cdot N(x_i \,|\, 0, 1) + p \cdot N(x_i \,|\, \mu, 1).$$

(a) Write down the complete data log-likelihood and derive the EM-algorithm for learning maximum likelihood estimates for μ and p.

(b) Implement the EM-algorithm. Load data `ex8_2.csv` and learn the maximum likelihood estimates for $\hat{\mu}$ and $\hat{p}$. Plot the distribution $P(x_i \,|\, \hat{\mu}, \hat{p})$. Does it match the observed data? Plot the log-likelihood of the observed data for each iteration (This should increase after each iteration)

Hint: You can use `simple_example.pdf` as a starting point.