Extra exercises for STK4030

Geir Storvik

Fall 2013

Exercise 1 (Classification and unlabeled data)
We will in this exercise see how estimation can be based on training data with known classes combined with data where the class labels are unknown.

Assume

$$\mathcal{T} = \{\boldsymbol{x}_{k,j}, k = 1, ..., K, j = 1, ..., n_k, \boldsymbol{x}_i, i = 1, ..., N\}$$

where $\boldsymbol{x}_{k,j}$ is an observation from class k while $\boldsymbol{x}_i$ is an observation with unknown class.

We will assume that $\pi_1, ..., \pi_K$ are all equal to $1/K$, while $f_k(\boldsymbol{x}; \boldsymbol{\theta})$ is the probability density for data from class k. We want to estimate $\boldsymbol{\theta}$ based on data $\mathcal{T}$.

(a) Show that

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{1}{K} f_k(\boldsymbol{x}_i; \boldsymbol{\theta})$$

and use this to derive an expression for the likelihood function $L(\boldsymbol{\theta}; \mathcal{T}) = p(\mathcal{T}|\boldsymbol{\theta})$ as well as the log-likelihood function $l(\boldsymbol{\theta}; \mathcal{T}) = \log[l(\boldsymbol{\theta}; \mathcal{T})]$ for $\mathcal{T}$.

Discuss estimation of $\boldsymbol{\theta}$ only based on $\{\boldsymbol{x}_{k,j}, k = 1, ..., K, j = 1, ..., n_k\}$ related to estimation of $\boldsymbol{\theta}$ based on the whole $\mathcal{T}$.

In order to estimate $\boldsymbol{\theta}$, we will use an iterative procedure called the *EM algorithm* (EM is a abbreviation for Expectation-Maximization). This is a algorithm that can be used when parts of the data is unobserved (or missing or incomplete). In our case we can think of the classes corresponding to $\boldsymbol{x}_i, i = 1, ..., N$ as missing. Let $\boldsymbol{C}_m = (C_1^m, ..., C_N^m)$ be the missing classes.

Denote by $p(\mathcal{T}, \boldsymbol{C}_m|\boldsymbol{\theta})$ the probability density for the "complete" data, that is when *all* classes are known. Treated as a function of $\boldsymbol{\theta}$ this would be the function to maximize if $\boldsymbol{C}_m$ was known. The first step of the EM-algorithm is to **E***stimate* $\log p(\mathcal{T}, \boldsymbol{C}_m|\boldsymbol{\theta})$ by its expectation:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathsf{E}^{\boldsymbol{C}_m}[\log p(\mathcal{T}, \boldsymbol{C}_m|\boldsymbol{\theta})|\mathcal{T}; \boldsymbol{\theta}'].$$

Here the expectation is under the model and our current best guess of $\boldsymbol{\theta}$ being equal to $\boldsymbol{\theta}'$. The next step is the **M***aximization* where a new value for $\boldsymbol{\theta}$ is obtained by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ wrt $\boldsymbol{\theta}$.

Specifically, the algorithms runs as follows:

Start with a sensible initial value of $\boldsymbol{\theta}^0$ for $\boldsymbol{\theta}$ (e.g. estimates based on only using the data with known classes). For $s = 1, 2, ...$, perform the following steps:

E-step : Find

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{s-1}) = \mathsf{E}^{\boldsymbol{C}_m}[\log p(\mathcal{T}, \boldsymbol{C}_m|\boldsymbol{\theta})|\mathcal{T}; \boldsymbol{\theta}^{s-1}]$$

M-step : Find $\boldsymbol{\theta}^s$ as the value which maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{s-1})$ wrt $\boldsymbol{\theta}$.

It can be shown that (see e.g. sec 8.5 in the text book) that at each iteration there will be an increase (or at least not a decrease) in $L(\boldsymbol{\theta}; \mathcal{T}) =$, that is

$$L(\boldsymbol{\theta}^s; \mathcal{T}) \geq L(\boldsymbol{\theta}^{s-1}; \mathcal{T})$$

If the (log-) likelihood function also is bounded (which it often is), the algorithm will converge to a (local) maximum.

(b) Show that for our situation

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{s-1}) = &\sum_{k=1}^{K}\sum_{j=1}^{n_k}[\log p_k(\boldsymbol{x}_{k,j}; \boldsymbol{\theta}) + \log \pi_k] + \\ &\sum_{i=1}^{N}\sum_{k=1}^{K} p(C_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}^{s-1})[\log p_k(\boldsymbol{x}_i; \boldsymbol{\theta}) + \log \pi_k] \end{aligned}$$

where $p(C_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}^{s-1})$ is the conditional probability for $C_i = k$ given that the true parameters are $\boldsymbol{\theta}^{s-1}$.

Hint: Write the log-likelihood for (x_i, C_i) as

$$\sum_{k=1}^{K} I(C_i = k)[\log p_k(x_i) + \log \pi_k]$$

where $I(C_i = k) = 1$ is 1 if $C_i = k$ and 0 otherwise.

(c) Show that maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{s-1})$ wrt $\pi_k, k = 1, ..., K$ gives

$$\hat{\pi}_k^s = \frac{n_k + \sum_{i=1}^{N} p(C_i = k|\boldsymbol{x}_i; \boldsymbol{\theta}^{s-1})}{n + N}$$

where $n = \sum_k n_k$. Comment on this result!

Hint: Remember that $\sum_k \pi_k = 1$. Use Lagrange's method in order to take this constraint into account.

In the rest of the exercise we will assume $p_k(\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. In this case $\boldsymbol{\theta} = \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, ..., K\}$.

(d) Show that

$$\hat{\boldsymbol{\mu}}_k^s = \frac{\sum_{j=1}^{n_k} \boldsymbol{x}_{k,j} + \sum_{i=1} p(C_i = k|\boldsymbol{x}_i;\boldsymbol{\theta}^{s-1})\boldsymbol{x}_i}{n_k + \sum_{i=1} p(C_i = k|\boldsymbol{x}_i;\boldsymbol{\theta}^{s-1})}.$$

by differentiating $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{s-1})$ wrt $\boldsymbol{\mu}_k$ and but to zero.

One can also show that (you do not need to do this)

$$\widehat{\boldsymbol{\Sigma}}_k^s = \frac{\sum_{j=1}^{n_k}(x_{k,j} - \hat{\boldsymbol{\mu}}_k)(x_{k,j} - \hat{\boldsymbol{\mu}}_k)^T + \sum_{i=1} p(C_i = k|\boldsymbol{x}_i;\boldsymbol{\theta}^{s-1})(x_i - \hat{\boldsymbol{\mu}}_k)(x_i - \hat{\boldsymbol{\mu}}_k)^T}{n_k + \sum_{i=1} p(C_i = k|\boldsymbol{x}_i;\boldsymbol{\theta}^{s-1})}.$$

Comment also on these results!

(e) On the course web page, there is an R-script called `EM_mix.R` which for $K = 2$ first simulate data with $n_k = 10$ and $N = 1000$ from a known set of parameters, followed by estimation only using data with known classes and thereafter the full data set. Try out this script many times and compare the performances of the different estimators. Comment on the results!

Exercise 2 (Analysis of the iris data)
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

The data is available in R names `iris`. We will in this exercise consider different methods for constructing a classification method for these data.

- Specify a method for comparing different classification methods (e.g splitting into training/test sets, cross-validation ...)
- Try out different classification methods you have learned on the iris data. Which method performs best?
- For the best method, also make a socalled *confussion matrix*, that is a matrix with rows the true classes and columns the predicted classes.