UNIVERSITY OF WARSAW

Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

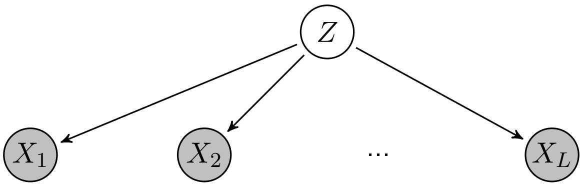
Statistical Data Analysis 2

Exercises # 2

Ewa Szczurek
Lukasz Kozlowski

September 19, 2018

Problem 1: The unsupervised naive Bayes model.
Let us assume we observe realizations of a random variable $X = (X_1, \ldots, X_L)$ with state space $\{0, 1\}$ and we want to cluster them into K distinct groups. For this we introduce a hidden random variable $Z \in \{1, \ldots, K\}$ indicating class membership. This is known as the unsupervised naive Bayes model. We can now get the class membership by computing $P(Z = k \mid X_1, \ldots, X_L)$ for any realization.



(a) How does the joint probability $P(X_1, \ldots, X_L, Z)$ factorize for this model?

(b) The model parameters are the class prior $P(Z)$, and the conditional probabilities $\theta_{n,kx} = P(X_n = x \mid Z = k)$ for each X_n. The prior we assume to be constant, which means you can use

$$P(X_1, \ldots, X_L, Z) = P(Z)P(X_1, \ldots, X_L \mid Z) \propto P(X_1, \ldots, X_L \mid Z).$$

We now want to learn θ from observed data. Show that the hidden or (complete-data) likelihood of the observed data $(X^{(1)}, \ldots, X^{(N)})$ and hidden data $(Z^{(1)}, \ldots, Z^{(N)})$ is

$$\prod_{i=1}^{N} \prod_{n=1}^{L} \prod_{k \in [K]} \prod_{x \in \{0,1\}} \theta_{n,kx}^{I_{n,kx}(Z^{(i)})},$$

where $I_{n,kx}(Z^{(i)})$is equal to 1 if and only if $X_n^{(i)} = x$ and $Z^{(i)} = k$ and 0 otherwise.

(c) Since we did not observe Z we also do not have the counts $I_{n,kx}(Z^{(i)})$. We will apply the EM algorithm to estimate the parameters without observing Z. For this, first write down the hidden log-likelihood. Then compute the expected value of $Z^{(i)}$ (E step). This is also called the responsibility of class k for observation $X_n^{(i)} = x$, written as $\gamma_{n,kx}^{(i)} = \mathrm{E}_{Z|X=x,\theta'}\left[I_{n,kx}(Z^{(i)})\right]$, where θ' is the current estimate of θ.

(d) Write down the expected hidden log-likelihood $\mathrm{E}_{Z|X,\theta'}[\ell_{\mathrm{hid}}(\theta)]$ in terms of the expected counts $N_{n,kx} = \sum_{i=1}^{N} \gamma_{n,kx}^{(i)}$. Now maximize this function (M-step) w.r.t the $\theta_{n,kx}$.

Problem 2: The EM algorithm and Gaussian mixtures

Let $n, K \in \mathbb{N}$ and consider an n-dimensional random vector X distributed according to a Gaussian mixture distribution with K components. In other words, the joint probability density function of X can be written as

$$f_X(x) = \sum_{k=1}^{K} \lambda_k \mathrm{Norm}(x \mid \mu_k, \Sigma_k)$$

where $\lambda_k \in [0,1]$ with $\sum_{k=1}^{K} \lambda_k = 1$ and $\mathrm{Norm}(x \mid \mu_k, \Sigma_k)$ stands for a multivariate Gaussian probability density function with mean vector μ_k and covariance matrix Σ_k, which is assumed to be *symmetric and nonsingular*. The goal of this exercise is to derive maximum likelihood estimators of all the parameters $\lambda_k, \mu_k, \Sigma_k$, $k = 1, ..., K$ (which we represent by $\theta := (\lambda_k, \mu_k, \Sigma_k)_{k=1,...,K}$), given a set of data points $x_1, ..., x_N$ which are *independent* samples of the random vector X. For reasons that will become apparent while you try to solve this exercise, this cannot be accomplished in closed form. However, approximate values of the maximum likelihood estimators can be calculated using the EM algorithm.

(a) Show that the *observed log-likelihood* function $\ell_{obs}(\theta) := \log(\prod_{i=1}^{N} f_X(x_i \mid \theta))$ of the sample $x_1, ..., x_N$ can be written as

$$\ell_{obs}(\theta) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \lambda_k \mathrm{Norm}(x_i \mid \mu_k, \Sigma_k) \right]$$

Furthermore, show[1] that

$$\frac{\partial \ell_{obs}}{\partial \mu_{kj}}(\theta) = \sum_{i=1}^{N} \frac{\lambda_k \mathrm{Norm}(x_i \mid \mu_k, \Sigma_k)(\Sigma_k^{-1}(x_i - \mu_k))_j}{\sum_{k=1}^{K} \lambda_k \mathrm{Norm}(x_i \mid \mu_k, \Sigma_k)}$$

and explain why it is in general very difficult to obtain a closed form expression for the maximum likelihood estimators of the parameters θ.

(b) Let us introduce a *latent* K-dimensional binary random vector Z satisfying $\sum_{k=1}^{K} z_k = 1$ and $z_k \in \{0, 1\}$. Thus, there are exactly K possible states for the vector Z according to which entry is nonzero. The probability assigned to each of these states is specified in terms of the mixing coefficients λ_k as

$$P(Z = z) = \prod_{k=1}^{K} \lambda_k^{z_k}$$

The value assumed by the latent variable specifies which of the K Gaussian distributions is used to generate a data point x. Therefore, $X \mid Z$ has the following probability density function

$$f_{X|Z}(x \mid Z = z) = \prod_{k=1}^{K} \mathrm{Norm}(x \mid \mu_k, \Sigma_k)^{z_k}$$

Now assume that every data point x_i is preceded by a realization z_i of the random variable Z_i. Show that the *log-likelihood function* $\ell_{hid}(\theta)$ *of the hidden data* $z_1, x_1, ..., z_N, x_N$ can be written as

$$\ell_{hid}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} (\log(\lambda_k) + \log(\mathrm{Norm}(x_i \mid \mu_k, \Sigma_k))) \qquad (1)$$

[1] Recall from vector calculus $(x^t A x)' = 2Ax$ for a symmetric matrix A.

Show that the expected value of the hidden log-likelihood (1) under the posterior probability $P(Z_1, ..., Z_N \mid X_1 = x_1, ..., X_N = x_N, \theta)$ satisfies

$$E_Z(\ell_{hid}(\theta)) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} (\log(\lambda_k) + \log(\text{Norm}(x_i \mid \mu_k, \Sigma_k)) \tag{2}$$

with

$$\gamma_{ik} = \frac{\lambda_k \text{Norm}(x_i \mid \mu_k, \Sigma_k)}{\sum\limits_{j=1}^{K} \lambda_j \text{Norm}(x_i \mid \mu_j, \Sigma_j)}$$

(γ_{ik} is the so called *responsibility* of component k for observation i.)

Hint: For the expectation of the hidden log-likelihood (1) under the posterior probability $P(Z_1, ..., Z_N \mid X_1 = x_1, ..., X_N = x_N, \theta)$ it holds

$$\begin{aligned} E_Z(\ell_{hid}(\theta)) &= E_Z(\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} (\log(\lambda_k) + \log(\text{Norm}(x_i \mid \mu_k, \Sigma_k)))) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} E_Z(z_{ik}) (\log(\lambda_k) + \log(\text{Norm}(x_i \mid \mu_k, \Sigma_k))) \end{aligned}$$

It can be shown that the probability density function of the random variables $Z_1, ..., Z_N \mid X_1 = x_1, ..., X_N = x_N, \theta$ can be factorized as

$$f_{Z_1,...,Z_N \mid X_1,...,X_N,\theta}(z_1, ..., z_N \mid X_1 = x_1, ..., X_N = x_N, \theta) = \prod_{i=1}^{N} f_{Z_i \mid X_i,\theta}(z_i \mid X_i = x_i, \theta)$$

Under this assumption, show

$$E_Z(z_{ik}) = \frac{\lambda_k \text{Norm}(x_i \mid \mu_k, \Sigma_k)}{\sum\limits_{j=1}^{K} \lambda_j \text{Norm}(x_i \mid \mu_j, \Sigma_j)}$$

With this result, the EM algorithm can be applied. After choosing some initial values θ^{old} for the parameters, the E step consists of calculating the responsibilities γ_{ik}. Keeping the values γ_{ik} fixed, the expectation (2) is maximized with respect to θ. This is the M step.

(c) Show that maximizing (2) yields the following expression for μ_k, $k = 1, ..., K$:

$$\mu_k = \frac{\sum\limits_{i=1}^{N} \gamma_{ik} x_i}{\sum\limits_{i=1}^{N} \gamma_{ik}}$$

(d) Show that maximizing (2) yields the following expression for Σ_k, $k = 1, ..., K$:

$$\Sigma_k = \frac{\sum\limits_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^t}{\sum\limits_{i=1}^{N} \gamma_{ik}}$$

To this end, show that maximizing the expression

$$\sum_{i=1}^{N}\sum_{k=1}^{K} E_Z(z_{ik})(\log(\lambda_k) + \log(\mathrm{Norm}(x_i \mid \mu_k, \Sigma_k)))$$

is equivalent to maximizing

$$\Gamma\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2\Gamma}\sum_{i=1}^{N}\gamma_{ij}(x_i-\mu_j)^t\Sigma_j^{-1}(x_i-\mu_j)\right)$$

where $\Gamma := \sum\limits_{i=1}^{N}\gamma_{ij}$. Furthermore, show that maximizing the last expression is equivalent to maximizing

$$-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2\Gamma}\sum_{i=1}^{N}(\sqrt{\gamma_{ij}}x_i - \sqrt{\gamma_{ij}}\mu_j)^t\Sigma_j^{-1}(\sqrt{\gamma_{ij}}x_i - \sqrt{\gamma_{ij}}\mu_j)$$

The latter maximization is (up to a coordinate transform) equivalent to maximizing the log-likelihood function of a sample of length N out of a multivariate Gaussian distribution $\mathrm{Norm}(x \mid \mu_j, \Sigma_j)$. The solution is the well-known result

$$\Sigma_k = \frac{1}{\Gamma}\sum_{i=1}^{N}\gamma_{ij}(x_i-\mu_j)(x_i-\mu_j)^t$$

which can be found in the literature. Its derivation is rather involved, meaning that you can use it without providing a proof.

(e) Show that maximizing (2) yields the following expression for λ_k, $k = 1, ..., K$:

$$\lambda_k = \frac{\sum\limits_{i=1}^{N}\gamma_{ik}}{N}$$

Hint: Since the parameters λ_k, $k = 1, ..., K$ obey the constraint $\sum_{k=1}^{K}\lambda_k = 1$, the maximization must be performed among the values that satisfy $\sum_{k=1}^{K}\lambda_k = 1$. Recall from calculus that this can be accomplished using a Lagrange multiplier α, i.e. maximizing the expression

$$\Lambda := E_Z(\ell_{hid}(\theta)) - \alpha(\sum_{k=1}^{K}\lambda_k - 1)$$

From the equations $\partial\Lambda/\partial\lambda_k = 0$, $k = 1, ..., K$ and $\partial\Lambda/\partial\alpha = 0$, eliminate α and solve for λ_k, $k = 1, ..., K$.

The values obtained constitute the θ^{new} which can be used to calculate new values of the γ_{ik} in a subsequent E step of the EM algorithm. Another M step follows. Both steps are iterated until the evaluation of (2) does not show significant increase. The sequence of parameter values θ_l obtained converges to a local maximum of the observed log-likelihood function $\ell_{obs}(\theta)$. Thus, we have overcome the hard problem of maximizing the *observed-likelihood* $\ell_{obs}(\theta)$ in task a) by estimating it from the *hidden-likelihood* $\ell_{hid}(\theta)$ for which we were able to derive closed form expressions for the maximum likelihood estimators of the parameters θ. The price we pay for this, is that we only get an approximate local solution for $\ell_{obs}(\theta)$.