EE-613: Probabilistic Generative Models, lab 2

Fall 2017

1 Maximum-A-Posteriori adaptation of a multivariate Gaussian

The goal is to study the estimation of the parameters ($\boldsymbol{\mu}$,Σ) using the Maximum A Posteriori (MAP) principle. Comparison with the Maximum-Likelihood estimator and the effect of the prior parameters will be studied.

Let's assume that we are give a set of i.i.d. observations $\mathcal{X} = \{\mathbf{x}_i, i = 1 \dots N\}$ that follow a multivariate Gaussian law $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, where the data points are of dimension d. It has been shown that the conjugate prior for the parameter $(\boldsymbol{\mu}, \Sigma)$ of this multivariate Gaussian likelihood distribution is the Normal-Inverse Wishart distribution:

$$p(\boldsymbol{\mu}, \Sigma|\mathbf{m}, \tau, V, \nu) = \mathcal{NIW}(\boldsymbol{\mu}, \Sigma|\boldsymbol{\beta}) = \mathcal{NIW}(\boldsymbol{\mu}, \Sigma|\mathbf{m}, \tau, V, \nu) = p(\boldsymbol{\mu}|\Sigma, \mathbf{m}, \tau)p(\Sigma|V, \nu) \quad (1)$$

where $\boldsymbol{\beta} = (\mathbf{m}, \tau, V, \nu)$ denotes the set of parameters defining the prior distribution, and the prior on the mean given the covariance is given by a normal distribution:

$$p(\boldsymbol{\mu}|\Sigma, \mathbf{m}, \tau) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, \frac{1}{\tau}\Sigma), \quad (2)$$

and the conjugate prior of the covariance is an inverse Wishart distribution given by:

$$p(\Sigma|V, \nu) = \mathcal{IW}(\Sigma|V, \nu) = B|\Sigma|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}tr(V\Sigma^{-1})\right) \quad (3)$$

where B is some normalization constant. According to the conjugacy definition, the posterior for the Gaussian parameters also follows a Normal-Inverse Wishart distribution given by:

$$p(\boldsymbol{\mu}, \Sigma|\mathcal{X}, \boldsymbol{\beta}) = \mathcal{NIW}(\boldsymbol{\mu}, \Sigma|\mathbf{m}_{new}, \tau_{new}, V_{new}, \nu_{new}) \quad (4)$$

with

$$\mathbf{m}_{new} = \frac{\tau\mathbf{m} + N\bar{\mathbf{x}}}{\tau + N}, \tau_{new} = \tau + N, \nu_{new} = \nu + N, V_{new} = V + NS + \frac{\tau N}{\tau + N}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})' \quad (5)$$

where $\bar{\mathbf{x}}$ denotes the sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i$, and $S = \frac{1}{N}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. Finally, from this distribution, the MAP estimates for the mean and covariance are given by

$$\boldsymbol{\mu}_{MAP} = \frac{\tau\mathbf{m} + N\bar{\mathbf{x}}}{\tau + N} \quad (6)$$

$$\Sigma_{MAP} = \frac{V + NS + \frac{\tau N}{\tau+N}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})'}{\nu - d + N} \quad (7)$$

(a) Draw the graphical model of the $(\boldsymbol{\mu}, \Sigma, \mathcal{X})$ variables (you can use the plate notation). Remind the conjugacy principle. Remind the MAP principle and thus to what correspond the above MAP estimates w.r.t. distributions provided above.

(b) Programming. You are given a squeleton in matlab `main.m` with missing parts. The data generation is written and some parameters settings for the wisharts are proposed. In addition, there is a function `Plot_GM.m` which can be used to display gaussian model parameters (mean and ellipse at one standard deviation to visualize the covariance). You are asked to program:

- a function which performs Maximum Likelihood estimation:

```
function [mu,Sigma]=MLGaussPara(X)
%%
% Performs Maximum Likelihood estimation of gaussian parameters
% Input: X: data points
% Output:mu,Sigma: estimated gaussian parameters
```

- a function which compute the MAP estimates of the multivariate Gaussian parameters under normal-wishart prior:

```
function [mu,Sigma]=MAPGaussPara(X,para)
%%
% Performs MAP estimation of a multivariate gaussian parameter using normal-wishart prior
%
%  See homework text
%
% Input: X: data points
%        para: structure containing the wishart parameters
%
% Output: mu,Sigma: estimated gaussian parameters
```

Assuming 100 points are generated, the following default parameters are assumed for the prior: $\mathbf{m} = (-10, -10)'$, $\tau = 100$, $\nu = 100$, and $V = 100 \times 10 I_2$ where I_2 denotes the 2×2 identity matrix.

(c) By inspecting the prior distribution on the mean and its MAP estimate, provide an interpretation of the $\mathbf{m}$ and τ parameters. Illustrate your answer by testing different values of τ, e.g $5, 100, 500$. What can you say about the special value $\tau = N$?

(d) Comment on the consequences of varying τ (e.g. towards 0, or towards values $\gg N$) on the MAP estimate of the covariance Σ_{MAP}. Why does it make sense?

(e) The MAP estimation of the covariance is also controlled through the Wishart parameters V and ν. It can be shown that the expected covariance Σ_E under the inverse Wishart law is given by $\frac{1}{\nu-d-1}V$ (write down the equation corresponding to what this sentence mean). To manually define the Wishart parameters V and ν, it might be thus more convenient to define Σ_E (as it is more easily interpretable w.r.t. tou our data) and then change ν to set V accordingly.

Try different values of ν (e.g. 25,100,1000) (use the above method for setting V), and visualize the obtained results. Given these results and looking at the formula of the MAP estimate of the covariance, explain the qualitative role of this parameter.

2 Gaussian Mixture Model (GMM) with Maximum-A-Posteriori adaptation

The goal is to study the GMM algorithm and how it can be used for model adaptation. The illustration context is on a simplified setting on how to estimate the Visual Focus of Attention (VFOA) of a person, i.e. to which target a person is looking at, given his head pose estimate.
Situation. Assume that scene consists of a room with the Nao robot, a human and 2 paintings. Nao is interacting with the human, commenting about the painting and having a questions & answers session. It is assumed that there are 3 possible VFOA targets for the human: Nao and the two paintings. The situation is illustrated in Fig. 1.

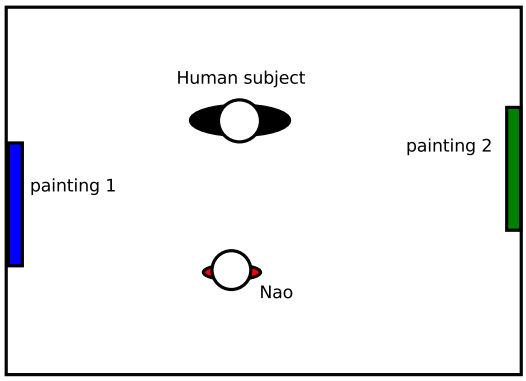

Figure 1: *Task overview.* A robot (Nao) is interacting with a human subject. From the video that looks at the subject, we can extract the person's head orientation $\mathbf{x} = (x_1, x_2)$, with x_1 representing the pan angle, i.e. looking right or left, and x_2 the tilt angle, looking up or down. From this observation, we would like to know whether the person looks at the robot, at the first painting, or at the second painting.

During the interaction with the robot, we assume the following generative process for the head poses of the person[1]

1. draw randomly a focus $\mathbf{z}$ according to the categorical distribution parameterized by $\boldsymbol{\pi}$:

$$\mathbf{z} \sim p(\mathbf{z} = k|\boldsymbol{\pi}) = \pi_k$$

   We will define that $k = 2$ corresponds to looking at painting 1, $k = 1$ corresponds to looking at painting 2, and $k = 3$ corresponds to looking at nao.

2. given the focus, draw the head pose from a Gaussian distribution:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z} = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

   We will denote $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \Sigma_k)$ and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\theta}_k, k = 1..K\}$.

[1]Note that in principle we should have considered the time variable, since the VFOA of a person normally remains the same for short segment of time.

You are given a main.m program and additionnal functions that you need to complete.

(a) You are given a training dataset (the matlab file Training.mat) containing complete observable data $(\mathbf{X}_{train}, \mathbf{Z}_{train})$ from different people. Use this dataset to learn the parameters $\boldsymbol{\theta}$ of the model, by completing the function:

```
function [pi,mu,Sigma]=GMMLabeledObservation(K,X,Z)
%%
% Performs Maximum Likelihood estimation of the GMM parameters given labeled observation (complete data)
% Input:
%   K : number of mixtures
%   X : (N x DimX) matrix of data points of DimX, with N number of points
%   Z : column vector of size N containing the component indices of X
% Output: pi,mu,Sigma: mixture weights and
%         estimated gaussian parameters for each of the component
```

You may want to use the function MLGaussPara to do the training.

(b) You are now given the sequences of head poses $\mathbf{X} = \{\mathbf{x}_i, i = 1, ..., N\}$ of a new person. Write down how you can recognize the VFOA for each of the head pose given the model.

Select the right option in the main file, and compute the obtained classification error for TestPerson1.mat and TestPerson2.mat.

(c) Unfortunately, the observations of a test person are potentially noisy, and different persons may use on average different head poses to look at the same targets (e.g. depending on whether the wear glasses for instance). In addition, during an interaction, people may look more or less at different paintings (e.g. if Nao does not speak about painting 2, people may not look at it). To handle those issues, we would thus like to use a generic model and adapt its parameters for the specific person and interaction.

Show that the distribution of head poses $p(\mathbf{x})$ of a person follows a GMM model. Then, complete the function GMM that learns the parameters of a GMM using the EM algorithm, as well as the related functions (see code).

```
function [pi,moy,Sigma]=GMMEM(K,X,para)
%%
% Performs Maximum Likelihood estimation of the GMM parameters using the EM algorithm
% Input:
%   K : number of mixtures
%   X : (N x DimX) matrix of data points of DimX, with N number of data points
%   para : parameter options
% Output: pi,moy,Sigma: mixture weights and
%         estimated gaussian parameters for each of the component
```

Set the right option in the main. Note the function

```
function [ pireorder,mureorder,sigmareorder] = Reorder(pitest,mutest,sigmatest,meansref)
```

that follows the GMM function. What does this function? why is it necessary to conduct the evaluation?
Commpute the classification accuracy for the test persons. Explain the results.

(d) In order to solve the issues raised in the previous question, we notice that using prior on each focus is a must for our problem. We therefore will perform MAP adaptation. We remind that MAP adaptation does not affect the E-step, where the distribution over the hidden variables is computed with the current estimates of the parameter values. In the M-step, the MAP maximizes the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + \log(p(\boldsymbol{\theta}|\boldsymbol{\Theta}))$, where $\boldsymbol{\Theta}$ are the parameters defining the prior distribution. For convenience, we use the appropriate conjugate distribution for each factor. For the categorical distribution over $\mathbf{z}$, we thus use a dirichlet $\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$. For each of the Gaussian, we use a normal-inverse-wishart prior, as given by Eq. 1, i.e. $p(\boldsymbol{\theta}_k|\boldsymbol{\beta}) = \mathcal{NIW}(\boldsymbol{\mu}_k, \Sigma_k|\boldsymbol{\beta}_k)$.

It can be shown that the MAP parameters of each of the conditional probability distribution at each M-step is then simply given by the MAP for each distribution separately, using as data the set of weighted samples:

$$\mathcal{X}_k = \{(\mathbf{x}_i) \text{ with weight } \gamma(z_{ik}), i = 1, ..., N\} \tag{8}$$

where the weights are estimated in the E-step. The MAP for the parameters $\boldsymbol{\pi}$ of the categorical distribution is given in the course. The MAP expression of the mean and covariance of each Gaussian is given by Eq. 6 and 7, modified to take into account the weighting factor. For instance, for the mean, we have:

$$\boldsymbol{\mu}_{MAP,k} = \frac{\tau_k \mathbf{m}_k + c_k \bar{\mathbf{x}}_k}{\tau_k + c_k} \text{ with } c_k = \sum_{i=1}^{N} \gamma(z_{ik}), \bar{\mathbf{x}}_k = \frac{1}{c_k} \sum_{i=1}^{N} \gamma(z_{ik}) \mathbf{x}_i$$

Complete the GMMadapt function to perform a MAP adaptation in the M-step rather than a ML estimation as in the standard EM.

(e) In the main.m file, select the GMM MAP option for computing the GMM parameters, and complete the part regarding the initialization of the Dirichlet parameters $\boldsymbol{\alpha}$ for the prior on the VFOA (i.e. $\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$), and the parameters $\boldsymbol{\beta}_k$ for each of the Gaussian using the parameters estimated from the training data and the analysis performed in Part 1.
Observe and comment how the different components and their weights change during the adaptation for each of the test person. Comment the classification results obtained, by contrasting them to those obtained in the supervised setting, or in the fully unsupervised case.
Based on the analysis of Section 1, try different options for setting the prior parameters. In a real setting, what would influence your choice for setting these parameters ?

(f) A person may not always look at the 3 VFOA targets. How would you change the model to account for head poses that would not correspond to any of the 3 predefined targets?