Weekly Exercise 4

Dr. Csaba Domokos and Lingni Ma
Technische Universität München, Computer Vision Group
April 26, 2016 (submission deadline: May 03, 2016)

The EM algorithm for mixtures of Gaussians (2 points)

Exercise 1 (**M step for $\mathbf{\Sigma}$**, 2 points)**.** Assuming n data samples $\{\mathbf{x}_n\}_{n=1}^N$, consider the log-likelihood function of a mixture of Gaussian model with K components

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_k(\mathbf{x}_n)\big(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\big) \ .$$

Show that the optimal choice with respect to the covariance matrices $\mathbf{\Sigma}_k$ for all $k = 1, \dots, K$ is given as

$$\operatorname{argmax}_{\mathbf{\Sigma}_k} \mathcal{L}(\boldsymbol{\theta}) = \frac{\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^{N}\gamma_k(\mathbf{x}_m)} \ .$$

(Hint: for a symmetric $\mathbf{X} \in \mathbb{R}^{n\times n}$ matrix and vector $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \mathbf{X}}\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T} \ ,$$

and for a non-singular matrix $\mathbf{Y} \in \mathbb{R}^{n\times n}$,

$$\frac{\partial}{\partial \mathbf{Y}}|\mathbf{Y}| = |\mathbf{Y}|\mathbf{Y}^{-1} \ .$$

Solution. We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\mathbf{\Sigma}_k$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Sigma}_k}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)\frac{1}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}\ \frac{\partial}{\partial \mathbf{\Sigma}_k}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\\ &= \sum_{n=1}^{N}\frac{\gamma_k(\mathbf{x})}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}\ \frac{\partial}{\partial \mathbf{\Sigma}_k}\left(\frac{1}{\sqrt{|2\pi\mathbf{\Sigma}_k|}}\exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big)\right) \ .\end{aligned}$$

Let us first calculate the following derivatives:

$$\frac{\partial}{\partial \mathbf{\Sigma}_k}\frac{1}{\sqrt{|2\pi\mathbf{\Sigma}_k|}} = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{\partial}{\partial \mathbf{\Sigma}_k}|\mathbf{\Sigma}_k|^{-\frac{1}{2}} = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{-1}{2}|\mathbf{\Sigma}_k|^{-\frac{3}{2}}|\mathbf{\Sigma}_k|\mathbf{\Sigma}_k^{-1} = \frac{-\mathbf{\Sigma}_k^{-1}}{2\sqrt{|2\pi\mathbf{\Sigma}_k|}} \ .$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \frac{-1}{2}(-\boldsymbol{\Sigma}_k^{-T})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_k^{-T} \\
&= \frac{1}{2}\exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \; .
\end{aligned}
$$

Now we are at the position to calculate the derivative of a Gaussian w.r.t. $\boldsymbol{\Sigma}$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left(\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big)\right) \\
=&\frac{-\boldsymbol{\Sigma}_k^{-1}}{2\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&+ \frac{1}{2}\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
=& -\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \frac{1}{2}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
=&\frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{2}\boldsymbol{\Sigma}_k^{-1}\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \; .
\end{aligned}
$$

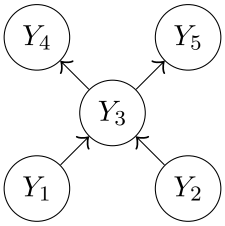
Setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$ to 0, we obtain

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x}_n)}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{2}\boldsymbol{\Sigma}_k^{-1}\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \\
&= \frac{\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \; .
\end{aligned}
$$

$$
\begin{aligned}
\frac{\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) &= 0 \\
\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} &= \sum_{m=1}^{N}\gamma_k(\mathbf{x}_m) \\
\frac{\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^{N}\gamma_k(\mathbf{x}_m)} &= \boldsymbol{\Sigma}_k \; .
\end{aligned}
$$

Graphical models (3 points)

Exercise 2 (Bayesian network, 1 point**).** Provide the factorization of $p(y_1, y_2, y_3, y_4, y_5)$ according to the following directed graphical model:



Solution. The joint distribution over all five variables is given as:

$$p(y_1, y_2, y_3, y_4, y_5) = p(y_1)p(y_2)p(y_3 \mid y_1, y_2)p(y_4 \mid y_3)p(y_5 \mid y_3) \ .$$

Exercise 3 (Proof of the Hammesley-Clifford theorem $\boldsymbol{\Sigma}$, 2 points**).** Assume an undirected graphical model $G = (\mathcal{V}, \mathcal{E})$ satisfying the local Markov property. Consider two non-connected nodes $a, b \in \mathcal{V}$, i.e. $(a, b) \notin \mathcal{E}$, and a subset of nodes $w \subset \mathcal{V}$ such that $a, b \notin w$. Show that

$$q(y_a \mid \mathbf{y}_w) \overset{\Delta}{=} p(y_a \mid \mathbf{y}_w, \mathbf{y}^*_{\mathcal{V} \setminus (w \cup \{a\})}) = q(y_a \mid y_b, \mathbf{y}_w) \ ,$$

where $p(\mathbf{y}_z, \mathbf{y}^*_{\bar{z}})$ is a joint probability.

Solution.

$$\begin{aligned} q(y_a \mid \mathbf{y}_w) \overset{\Delta}{=} & p(y_a \mid \mathbf{y}_w, \mathbf{y}^*_{\mathcal{V} \setminus (w \cup \{a\})}) \\ = & p(y_a \mid y_b^*, \mathbf{y}_w, \mathbf{y}^*_{\mathcal{V} \setminus (w \cup \{a,b\})}) \overset{a \perp\!\!\!\perp b}{=} p(y_a \mid y_b, \mathbf{y}_w, \mathbf{y}^*_{\mathcal{V} \setminus (w \cup \{a,b\})}) \\ \overset{\Delta}{=} & q(y_a \mid y_b, \mathbf{y}_w) \ . \end{aligned}$$

Programming (6 points)

Exercise 4 (Gaussian Mixture Model Estimation, 6 points**).** Train two Gaussian mixture models to model the foreground and background probabilities of the pixels based on their intensity. Apply the trained trained model to the input image and segment the foreground and background. The testing image is shown in Figure 1 (a) (you can download the images from `supp_04.zip`). The specific requirement is as follows.

1. Write a program to take in argument (x_1, y_1, x_2, y_2), which specify the bounding box of the foreground as shown in in Figure 1 (b). Model the foregreound probability $p_F(I), I \in \mathbb{R}^3$ using all pixels inside the bounding box, where $p_F(I)$ is a GMM model with 5 Gaussian components and I is the RGB values of the pixel.

2. Based on the same bounding box, model the background probability $p_B(I), I \in \mathbb{R}^3$ using all pixels outside the bounding box, where $p_B(I)$ is a GMM model with 5 Gaussian components.

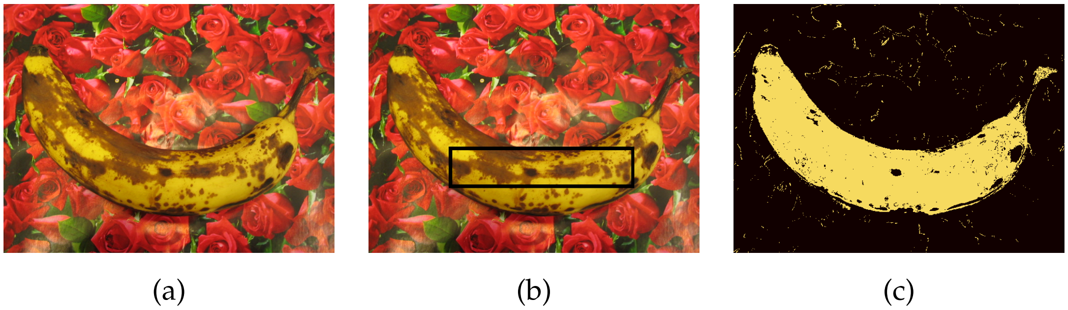


Figure 1: (a) test image. (b) bounding box for fourground training. (c) foregound segmentation based the trained model $p_F(I)$.

3. Segment the input image into foreground and background use the two distributions you obtained, respectively. Compare the results.

Hints: the bounding box should mostly contain the banana. You can initialize the GMM with random Gaussian kernels. Note that the covariance matrix should not be singular, during any time of GMM training. If the covariance matrix does become singular, you can restart the estimation from a different initialization all over again. Alternatively, can you think of way to better initialize the GMM?