TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis

Machine Learning for Computer Vision
Summer term 2017

June 17, 2017
Topic: Laplace Approximation, K-Means, EM

Exercise 1: Laplace Approximation

In Gaussian Process classification, we cannot integrate exactly over the parameters $\mathbf{w}$.

a) The integral of the predictive distribution becomes analytically intractable because the posterior distribution is no longer Gaussian. Therefore we don't have a closed form solution as in regression.
There are basically two approaches to tackle this problem. One is to approximate the true posterior with sampling methods. The other is to use analytical approximations which assume a Gaussian posterior. There are three common methods under this approach:

- Laplace approximation
- Expectation Propagation
- Variational Inference

b) The goal is to find a Gaussian distribution $q(z)$ with mean equal to a mode of $p(z)$. In other words, we first want to find a point $z_0 \in \mathbb{R}^d$ for which the gradient of $p(z)$ is zero. For now we can ignore the normalizer and work with $f(z)$.

The Laplace Approximation considers a second-order Taylor expansion of the logarithm of $f(z)$ centered at z_0:

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}(z - z_0)^T A(z - z_0) \tag{1}$$

where $A = -\nabla\nabla \ln f(z_0)$ is the Hessian matrix at z_0.

The first-order term does not appear as z_0 is a local maximum ($\nabla f(z_0) = 0$).

If we take the exponential on both sides, we approximate $f(z)$ as:

$$f(z) \approx f(z_0) \exp\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\} \tag{2}$$

With this we can easily write $q(z)$ as a normal distribution with mean z_0 and covariance A^{-1}. The only thing left is to estimate the normalizer which we can do by inspection. The Laplace approximation then is:

$$q(z) = \frac{1}{(2\pi)^{(d/2)}|A|^{1/2}} \exp\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\} \tag{3}$$

Exercise 2: Expectation-Maximization for GMM

In the standard EM algorithm, we first define the responsibilities γ as

$$\gamma_{nk} = p(z_{nk} = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \quad , z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1$$

a) Find the optimal means, covariances and mixing coefficients that maximize the data likelihood. How can we interpret the results?

We want to maximize the data likelihood, so as usual we minimize the negative log-likelihood:

$$-\mathcal{LL} = -\log p(X|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = -\log \prod_n \sum_k \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{4}$$

This time we minimize 3 times independently with respect to the means, the covariances and the mixture coefficients:

$$\mu_k^* = \arg\min_{\mu_k} -\mathcal{LL} \tag{5}$$
$$\Sigma_k^* = \arg\min_{\Sigma_k} -\mathcal{LL} \tag{6}$$
$$\pi_k^* = \arg\min_{\pi_k} -\mathcal{LL} \tag{7}$$

In the following, to avoid confusion of sums and covariances, we denote covariance Σ_k as C_k. To simplify some expressions, let us agree on the following notation:

$$\mathcal{N}_{nk} \equiv \mathcal{N}(x_n|\mu_k, C_k) \tag{8}$$
$$Z_k \equiv ((2\pi)^d |C_k|)^{1/2} \tag{9}$$
$$D_{nk} \equiv (x_n - \mu_k)^T C_k^{-1} (x_n - \mu_k) \tag{10}$$
$$\text{Therefore } \mathcal{N}_{nk} = Z_k^{-1} \exp\{-\frac{1}{2} D_{nk}\} \tag{11}$$

Thus, we have:

$$\begin{aligned} -\mathcal{LL} &= -\sum_n \log \sum_k \pi_k \mathcal{N}_{nk} \\ &= -\sum_n \log \sum_k \pi_k Z_k^{-1} \exp(-\frac{1}{2} D_{nk}) \end{aligned}$$

Solving for the means:

$$\frac{\partial \mathcal{L}\mathcal{L}}{\partial \mu_k} = \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k Z_k^{-1} \exp(-\frac{1}{2} D_{nk})}{\partial \mu_k} \tag{12}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k Z_k^{-1} \frac{\partial \exp(-\frac{1}{2} D_{nk})}{\partial \mu_k} \tag{13}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k Z_k^{-1} \exp(-\frac{1}{2} D_{nk}) C_k^{-1} (x_n - \mu_k) \tag{14}$$

$$= \sum_n \frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} C_k^{-1} (x_n - \mu_k) \tag{15}$$

$$= \sum_n \gamma_{nk} C_k^{-1} (x_n - \mu_k) \tag{16}$$

$$\tag{17}$$

Setting $-\frac{\partial \mathcal{L}\mathcal{L}}{\partial \mu_k} \stackrel{!}{=} 0$ gives us:

$$\sum_n \gamma_{nk} C_k^{-1} \mu_k = \sum_n \gamma_{nk} C_k^{-1} x_n \tag{18}$$

$$C_k^{-1} \mu_k \sum_n \gamma_{nk} = C_k^{-1} \sum_n \gamma_{nk} x_n \tag{19}$$

$$C_k^{-1} \mu_k \sum_n \gamma_{nk} = C_k^{-1} \sum_n \gamma_{nk} x_n \tag{20}$$

$$\mu_k \sum_n \gamma_{nk} = \sum_n \gamma_{nk} x_n \tag{21}$$

$$\mu_k = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}} \tag{22}$$

Solving for the covariances:

$$\frac{\partial \mathcal{L}\mathcal{L}}{\partial C_k} = \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k Z_k^{-1} \exp(-\frac{1}{2} D_{nk})}{\partial C_k} \tag{23}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \frac{\partial Z_k^{-1} \exp(-\frac{1}{2} D_{nk})}{\partial C_k} \tag{24}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \left(\frac{\partial Z_k^{-1}}{\partial C_k} \exp(-\frac{1}{2} D_{nk}) + Z_k^{-1} \frac{\partial \exp(-\frac{1}{2} D_{nk})}{\partial C_k} \right) \tag{25}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \left((-\frac{1}{2} Z_k^{-1} C_k^{-1}) \exp(-\frac{1}{2} D_{nk}) + \frac{1}{2} Z_k^{-1} \exp(-\frac{1}{2} D_{nk}) C_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) \tag{26}$$

$$= (-\frac{1}{2}) \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k Z_k^{-1} \exp(-\frac{1}{2} D_{nk}) \left(C_k^{-1} - C_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) \tag{27}$$

$$= (-\frac{1}{2}) \sum_n \gamma_{nk} \left(C_k^{-1} - C_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) \tag{28}$$

$$\tag{29}$$

Here, we used the derivative of the determinant as follows:

$$\frac{\partial Z_k^{-1}}{\partial C_k} = \frac{\partial ((2\pi)^d |C_k|)^{-\frac{1}{2}}}{\partial C_k} = ((2\pi)^d)^{-\frac{1}{2}} \frac{\partial (|C_k|)^{-\frac{1}{2}}}{\partial C_k} \tag{30}$$

$$= ((2\pi)^d)^{-\frac{1}{2}} (-\frac{1}{2}) |C_k|^{-\frac{3}{2}} \frac{\partial (|C_k|)}{\partial C_k} = ((2\pi)^d)^{-\frac{1}{2}} (-\frac{1}{2}) |C_k|^{-\frac{3}{2}} |C_k| (C_k^{-1})^T \tag{31}$$

$$= (-\frac{1}{2}) ((2\pi)^d)^{-\frac{1}{2}} |C_k|^{-\frac{1}{2}} C_k^{-1} = -\frac{1}{2} Z_k^{-1} C_k^{-1} \tag{32}$$

and the derivative of the Mahalanobis distance as:

$$\frac{\partial x^T C^{-1} x}{\partial C} = -C^{-T} x x^T C^{-T} = -C^{-1} x x^T C^{-1} \tag{33}$$

Setting $-\frac{\partial \mathcal{L}\mathcal{L}}{\partial C_k} \stackrel{!}{=} 0$ gives us:

$$\sum_n \gamma_{nk} C_k^{-1} = \sum_n \gamma_{nk} C_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \tag{34}$$

$$C_k^{-1} \sum_n \gamma_{nk} = C_k^{-1} \sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \tag{35}$$

$$\sum_n \gamma_{nk} = \sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \tag{36}$$

$$C_k = \frac{\sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n \gamma_{nk}} \tag{37}$$

Solving for the mixture coefficients: Here we must take into account that $\sum_k \pi_k = 1$. We enforce this constraint with a Lagrange multiplier. Our objective then becomes:

$$\mathcal{LL}' = \mathcal{LL} + \lambda(\sum_k \pi_k - 1) \tag{38}$$

where $\lambda < 0$.

Deriving w.r.t. π_k, we get

$$\frac{\partial \mathcal{LL}'}{\partial \pi_k} = \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k \mathcal{N}_{nk}}{\partial \pi_k} + \lambda \tag{39}$$

$$= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \mathcal{N}_{nk} + \lambda \tag{40}$$

$$= \sum_n \frac{\gamma_{nk}}{\pi_k} + \lambda \tag{41}$$

Setting equal to zero and solving for λ, we get

$$\lambda = -\sum_n \frac{\gamma_{nk}}{\pi_k} \tag{42}$$

$$\lambda \pi_k = -\sum_n \gamma_{nk} \tag{43}$$

$$\sum_k \lambda \pi_k = -\sum_k \sum_n \gamma_{nk} \tag{44}$$

$$\lambda = -N \tag{45}$$

Now we can plug this back to the objective and actually solve for π_k:

$$\frac{\partial \mathcal{LL}'}{\partial \pi_k} = \sum_n \frac{\gamma_{nk}}{\pi_k} - N \stackrel{!}{=} 0 \tag{46}$$

$$\frac{1}{\pi_k} \sum_n \gamma_{nk} = N \tag{47}$$

$$\pi_k = \frac{\sum_n \gamma_{nk}}{N} = \frac{N_k}{N} \tag{48}$$

We can interpret these results as weighted averages of means and covariances, the weights corresponding to the responsibilities γ_{nk}. The mixture coefficients π_k are simply the ratio of data points explained by each component.

Exercise 3: K-Means Compression and EM for GMM

See code.