Advanced Machine Learning
Summer 2019

Part 15 – Latent Variable Models II
06.06.2019

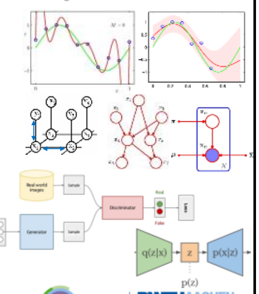
Prof. Dr. Bastian Leibe

RWTH Aachen University, Computer Vision Group
http://www.vision.rwth-aachen.de

Course Outline

- Regression Techniques
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
- Deep Reinforcement Learning
- Probabilistic Graphical Models
  - Bayesian Networks
  - Markov Random Fields
  - Inference (exact & approximate)
  - Latent Variable Models
- Deep Generative Models
  - Generative Adversarial Networks
  - Variational Autoencoders

$f : \mathcal{X} \rightarrow \mathbb{R}$



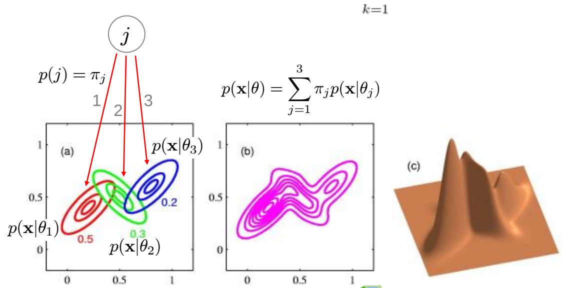
Topics of This Lecture

- Recap: Mixtures of Gaussians and General EM
  - Mixtures of Gaussians
  - General EM
- Mixtures of Gaussians revisited
  - General EM derivation
- The EM algorithm in general
  - Generalized EM
  - Relation to Variational inference

Recap: Mixtures of Gaussians

- "Generative model" $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

$p(\mathbf{x}|\theta) = \sum_{j=1}^{3} \pi_j p(\mathbf{x}|\theta_j)$

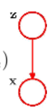


Slide credit: Bernt Schiele
Image source: C.M. Bishop, 2006

Recap: GMMs as Latent Variable Models

- Write GMMs in terms of latent variables $\mathbf{z}$
  - Marginal distribution of $\mathbf{x}$

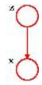
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



- Advantage of this formulation
  - We have represented the marginal distribution in terms of latent variables $\mathbf{z}$.
  - Since $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, there is a corresponding latent variable $\mathbf{z}_n$ for each data point $\mathbf{x}_n$.
  - We are now able to work with the joint distribution $p(\mathbf{x}, \mathbf{z})$ instead of the marginal distribution $p(\mathbf{x})$.
  - $\Rightarrow$ *This will lead to significant simplifications…*

Recap: Sampling from a Gaussian Mixture

- MoG Sampling
  - We can use ancestral sampling to generate random samples from a Gaussian mixture model.
    1. Generate a value $\hat{\mathbf{z}}$ from the marginal distribution $p(\mathbf{z})$.
    2. Generate a value $\hat{\mathbf{x}}$ from the conditional distribution $p(\mathbf{x}|\hat{\mathbf{z}})$.



Samples from the joint $p(\mathbf{x}, \mathbf{z})$ | Samples from the marginal $p(\mathbf{x})$ | Evaluating the responsibilities $\gamma(z_{nk})$

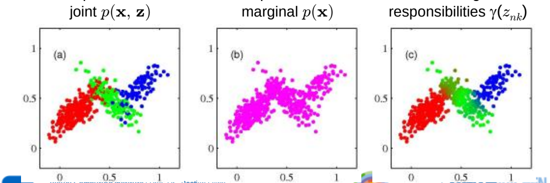


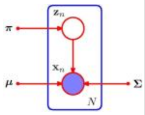
Image source: C.M. Bishop, 2006

Recap: Gaussian Mixtures Revisited

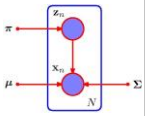
- Applying the latent variable view of EM
  - Goal is to maximize the log-likelihood using the observed data $\mathbf{X}$

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

  - Corresponding graphical model:



  - Suppose we are additionally given the values of the latent variables $\mathbf{Z}$.
  - The corresponding graphical model for the complete data now looks like this:



$\Rightarrow$ Straightforward to marginalize…

Image source: C.M. Bishop, 2006

Recap: Alternative View of EM

- In practice, however,…
  - We are not given the complete data set $\{\mathbf{X},\mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$. All we can compute about $\mathbf{Z}$ is the posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$.
  - Since we cannot use the complete-data log-likelihood, we consider instead its expected value under the posterior distribution of the latent variables:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

  - This corresponds to the E-step of the EM algorithm.
  - In the subsequent M-step, we then maximize the expectation to obtain the revised parameter set θ^{new}.

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

  where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

  4. While not converged, let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

Recap: MAP-EM

- Modification for MAP
  - The EM algorithm can be adapted to find MAP solutions for models for which a prior $p(\boldsymbol{\theta})$ is defined over the parameters.
  - Only changes needed:

  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \log p(\boldsymbol{\theta})$$

$\Rightarrow$ Suitable choices for the prior will remove the ML singularities!

Recap: Monte Carlo EM

- EM procedure
  - M-step: Maximize expectation of complete-data log-likelihood

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{Z}$$

  - For more complex models, we may not be able to compute this analytically anymore…
- Idea
  - Use sampling to approximate this integral by a finite sum over samples $\{\mathbf{Z}^{(l)}\}$ drawn from the current estimate of the posterior

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \sim \frac{1}{L} \sum_{l=1}^{L} \log p(\mathbf{X}, \mathbf{Z}^{(l)}|\boldsymbol{\theta})$$

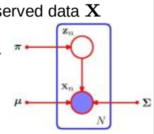
  - This procedure is called the Monte Carlo EM algorithm.

Gaussian Mixtures Revisited

- Applying the latent variable view of EM
  - Goal is to maximize the log-likelihood using the observed data $\mathbf{X}$

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

  - Corresponding graphical model:



  - Suppose we are additionally given the values of the latent variables $\mathbf{Z}$.
  - The corresponding graphical model for the complete data now looks like this:

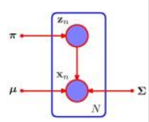


Image source: C.M. Bishop, 2006

Topics of This Lecture

- Recap: Mixtures of Gaussians and General EM
  - Mixtures of Gaussians
  - General EM
- Mixtures of Gaussians revisited
  - General EM derivation
- The EM algorithm in general
  - Generalized EM
  - Relation to Variational inference

Gaussian Mixtures Revisited

- Maximize the likelihood
  - For the complete-data set $\{\mathbf{X},\mathbf{Z}\}$, the likelihood has the form

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_{nk}}$$

  - Taking the logarithm, we obtain

$$\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\log\pi_k + \log\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

  - Compared to the incomplete-data case, the order of the sum and logarithm has been interchanged.

  $\Rightarrow$ *Much simpler solution to the ML problem*.
  - Maximization w.r.t. a mean or covariance is exactly as for a single Gaussian, except that it involves only the subset of data points that are “assigned” to that component ($z_{nk} = 1$).

Gaussian Mixtures Revisited

- Maximization w.r.t. mixing coefficients
  - More complex, since the π_k are coupled by the summation constraint

$$\sum_{j=1}^{K}\pi_j = 1$$

  - Solve with a Lagrange multiplier

$$\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

  - Solution (after a longer derivation):

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} z_{nk}$$

  $\Rightarrow$ The complete-data log-likelihood can be maximized trivially in closed form.

Gaussian Mixtures Revisited

- In practice, we don’t have values for the latent variables
  - Consider the expectation w.r.t. the posterior distribution of the latent variables instead.
  - The posterior distribution takes the form

$$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]^{z_{nk}}$$

  and factorizes over n, so that the $\{\mathbf{z}_n\}$ are independent under the posterior.
  - Expected value of indicator variable z_{nk} under the posterior.

$$\mathbb{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk}\left[\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]^{z_{nk}}}{\sum_{z_{nj}}\left[\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\right]^{z_{nj}}} = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} = \gamma(z_{nk})$$

Gaussian Mixtures Revisited

- Continuing the estimation
  - The expected value of the complete-data log-likelihood is therefore

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma z_{nk}\left\{\log\pi_k + \log\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

- Putting everything together
  - Start by choosing some initial values for $\boldsymbol{\mu}^{old}$, $\boldsymbol{\Sigma}^{old}$, and $\boldsymbol{\pi}^{old}$.
  - Use these to evaluate the responsibilities (the E-Step).
  - Keep the responsibilities fixed and maximize the above for $\boldsymbol{\mu}^{new}$, $\boldsymbol{\Sigma}^{new}$, and $\boldsymbol{\pi}^{new}$ (the M-Step).
  - This leads to the familiar closed-form solutions for $\boldsymbol{\mu}^{new}$, $\boldsymbol{\Sigma}^{new}$, and $\boldsymbol{\pi}^{new}$.

$\Rightarrow$ *This is precisely the EM algorithm for Gaussian mixtures as derived before. But we can now also apply it to other distributions.*

Topics of This Lecture

- Recap: Mixtures of Gaussians and General EM
  - Mixtures of Gaussians
  - General EM
- Mixtures of Gaussians revisited
  - General EM derivation
- The EM algorithm in general
  - Generalized EM
  - Relation to Variational inference

The EM Algorithm in General

- General formulation
  - Given a probabilistic model with observed variables $\mathbf{X}$, hidden variables $\mathbf{Z}$ and parameters $\boldsymbol{\theta}$.
  - Our goal is to maximize the likelihood given by

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

  - However, a direct optimization of $p(\mathbf{X}|\boldsymbol{\theta})$ is often difficult. Optimization of the complete-data log-likelihood $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is significantly easier.

The EM Algorithm in General

- Decomposition
  - Introduce a distribution $q(\mathbf{Z})$ over the latent variables. For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

  - where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

$$KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

  - (Proof on extra slide set)

Analysis of this Result

- Decomposition
  - For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

$$KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

- Notes (1)
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is a functional of the distribution $q(\mathbf{Z})$ and a function of the parameters $\boldsymbol{\theta}$.
  - A functional is an operator that takes as input a function and outputs again a function.

Analysis of this Result

- Decomposition
  - For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

$$KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

- Notes (2)
  - $KL(q \parallel p)$ is the Kullback-Leibler divergence between the distribution $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.
  - The KL divergence satisfies $KL(q \parallel p) \geq 0$ with $= 0$ iff $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.

Analysis of this Result

- Decomposition
  - For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

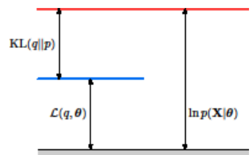
$$KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

- Notes (3)
  - It therefore follows that $\mathcal{L}(q, \boldsymbol{\theta}) \leq \log p(\mathbf{X}|\theta)$.
  - In other words: $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\log p(\mathbf{X}|\theta)$.
  - We can now use this result in order to analyze how EM works…

Analysis of EM

- Decomposition

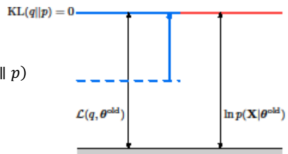
$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$



- Interpretation
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\log p(\mathbf{X}|\boldsymbol{\theta})$.
  - The approximation comes from the fact that we use an approximative distribution $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ instead of the (unknown) real posterior.
  - The KL divergence measures the difference between the approximative distribution $q(\mathbf{Z})$ and the real posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.
  - In every EM iteration, we try to make this difference smaller.

Analysis of EM

- Decomposition

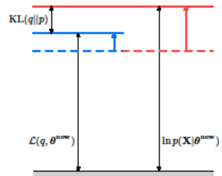


$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

- E-Step
  - Suppose the current value of the parameter vector is $\boldsymbol{\theta}^{old}$.
  - The E-step maximizes the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ w.r.t. $q(\boldsymbol{Z})$ while holding $\boldsymbol{\theta}^{old}$ fixed.
  - The solution to this maximization problem of $\log p(\mathbf{X}|\boldsymbol{\theta}^{old})$ will occur when the KL divergence vanishes, i.e. when $q(\boldsymbol{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$.
  - In this case, the lower bound equals the log-likelihood.

Analysis of EM

- Decomposition

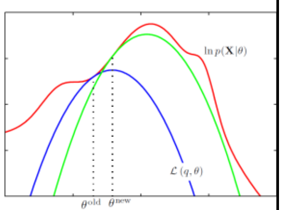


$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p)$$

- M-Step
  - In the M-step, the distribution $q(\boldsymbol{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{new}$.
  - This causes the lower bound $\mathcal{L}$ to increase (unless it is already at maximum), which will cause the log-likelihood to increase.
  - Because $q(\boldsymbol{Z})$ is determined using the old parameter values, it will not equal the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{new})$ and there will be a non-zero KL divergence.

Analysis of EM

- Visualization in the space of parameters

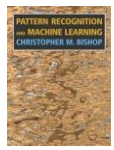


- The EM algorithm alternately
  - Computes a lower bound on the log-likelihood for the current parameters values
  - And then maximizes this bound to obtain the new parameter values.

References and Further Reading

- More information about EM and MoG estimation is available in Chapter 9 of Bishop's book (recommendable to read).

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



- Additional information
  - A.P. Dempster, N.M. Laird, D.B. Rubin, „Maximum-Likelihood from incomplete data via EM algorithm", In J. Royal Statistical Society, Series B. Vol 39, 1977
  - J.A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", TR-97-021, ICSI, U.C. Berkeley, CA,USA