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

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"
  \[ p(x) = \sum_{k=1}^{K} \tau_k \mathcal{N}(x|\mu_k, \Sigma_k) \]

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 \( z \) from the marginal distribution \( p(z) \).
  2. Generate a value \( x \) from the conditional distribution \( p(x|z) \).

Recap: GMMs as Latent Variable Models

- Write GMMs in terms of latent variables \( z \)
  - Marginal distribution of \( x \)
    \[ p(x) = \sum_{z} p(x, z) = \sum_{z} p(x|z) \pi(z) = \sum_{k=1}^{K} \pi(z_k) \mathcal{N}(x|\mu_k, \Sigma_k) \]

  - Advantage of this formulation
    - We have represented the marginal distribution in terms of latent variables \( z \).
    - Since \( p(x) = \sum_{z} p(x, z) \), there is a corresponding latent variable \( z \) for each data point \( x \).
    - We are now able to work with the joint distribution \( p(x, z) \) instead of the marginal distribution \( p(x) \).
    \[ \Rightarrow \text{This will lead to significant simplifications...} \]
Recap: Gaussian Mixtures Revisited

- Applying the latent variable view of EM
  - Goal is to maximize the log-likelihood using the observed data $X$
  $$\log p(X|\theta) = \log \left\{ \sum_Z p(X, Z|\theta) \right\}$$
  - Corresponding graphical model:
  - Suppose we are additionally given the values of the latent variables $Z$.
  - The corresponding graphical model for the complete data now looks like this:
  \[ \Rightarrow \text{Straightforward to marginalize…} \]

Recap: Alternative View of EM

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

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\theta^{(0)}$
  2. **E-step**: Evaluate $p(Z|X, \theta^{(d)})$
  3. **M-step**: Evaluate $\theta^{(n+1)}$ given by
     $$\theta^{(n+1)} = \arg \max_\theta Q(\theta, \theta^{(d)})$$
     where
     $$Q(\theta, \theta^{(d)}) = \sum_Z p(Z|X, \theta^{(d)}) \log p(X, Z|\theta)$$
  4. While not converged, let $\theta^{(d)} \leftarrow \theta^{(n+1)}$ 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 $p(\theta)$ is defined over the parameters.
  - Only changes needed:
  2. **E-step**: Evaluate $p(Z|X, \theta^{(d)})$
  3. **M-step**: Evaluate $\theta^{(n+1)}$ given by
     $$\theta^{\text{new}} = \arg \max_\theta Q(\theta, \theta^{(d)}) + \log p(\theta)$$
  \[ \Rightarrow \text{Suitable choices for the prior will remove the ML singularities!} \]

Recap: Monte Carlo EM

- EM procedure
  - **M-step**: Maximize expectation of complete-data log-likelihood
    $$Q(\theta, \theta^{(d)}) = \int p(Z|X, \theta^{(d)}) \log p(X, Z|\theta) dZ$$
  - 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 $\{Z^{(i)}\}$ drawn from the current estimate of the posterior
  $$Q(\theta, \theta^{(d)}) \approx \frac{1}{L} \sum_{i=1}^L \log p(X, Z^{(i)}|\theta)$$
  - 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 $X$
  $$\log p(X|\theta) = \log \left\{ \sum_Z p(X, Z|\theta) \right\}$$
  - Corresponding graphical model:
  - Suppose we are additionally given the values of the latent variables $Z$.
  - The corresponding graphical model for the complete data now looks like this:
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

Mixtures of Gaussians revisited
- General EM derivation
  - Mixtures of Gaussians
  - Compared to the incomplete case, the order of the sum and logarithm has been Interchanged.
  - Much simpler solution to the ML problem

Maximization w.r.t. mixing coefficients
- More complex, since the \( \pi_k \) are coupled by the summation constraint
  \[ \sum_{k=1}^{K} \pi_k = 1 \]
- Solve with a Lagrange multiplier
  \[ \log p(X, Z | \mu, \Sigma, \pi) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right) \]
- 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.

Continuing the estimation
- The expected value of the complete-data log-likelihood is therefore
  \[ \mathbb{E}_q[\log p(X, Z | \mu, \Sigma, \pi)] = \sum_{k=1}^{K} \gamma_{nk} \log \pi_k + \log p(x_n | \mu_k, \Sigma_k) \]
- Putting everything together
  - Start by choosing some initial values for \( \mu^{old}, \Sigma^{old}, \text{ and } \pi^{old} \)
  - Use these to evaluate the responsibilities (the E-Step)
  - Keep the responsibilities fixed and maximize the above for \( \mu^{new}, \Sigma^{new}, \text{ and } \pi^{new} \) (the M-Step)
  - This leads to the familiar closed-form solutions for \( \mu^{new}, \Sigma^{new}, \text{ and } \pi^{new} \)
\( \Rightarrow \) This is precisely the EM algorithm for Gaussian mixtures as derived before. But we can now also apply it to other distributions.
**The EM Algorithm in General**

- **General formulation**
  - Given a probabilistic model with observed variables $X$, hidden variables $Z$ and parameters $\theta$.
  - Our goal is to maximize the likelihood given by
    \[ p(X|\theta) = \sum_Z p(X,Z|\theta) \]
  - However, a direct optimization of $p(X|\theta)$ is often difficult. Optimization of the complete-data log-likelihood $p(X,Z|\theta)$ is significantly easier.

**Analysis of this Result**

- **Decomposition**
  - For any choice of $q(Z)$, the following decomposition holds
    \[ \log p(X|\theta) = L(q,\theta) + KL(q \parallel p) \]
    \[ L(q,\theta) = \sum_Z q(Z) \log \left( \frac{p(X,Z|\theta)}{q(Z)} \right) \]
    \[ KL(q \parallel p) = -\sum_Z q(Z) \log \left( \frac{p(X,Z|\theta)}{q(Z)} \right) \]
  - It therefore follows that $L(q,\theta) \leq \log p(X|\theta)$.
  - In other words: $L(q,\theta)$ is a lower bound on $\log p(X|\theta)$.
  - We can now use this result in order to analyze how EM works…

**Analysis of EM**

- **Decomposition**
  - For any choice of $q(Z)$, the following decomposition holds
    \[ \log p(X|\theta) = L(q,\theta) + KL(q \parallel p) \]
  - $L(q,\theta)$ is a lower bound on $\log p(X|\theta)$.
  - The approximation comes from the fact that we use an approximative distribution $q(Z) = p(Z|X,a^{old})$ instead of the (unknown) real posterior.
  - The KL divergence measures the difference between the approximative distribution $q(Z)$ and the real posterior $p(Z|X,\theta)$.
  - In every EM iteration, we try to make this difference smaller.
Analysis of EM

• Decomposition

\[
\log p(\mathbf{X} | \theta) = \mathcal{L}(q, \theta) + KL(q \| p)
\]

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

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

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