Advanced Machine Learning
Summer 2019
Part 16 – Latent Variable Models III
19.06.2019

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: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Recap: General EM Algorithm

1. Choose an initial setting for the parameters \( \theta^{(0)} \)
2. E-step: Evaluate \( p(Z|X, \theta^{(l)}) \)
3. M-step: Evaluate \( \theta^{(l+1)} \) given by
   \[
   \theta^{(l+1)} = \arg \max_{\theta} Q(\theta, \theta^{(l)})
   \]
   where
   \[
   Q(\theta, \theta^{(l)}) = \sum_Z p(Z|X, \theta^{(l)}) \log p(X, Z|\theta)
   \]
4. While not converged, let \( \theta^{(l+1)} \) and return to step 2.

Recap: The EM Algorithm in General

- Decomposition
  - Introduce a distribution \( q(Z) \) over the latent variables. For any choice of \( q(Z) \), the following decomposition holds
    \[
    \log p(X|\theta) = L(q, \theta) + KL(q \parallel p)
    \]
    where
    \[
    L(q, \theta) = \sum_Z q(Z) \log \left( \frac{p(X,Z|\theta)}{q(Z)} \right)
    \]
    \[
    KL(q \parallel p) = -\sum_Z q(Z) \log \left( \frac{p(X,Z|\theta)}{q(Z)} \right)
    \]
    \( KL(q \parallel p) \) is the Kullback-Leibler divergence between the distribution \( q(Z) \) and the posterior distribution \( p(Z|X, \theta) \).
    \( L(q, \theta) \) is a functional of the distribution \( q(Z) \) and a function of the parameters \( \theta \). Since \( KL \geq 0 \), \( L(q, \theta) \) is a lower bound on \( \log p(X|\theta) \).

Recap: Analysis of EM

- Decomposition
  \[
  \log p(X|\theta) = L(q, \theta) + KL(q \parallel p)
  \]
- Interpretation
  - \( L(q, \theta) \) is a lower bound on \( \log p(X|\theta) \).
    - The approximation comes from the fact that we use an approximative distribution \( q(Z) = p(Z|X, \theta^{(l)}) \) 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.

Recap: Analysis of EM
Recap: 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 minimizes this bound to obtain the new parameter values.

Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Motivation

- Recall: Bayesian estimation
\[ p(x|X) = \int p(x|\theta)p(\theta) \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta) d\theta} d\theta \]
- So far, we have only done this for Gaussian distributions, where the integrals could be solved analytically.
- Now, let’s also examine other distributions...

Conjugate Priors

- Problem: How to evaluate the integrals?
  - We will see that if likelihood and prior have the same functional form \( c f(x) \), then the analysis will be greatly simplified and the integrals will be solvable in closed form.
  - Such an algebraically convenient choice is called a conjugate prior.
  - Whenever possible, we should use it.
  - To do this, we need to know for each probability distribution what is its conjugate prior. ⇒ Topic of this lecture.
  - What to do when we cannot use the conjugate prior?
    ⇒ Use approximate inference methods.

The Multinomial Distribution

- Joint distribution over \( m_1, \ldots, m_K \) conditioned on \( \mu \) and \( N \)
  \[ \text{Mult}(m_1, m_2, \ldots, m_K | \mu, N) = \binom{N}{m_1, m_2, \ldots, m_K} \prod_{k=1}^{K} m_k^{m_k} \]
  with the normalization coefficient
  \[ \binom{N}{m_1, m_2, \ldots, m_K} = \frac{N!}{m_1! m_2! \cdots m_K!} \]
- Properties
  \[ \mathbb{E}[m_k] = \frac{N \mu_k}{N} \]
  \[ \text{var}[m_k] = \frac{N \mu_k (1 - \mu_k)}{N} \]
  \[ \text{cov}[m_k, m_l] = -\frac{N \mu_k \mu_l}{N} \]

Bayesian Multinomial

- Conjugate prior for the Multinomial
  - Introduce a family of prior distributions for the parameters \( \{\mu_k\} \) of the Multinomial.
  - The conjugate prior is given by
    \[ p(\mu | \alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \]
    with the constraints
    \[ \forall k: 0 \leq \mu_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1 \]
The Dirichlet Distribution

- **Dirichlet Distribution**
  - Multivariate generalization of the Beta distribution
  \[
  \text{Dir}(\alpha) = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}{\Gamma(\sum_k \alpha_k)} \prod_k \alpha_k^{\alpha_k-1}
  \]
  with \( \alpha_0 = \sum_k \alpha_k \)

- **Properties**
  - Conjugate prior for the Multinomial.
  - The Dirichlet distribution over \( K \) variables is confined to a \( K-1 \) dimensional simplex.
  - Expectations:
    \[
    E[\alpha_k] = \alpha_k \alpha_0 / \alpha_0 \\
    \text{var}[\alpha_k] = \alpha_k (\alpha_0 - \alpha_k) / \alpha_0 (\alpha_0 + 1) \\
    \text{cov}[\alpha_j, \alpha_k] = -\alpha_0 \alpha_j \alpha_k / \alpha_0 (\alpha_0 + 1)
    \]

Recap: The Gaussian Distribution

- **One-dimensional case**
  - Mean \( \mu \)
  - Variance \( \sigma^2 \)
  \[
  N(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}
  \]

- **Multi-dimensional case**
  - Mean \( \mu \)
  - Covariance \( \Sigma \)
  \[
  N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} \sqrt{\det \Sigma}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
  \]

Bayesian Inference for the Gaussian

- **Univariate conjugate priors**
  - \( \sigma^2 \) known, \( \mu \) unknown: \( p(\mu) \text{ Gaussian} \)
  \[
  p(X | \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_n (x_n - \mu)^2 \right\}
  \]
  - \( \mu \) is known, \( \lambda \) unknown: \( p(\lambda) \text{ Gamma} \)
  \[
  p(X | \lambda) \propto \lambda^{N/2} \exp \left\{ -\frac{\lambda}{2} \sum_n (x_n - \mu)^2 \right\}
  \]
  - both \( \mu \) and \( \lambda \) unknown: \( p(\mu, \lambda) \text{ Gaussian-Gamma} \)
  \[
  p(X | \mu, \lambda) = \left[ \frac{1}{\lambda^{1/2}} \exp \left\{ -\frac{\lambda}{2} \right\} \right]^N \exp \left\{ -\frac{\lambda}{2} \sum_n (x_n - \mu)^2 \right\}
  \]

Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Towards a Full Bayesian Treatment...

- **Mixture models**
  - We have discussed mixture distributions with \( K \) components
  \[
  p(X | \theta) = \sum_Z p(X, Z | \theta)
  \]
  - So far, we have derived the ML estimates \( \Rightarrow \text{EM} \)
  - Introduced a prior \( p(\theta) \) over parameters \( \Rightarrow \text{MAP-EM} \)
  - One question remains open: how to set \( K \)?
    \( \Rightarrow \) Let's also set a prior on the number of components...
Bayesian Mixture Models

- Let's be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable $z_n$ as indicator which component data point $x_n$ belongs to.

$$z_n | \pi \sim \text{Multinomial}(\pi)$$
$$x_n | z_n = k \sim N(\mu_k, \Sigma_k)$$

- This is similar to the graphical model we've used before, but now the $\pi$ and $\theta_l = (\mu_k, \Sigma_k)$ are also treated as random variables.
- What would be suitable priors for them?

Bayesian Mixture Models

- Full Bayesian Treatment
  - Given a dataset, we are interested in the cluster assignments

$$p(Z|X) = \frac{p(X|Z)p(Z)}{\sum_K p(X|Z_K)p(Z_K)}$$

where the likelihood is obtained by marginalizing over the parameters $\theta$

$$p(X|Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_n | z_{nk}, \theta_k)$$

- The posterior over assignments is intractable!
  - Denominator requires summing over all possible partitions of the data into $K$ groups!
  - Need efficient approximate inference methods to solve this...

Recap: The Dirichlet Distribution

- Dirichlet Distribution
  - Conjugate prior for the Categorical and the Multinomial distrib.

$$\text{Dir}(\mu | \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \alpha_k^{\mu_k - 1}$$

- Symmetric version (with concentration parameter $\alpha$)

$$\text{Dir}(\mu | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \alpha_k^{\mu_k - 1}$$

- Properties

$$E[\mu_k] = \frac{\alpha_k}{\alpha_0}$$

$$\text{var}[\mu_k] = \frac{\alpha_k (\alpha_0 - \alpha_k)}{\alpha_0^2 (\alpha_0 + 1)}$$

$$\text{cov}[\mu_k, \mu_l] = \frac{\alpha_k \alpha_l}{\alpha_0^2 (\alpha_0 + 1)}$$

- More structure appears as more points are drawn
Visual Computing Institute | Prof. Dr. Bastian Leibe
Advanced Machine Learning
Part 16 – Latent Variable Models III

**Dirichlet Samples**

- Effect of concentration parameter $\alpha$
  - Controls sparsity of the resulting samples

- Finite mixture of $K$ components
  \[ p(x_n | \theta) = \sum_{k=1}^{K} \pi_k p(x_n | \theta_k) \]
  - The distribution of latent variables $x_n$ given $\pi$ is multinomial
  \[ p(x | \pi) = \prod_{k=1}^{K} \pi_k^{n_k} \]
  - Assume mixing proportions have a given symmetric conjugate Dirichlet prior

**Mixture Model with Dirichlet Priors**

- Integrating out the mixing proportions $\pi$:
  \[ p(x) = \int p(x | \pi)p(\pi) d\pi = \int \prod_{k=1}^{K} \pi_k^{N_k} \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} d\pi \]
  - This is again a Dirichlet distribution (reason for conjugate priors)

- Conditional probabilities
  - Let’s examine the conditional of $x_n$ given all other variables $p(z_{nk} = 1 | z_{-n}, \alpha) = \frac{p(z_{nk} = 1, z_{-n}, \alpha)}{p(z_{-n}, \alpha)}$ where $z_n$ denotes all indices except $n$. 
Finite Dirichlet Mixture Models

- Conditional probabilities: Finite $K$
  \[ p(z_{nk} = 1|z_{-n}, \alpha) = \frac{N_{n,k} + \alpha/K}{N - 1 + \alpha}, \quad \text{with} \quad N_{n,k} = \sum_{i=1, j \neq n}^{N} z_{ik} \]

- This is a very interesting result. Why?
  - We directly get a numerical probability, no distribution.
  - The probability of joining a cluster mainly depends on the number of existing entries in a cluster.
    \[ \Rightarrow \text{The more populous a class is, the more likely it is to be joined!} \]
  - In addition, we have a base probability of also joining as-yet empty clusters.

- This result can be directly used in Gibbs Sampling…
(see later derivation)

Infinite Dirichlet Mixture Models

- Conditional probabilities: Finite $K$
  \[ p(z_{nk} = 1|z_{-n}, \alpha) = \frac{N_{n,k} + \alpha/K}{N - 1 + \alpha}, \quad \text{with} \quad N_{n,k} = \sum_{i=1, j \neq n}^{N} z_{ik} \]

- Conditional probabilities: Infinite $K$
  - Taking the limit as $K \rightarrow \infty$ yields the conditionals
  \[ p(z_{nk} = 1|z_{-n}, \alpha) = \begin{cases} \frac{N_{n,k}}{N - 1 + \alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N - 1 + \alpha} & \text{if all } k \text{ not represented} \end{cases} \]

  - Left-over mass $\alpha \Rightarrow$ countably infinite number of indicator settings

Discussion

- Infinite Mixture Models
  - What we have just seen is a first example of a Dirichlet Process.
  - DPs allow us to work with models that have an infinite number of components.
  - This will raise a number of issues
    - How to represent infinitely many parameters?
    - How to deal with permutations of the class labels?
    - How to control the effective size of the model?
    - How to perform efficient inference?
  \[ \Rightarrow \text{More background needed here!} \]
  - DPs are a very interesting class of models, but would take us too far here.
  - If you’re interested in learning more about them, take a look at the Advanced ML slides from Winter 2012.

Topics of This Lecture

- Recap: General EM
- Bayesian Estimation Revisited
  - Conjugate priors
  - Probability distributions
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
- Approximate Inference for Bayesian Mixture Models
  - Gibbs Sampler

Recap: Gibbs Sampling

- Approach
  - MCMC-algorithm that is simple and widely applicable.
  - May be seen as a special case of Metropolis-Hastings.
- Idea
  - Sample variable-wise: replace $x_i$ by a value drawn from the distribution $p(z|x_i)$.
    \[ \Rightarrow \text{This means we update one coordinate at a time.} \]
  - Repeat procedure either by cycling through all variables or by choosing the next variable.
- Properties
  - The algorithm always accepts!
  - Completely parameter free.
  - Can also be applied to subsets of variables.
Gibbs Sampling for Finite Mixtures

- **Standard finite mixture sampler**
  - Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample new parameters as follows:
  1. Independently assign each point $x_n$ to one of the $K$ clusters by sampling the variables $z_n$ from the multinomial distributions
  
  
  $z_n^{(t)} = \arg\max_{k=1}^{K} \left[ \frac{\pi_k^{(t-1)} \cdot g(x_n | \theta_k^{(t-1)})}{Z_n^{(t-1)}} \right]$

  2. Sample new mixture weights from the Dirichlet distribution
  
  
  $\pi^{(t)} \sim \text{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K)$

  3. For each of the $K$ clusters, independently sample new parameters from the conditional of the assigned observations
  
  $\theta_k^{(t)} \sim p(\theta | \{x_n | z_n = k\}, H) $

- **We need approximate inference here**
  - **Gibbs Sampling**: Conditionals are simple to compute
  
  
  $p(z_n = k | \text{others}) \propto \sum_{k=1}^{K} \pi_k N(x_n | \mu_k, \Sigma_k) ^{z_n}$

  $\pi | z \sim \text{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K)$

  $\mu_k, \Sigma_k | \text{others} \sim \mathcal{N}(\mu', \Sigma', \sigma')$

- **However, this will be rather inefficient...**
  - In each iteration, algorithm can only change the assignment for individual data points.
  - There are often groups of data points that are associated with high probability to the same component. => Unlikely that group is moved.
  - Better performance by collapsed Gibbs sampling which integrates out the parameters $\pi, \mu, \Sigma$. 

- **More efficient algorithm**
  - Conjugate priors allow analytic integration of some parameters
  - Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)

- **Procedure**
  - The model implies the factorization
  
  
  $p(z_n | x_n, \alpha, H) \propto p(z_n | \alpha) p(x_n | z_n, H)$

  - Derive
  
  
  $p(z | \alpha) = \int p(z | \pi) p(\pi | \alpha) d\pi$

- **Collapsed Finite Bayesian Mixture**
Recap: Mixture Models with Dirichlet Priors

- Integrating out the mixing proportions $\pi$
  
  $$p(z|\alpha) = \int p(z|\pi)p(\pi|\alpha)d\pi$$

  $$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}$$

- Conditional probabilities
  
  - Examine the conditional of $z_n$ given all other variables $z_{-n}$

  $$p(z_n = 1|z_{-n}, \alpha) = \frac{p(z_n = 1, z_{-n}|\alpha) p(z_{-n}|\alpha)}{p(z_{-n}|\alpha)} = \frac{N_{n,k} + \alpha/K}{N - 1 + \alpha}$$

  $N_{n,k}$ def $\sum_{i=1}^{N} z_{ni}$

  ⇒ The more populous a class is, the more likely it is to be joined.

Collapsed (Rao-Blackwellized) Finite Mixture Sampler

- Algorithm

  1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
  2. Set $z = x^{(1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, sequentially resample $z_i$ as follows:
    a. For each of the $K$ clusters, determine the predictive likelihood (this can be computed from cached sufficient statistics)
      
      $$p_k(x_i|z_{-i}, H) = p(x_i|z_{-i} = 1, m \neq n), H)$$
    b. Sample a new assignment $z_i$ from the multinomial distribution
      
      $$z_i \sim \sum_{k=1}^{K} \frac{N_{n,k} + \alpha/K}{N - 1 + \alpha} p_k(x_i|z_{-i}, H)$$
    c. Update cached sufficient statistics to reflect assignment $z_i$.
  3. Set $x^{(t)} = z$. Optionally, mixture parameters may be sampled via steps 2-3 of the standard finite mixture sampler.

Standard vs. Collapsed Samplers

⇒ Collapsed sampler converges much more quickly.

- Theorem (Rao-Blackwell)

  "Analytical marginalization of some variables from a joint distribution always reduces the variance of later estimates."

Discussion

-Collapsed Gibbs sampling
  
  Integrates out the parameters $\pi, \mu, \Sigma$

  $$p(z_{nk} = 1|\text{others}) \propto \frac{(N_{n,k} + \alpha/K)}{N - 1 + \alpha} p_k(x_n|z_{-n}, H)$$

- Properties
  
  - Can change all assignments in each iteration.
  - Able to move entire groups between clusters.
  - Faster convergence, less likely to get stuck.

References and Further Reading

- Unfortunately, there are currently no good introductory textbooks on the Dirichlet Process. We therefore recommend a number of tutorial papers on their different aspects.

  - One of the best available general introductions
  - A gentle introductory tutorial (recommended 1st read)
  - Good overview of MCMC methods for DPMMs