Imperial College London

COURSEWORK

PROBABILISTIC INFERENCE (CO-493)

Mean Field Variational Inference

Mean Field Variational Inference

Coursework Description

Variational inference is a powerful tool in modern machine learning that frames the problem of approximating intractable densities as an optimization problem. The goal of this coursework is to build an intuitive understanding of variational inference by implementing the necessary steps for approximate inference in a Gaussian Mixture Model (GMM).

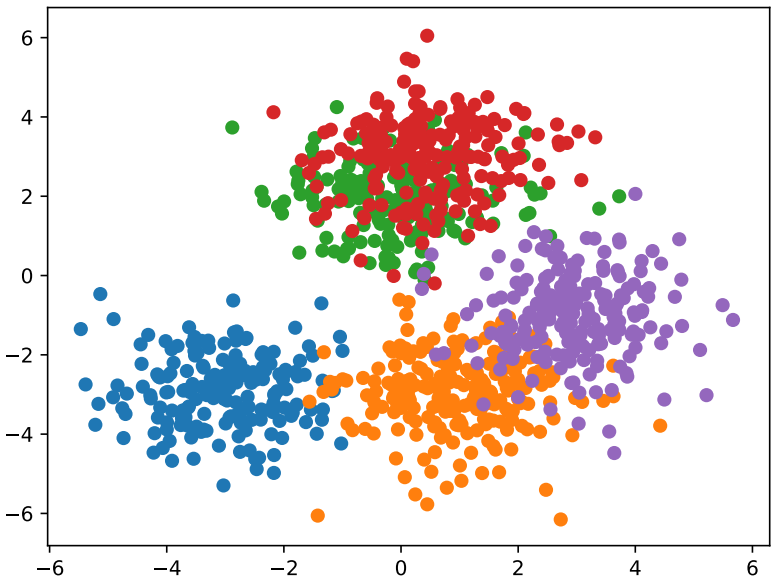

Figure 1: Samples from a 2D mixture of Gaussians with five components.

You are given examplary 2D data from a mixture of Gaussians with five components in Figure 1. The aim is to model the dataset of a D-dimensional GMM with K components. The generative process is specified as:

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}),\ k = 1, \ldots, K, \tag{1}$$

$$\boldsymbol{c}_i \sim \text{Cat}(1/K),\ i = 1, \ldots, N, \tag{2}$$

$$\boldsymbol{x}_i | \boldsymbol{c}_i, \boldsymbol{\mu} \sim \prod_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I})^{c_{ik}},\ i = 1, \ldots, N,\ k = 1, \ldots, K, \tag{3}$$

where $\boldsymbol{x}_i \in \mathbb{R}^D$ are the observations, $\boldsymbol{c}_i \in \{0, 1\}^K$ are one-hot vectors indicating mixture component assignment, $\boldsymbol{\mu}_k \in \mathbb{R}^D$ are the means of the mixture components, $\sigma^2 \in \mathbb{R}$ is the prior variance on the means which is assumed to be fixed across dimensions and $\boldsymbol{I}$ denotes the identity matrix.

Note that we have only specified a prior distribution over the means of the mixture components and the variance of the observations is assumed to be known and fixed. This is a (partially) Bayesian GMM where the latent variables are the means of the mixture components $\boldsymbol{\mu}_k$ and the cluster assignment variables $\boldsymbol{c}_i$.

Given the dataset $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ the objective is to approximate the posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{c}|\boldsymbol{X})$. To do so, we use a **mean field** variational approximation (all the latent variables factorize):

$$q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(\boldsymbol{c}_i) \tag{4}$$

where $q(\boldsymbol{\mu}_k)$ is a Gaussian distribution on the mean of the k-th mixture component, $q(\boldsymbol{c}_i)$ is a categorical distribution on the mixture assignment of the i-th observation. Formally, we have

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{m}_k, \text{diag}(\boldsymbol{s}_k^2)),\ k = 1, \ldots, K \tag{5}$$
$$\boldsymbol{c}_i \sim \text{Cat}(\boldsymbol{\pi}_i),\ i = 1, \ldots, N \tag{6}$$

where $\boldsymbol{m}_k \in \mathbb{R}^D$, $\boldsymbol{s}_k^2 \in \mathbb{R}^D$ and $\boldsymbol{\pi}_i \in [0,1]^K$ are variational parameters that we are going to optimize. Note that $\text{diag}(\boldsymbol{s}_k^2)$ is a diagonal $D \times D$ matrix, i.e. we have assumed that the means of the mixture components factorize across dimensions.

Coordinate ascent variational inference (CAVI) is an iterative algorithm that finds a local optimum of the evidence lower bound (ELBO). CAVI takes advantage of the mean field assumption and updates each factor according to:

$$q^*(z_j) \propto \exp\left(\mathbb{E}_{-j}[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{X})]\right) \tag{7}$$

where the subscript $-j$ denotes all variables except the j-th one, so that the expectation is with respect to all variational factors except j.

File Descriptions / Submission

- `vi_assignment.py`: Skeleton file that will be graded. Put all your code in this file.
- `vi_extra.ipynb`: Jupyter notebook where you can plot the data, run CAVI and visualize the variational distribution after implementing the code.

Submit your final version to CATe via the LabTS system. **Use Python 3.6 for your implementation.**

The Evidence Lower Bound

The ELBO is a lower bound on the log-marginal likelihood $\log p(\boldsymbol{X}) = \log \int_{\mu,c} p(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{c})$ which is generally intractable. The ELBO is a key feature of variational inference and is given by:

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{c})}\left[\frac{p(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{c})}{q(\boldsymbol{\mu}, \boldsymbol{c})}\right] \tag{8}$$

Task 1: 20 Marks

Complete the function `log_joint()` which returns the expectation of the log-density of the model under the variational distribution: $\mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{c})}\Big[\log p(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{c})\Big]$.

NOTE: Make sure the function returns the value after throwing away constant terms, meaning terms that do not depend on any variational parameters.
NOTE: An n-dimensional Gaussian with mean μ and diagonal covariance matrix $\text{diag}(\sigma^2) = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ is the same as a collection of n independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.

Task 2: 15 Marks

Complete the function `log_var()` which returns the expectation of the log-variational-density of the model under the variational distribution: $\mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{c})}\Big[\log q(\boldsymbol{\mu}, \boldsymbol{c})\Big]$.

NOTE: Make sure the function returns the value after throwing away constant terms, meaning terms that do not depend on any model parameters or variational parameters.

Task 3: 5 Marks

Complete the `elbo()` function using the `log_joint()` and `log_var()` functions from above.

Variational Parameter Updates

Equation (7) is the general form for the optimal updates. In order to derive the updates you can plug in the model definition, throw away constant terms and perform the relevant expectations.

Task 4: 20 Marks

Complete the function `update_pi()` which returns the normalized probabilities of cluster assignments.

NOTE: After throwing away constant terms you are left with an expression $\boldsymbol{\pi}_i \propto [\exp(p_1),..,\exp(p_K)]$. Normalize by dividing by the sum across K for each i.

Task 5: 20 Marks

Complete the function `update_m()` which returns the updated mean parameters for $q(\boldsymbol{\mu})$.

NOTE: You might find the Gaussian distribution in its canonical form useful for the derivation:

$$p(x|\eta,\Lambda) = \alpha \exp\{\eta^T x + \frac{1}{2} x^T \Lambda x\} \tag{9}$$

Coverting the canonical parameters η, Λ to moment parameters, we have $\mu = \Lambda^{-1}\eta$ and $\sigma^2 = \Lambda^{-1}$. α is a normalizing constant.
NOTE: Remember to throw away constant terms.

Task 6: 20 Marks

Complete the function `update_s2()` which returns the updated variance parameters for $q(\boldsymbol{\mu})$.

NOTE: Remember to throw away constant terms.