Lecture Pattern Analysis

Part 10: Variational Inference

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
June 10, 2022

Introduction

- Recall that GMMs are fitted in a Bayesian framework by optimizing for $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$
- A further extension also includes the model selection task, i.e., to automatically select the number of mixture components K
- However, the joint distribution is intractable, hence we will need to discuss advanced inference techniques for that
    1. Variational inference: approximates the intractable function with a simpler, tractable function
    2. Markov Chain Monte Carlo Sampling: approximates the intractable function with a finite number of samples
- In this lecture, we use variational inference (VI)[1]

[1] This lecture refers to Bishop Chapter 10 until 10.2.4 (VI), Bishop Appendix D (Variational Calculus), and Bishop 9.4 (EM via KL divergence), and the excellent blog entry here: https://mpatacchiola.github.io/blog/2021/01/25/intro-variational-inference.html

Illustration: What is Variational Inference — Example Tasks

- VI = Calculus to find a nested function that optimizes another function

- Example nested function: length(pathBetween(A, B))
  Q: What function pathBetween() gives the shortest length for any input A, B?
  A: The line between A and B

- Example nested function: Entropy $H[p] = -\int p(x) \log p(x) \,\mathrm{d}x$
  Q: What PDFs $p(x)$ minimize/maximize the integral?
  A: Any Dirac impulse/uniform distribution

- Our nested function: $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$
  Q: What distribution $p(\mathbf{x}, \mathbf{z})$ maximizes the sum?
  A: ?

Variational Calculus (cf. Bishop Appendix D)

- Let $F[y]$ denote a functional, i.e., an operator that takes a function $y(x)$ and returns an output value F
- Our (school-) calculus: find x that maximizes $y(x)$
- Var. calculus is analogous! Find function $y(x)$ to max./min. functional $F[y]$
  - The optimum is found through the functional derivative
  - Derivatives can be calculated from ϵ changes (“variations”!) in the input
- The functional derivative of $F[y]$ with respect to $y(x)$ is denoted as $\delta F / \delta y(x)$, s.t.

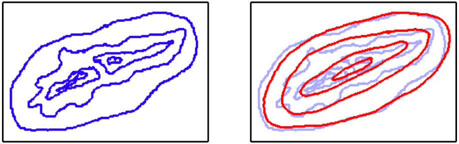
$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) \, \mathrm{d}x + O(\epsilon^2) \ , \qquad (1)$$

  where we do a small change $\epsilon\eta(x)$ with $\eta(x)$ being an arbitrary function of x.
- For optimization, we seek a functional that is stationary for arbitrary $\eta(x)$, s.t.

$$\int \frac{\delta F}{\delta y(x)} \eta(x) \, \mathrm{d}x = 0 \qquad (2)$$

Approximations in Variational Calculus

- Theoretically, solutions from Variational Calculus are exact (“there is nothing approximate”)
- However, in most applications, including Bayesian Inference, the framework is attractive because it directly allows to do approximations
- More specifically, when $y(x)$ is difficult to optimize within $F[y]$, it can be replaced by a more convenient function $q(x)$
- Example: Left: function with a complicated integral. Right: Gaussian



- The approximation error between y and q is quantified with the Kullback-Leibler (KL) Divergence $\mathrm{KL}(y||q)$

Approximation in Variational Inference for GMM Fitting

- In GMM fitting with the EM algorithm, it is difficult to directly optimize $p(\mathbf{X}|\boldsymbol{\theta})$, and easier to optimize the complete-data log-likelihood $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$.
  We use this difficulty as a precondition for applying VI.
- In our specific task, we seek to find the marginal probability $p(\mathbf{X})$,

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{3}$$

  where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \tag{4}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \ . \tag{5}$$

  (the full derivation is in Bishop Sec. 9.4)
- There are two problems: calculating $p(\mathbf{X})$ is intractable, and calculating the KL divergence is intractable

The Trick with the Evidence Lower Bound (ELBO)

- There are two problems: calculating $p(\mathbf{X})$ is intractable, and calculating the KL divergence is intractable
- The trick to deal with this is to use the identity

$$\mathrm{KL}(q\|p) = \mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) = \mathbb{E}_q\left[\log q(\mathbf{Z}) - \log p(\mathbf{X},\mathbf{Z})\right] + \log p(\mathbf{x}) \tag{6}$$

and to rearrange these terms to

$$\mathbb{E}_q\left[\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z})\right] = \log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) \tag{7}$$

- Here, the right-hand side has the intractable terms, where we seek to maximize $\log p(\mathbf{x})$ and to minimize $\mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X}))$
- The left-hand side is tractable, and is called the “Evidence Lower Bound” (ELBO): Maximizing ELBO also optimizes the terms on the right-hand side.
- A good derivation is here: https://mpatacchiola.github.io/blog/2021/01/25/intro-variational-inference.html

Introduction to GMM Fitting via VI

- We use samples $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and latent variables $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$
- Let us rewrite the GMM equations with explicit parameters,

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{8}$$

and

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Lambda_k^{-1})^{z_{nk}} \quad \text{where} \quad \Lambda = \Sigma^{-1} \tag{9}$$

- The Bayesian way of model selection is to add priors to the parameters
- Hence, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, Λ require suitable priors

Remarks on Priors

- Priors for $\boldsymbol{\mu}_i$, Λ_i, π_i allow to “draw” new components, and to regularize fits on limited data
- Important: if likelihood and prior are **conjugate**, their product is from the same family of distributions as the likelihood
- Conjugate priors make the calculation **much** easier — always choose conjugate priors!.
- Priors bring additional parameters
  - The prior often has one more parameter than its associated likelihood — so nothing gained?
  - Not quite: the influence of the prior’s parameters is very indirect, and further reduces with increasing dataset size
  - Hence, these parameters can be rather generic (“1”, “mean of the data”, ...)

Specific Priors: Dirichlet Distribution for the Mixing Coefficients

- The mixing coefficients $\boldsymbol{\pi}$ forms a **multinomial distribution** that can be seen as the relative number of samples that belong to a component
- The conjugate prior of the multinomial distribution is the Dirichlet distribution[2]

$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{10}$$

with k identical positive values α_0 in $\boldsymbol{\alpha}_0$, i.e., $\boldsymbol{\alpha}_0 = (\alpha_0, \ldots, \alpha_0) \in \mathbb{R}_+^k$, and a scaling factor

$$C(\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum\limits_{k=1}^{K} \alpha_k\right)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_k)} \tag{11}$$

[2] A list of distributions together with short descriptions and their role as conjugate prior (if applicable) is in Appendix B in the book by Bishop

Specific Priors: Gaussian-Wishart for Mean and Precision Matrix

- The prior for mean and precision matrix Λ (recall $\Lambda = \Sigma^{-1}$) is

$$p(\boldsymbol{\mu}, \Lambda) = p(\boldsymbol{\mu}|\Lambda)p(\Lambda) \tag{12}$$

$$= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1}\right)\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0) \tag{13}$$

where $\mathbf{m}_0$, β, $\mathbf{W}_0$, ν_0 are hyperpriors, and[3]

$$\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0) = B(\mathbf{W}_0, \nu_0)|\Lambda|^{(\nu_0-D-1)/2}\text{exp}\left(-\frac{1}{2}\text{Tr}(\mathbf{W}_0^{-1}\Lambda)\right) \tag{14}$$

with feature dimensionality D and the normalizing constant

$$B(\mathbf{W}_0, \nu_0) = |\mathbf{W}_0|^{-\nu_0/2}\left(2^{\nu_0 D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right)^{-1} \tag{15}$$

[3] Formally, the Wishart distribution is the distribution of sample covariance matrices. As stated also later in this lecture, you do not need to memorize this equation.

Joint Distribution and Variational Approximation

- With our priors, the joint GMM distribution can be written as

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\Lambda)p(\Lambda) \tag{16}$$

- We use Bishop's variational framework (Sec. 10-10.2.1) for inference
- Note that this is just an illustration. The details are omitted on purpose[4].
- Bishop's variational framework approximates the distribution $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$ using a distribution $q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$ with the independence assumption

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) \tag{17}$$

- After several calculations, the solution is again an EM algorithm

[4] However, I cordially invite those of you who are curious to know more to a special meeting, where we can go through this Section in full detail

EM Solution: Expectation Step

- Calculate responsibilities

$$r_{ik} = \mathbb{E}[z_{nk}] = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \tag{18}$$

where

$$\rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\Lambda_k}[(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\Lambda_k(\mathbf{x}_n - \boldsymbol{\mu}_k)] \tag{19}$$

- Our EM algorithm for standard GMM fitting looks somewhat simpler.
- However, the main difference here is that we operate on expectations over distributions (which are induced from the priors).
- The next slide lists the expanded equations for the expectations, but again only for illustration

EM Solution: Expectation Step / Expanded Equations

- Expanded equations for the expectations:

$$\mathbb{E}_{\boldsymbol{\mu}_k,\Lambda_k}[(\mathbf{x}_n-\boldsymbol{\mu}_k)^\mathsf{T}\Lambda_k(\mathbf{x}_n-\boldsymbol{\mu}_k)] = D\beta_k^{-1}+\nu_k(\mathbf{x}_n-\mathbf{m}_k)^\mathsf{T}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k) \tag{20}$$

$$\mathbb{E}[\ln|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right)+D\ln 2+\ln|\mathbf{W}_k| \tag{21}$$

$$\mathbb{E}[\ln\pi_k] = \psi(\alpha_k)-\sum_{k=1}^{K}\alpha_k \tag{22}$$

where $\psi(\alpha)$ is the digamma function,

$$\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln\Gamma(\alpha) \tag{23}$$

EM Solution: Maximization Step / Mixing Coefficients

- The distribution of mixing coefficients is updated via the Dirichlet distribution

$$q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{24}$$

where each component α_k of $\boldsymbol{\alpha}$ is updated with the sum of its responsibilities

$$\alpha_k = \alpha_0 + N_k \tag{25}$$

with

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{26}$$

EM Solution: Maximization Step / Means and Precision

- The distribution of means and precision matrices is updated with

$$q^*(\boldsymbol{\mu}_k, \Lambda_k) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \Lambda_k)^{-1}\right) \mathcal{W}(\Lambda_k | \mathbf{W}_k, \nu_k) \qquad (27)$$

where

$$\beta_k = \beta_0 + N_k \qquad (28)$$

$$\nu_k = \nu_0 + N_k + 1 \qquad (29)$$

$$\mathbf{m}_k = \frac{1}{\beta_k}(\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k) \qquad (30)$$

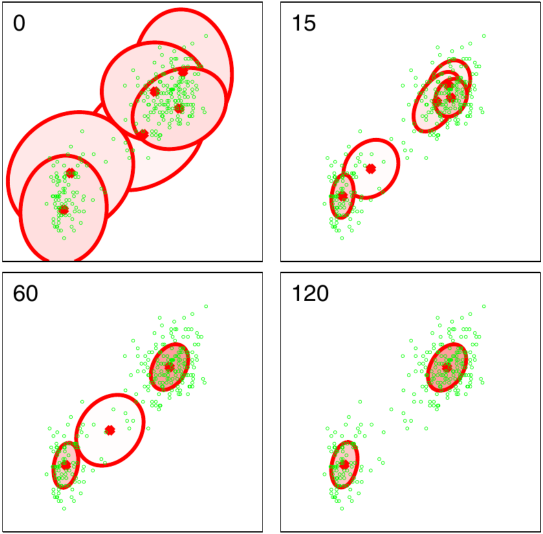
$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \qquad (31)$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^{\mathsf{T}} \qquad (32)$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^{\mathsf{T}} \qquad (33)$$

Example Fit

Figure 10.6 Variational Bayesian mixture of $K = 6$ Gaussians applied to the Old Faithful data set, in which the ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The number in the top left of each diagram shows the number of iterations of variational inference. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.



Remarks

- The priors can be thought of as regularizers that prevent overfitting
- General behavior:
  - Few data points: the priors dominate the result
  - Many data points: the data dominates the result
- Unnecessary components are automatically removed:
  α_k with (almost) zero responsibility approaches the uniform start value α_0
- Hence, it makes sense to start with a larger-than-expected number of clusters, and to remove at the end those clusters with $\alpha_k \approx \alpha_0$

- Markov Chain Monte-Carlo methods are an alternative for Bayesian inference
- One specific example is Gibbs sampling. Here,
  1. Randomly select one sample
  2. Calculate the conditional distributions of the parameters without that sample
  3. Randomly assign the sample to an existing or new cluster with probability of the cluster likelihood
  4. Goto 1. (theoretically until infinity)