CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

L23: Expectation Maximization/Majorize-Minimize/Concave-convex procedure

Goals of this lecture

- Generalizing the kMeans/Gaussian mixture model algorithm
- Getting to know the Concave-convex procedure (CCCP)

Reading material:

- C. Bishop; Pattern Recognition and Machine Learning; Chapter 9.3, 9.4
- Yuille and Rangarajan; Concave Convex Procedure (CCCP); NIPS 2001
- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 11

Recap:

$$\min_{\pi,\mu,\sigma} -\sum_{i\in\mathcal{D}} \ln \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)}_{\sum_{\boldsymbol{z}_i} \underbrace{p(x^{(i)}|\boldsymbol{z}_i)p(\boldsymbol{z}_i)}_{p(x^{(i)},\boldsymbol{z}_i)}} \quad \text{s.t.} \quad \sum_{k=1}^{K}\pi_k = 1, \quad \pi_k \geq 0$$

More generally: (ignoring $\sum_i$)

$$\ln p_\theta(x^{(i)}) = \ln \sum_{\boldsymbol{z}_i} p_\theta(x^{(i)}, \boldsymbol{z}_i)$$

Two options:

- Empirical Lower Bound (ELBO)
- Concave-Convex Procedure/Majorize-Minimize

End up being identical

Empirical Lower Bound:

Goal: maximize likelihood

$$\ln p_\theta(x^{(i)}) = \ln \sum_{\boldsymbol{z}} p_\theta(x^{(i)}, \boldsymbol{z})$$

Let's introduce distribution $q(\boldsymbol{z})$ and rewrite:

$$\ln p_\theta(x^{(i)}) = \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z})) + D_{\mathrm{KL}}(q(\boldsymbol{z}), p_\theta(\boldsymbol{z}|x^{(i)}))$$

where

$$\begin{aligned}
\mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z})) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{p_\theta(x^{(i)}, \boldsymbol{z})}{q(\boldsymbol{z})} \\
D_{\mathrm{KL}}(q(\boldsymbol{z}), p_\theta(\boldsymbol{z}|x^{(i)})) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|x^{(i)})}
\end{aligned}$$

D_{KL}: Kullback-Leibler divergence

Jensen's inequality:

$$f \text{ convex:} \qquad f\left(\sum_{\boldsymbol{z}} q(\boldsymbol{z}) g(\boldsymbol{z})\right) \leq \sum_{\boldsymbol{z}} q(\boldsymbol{z}) f(g(\boldsymbol{z}))$$

$$f \text{ concave:} \qquad f\left(\sum_{\boldsymbol{z}} q(\boldsymbol{z}) g(\boldsymbol{z})\right) \geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}) f(g(\boldsymbol{z}))$$

Consequence for D_{KL}:

$$\begin{aligned} -D_{\text{KL}}(q(\boldsymbol{z}), p_\theta(\boldsymbol{z}|x^{(i)})) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{p_\theta(\boldsymbol{z}|x^{(i)})}{q(\boldsymbol{z})} \\ &\leq 0 \qquad \text{(pull out ln)} \end{aligned}$$

Kullback-Leibler divergence is non-negative

Consequence for log-likelihood:

$$\ln p_\theta(x^{(i)}) = \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z})) + D_{\mathrm{KL}}(q(\boldsymbol{z}), p_\theta(\boldsymbol{z}|x^{(i)}))$$

Lower bound:

$$\ln p_\theta(x^{(i)}) \geq \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z}))$$

Idea: instead of maximizing log-likelihood, let's maximize lower bound:

$$\max_{q,\theta} \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z}))$$

How: alternating optimization w.r.t. q and θ

Alternating optimization:

$$\max_{q,\theta} \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z}))$$

- Maximize w.r.t. q:

$$\implies \quad q(\boldsymbol{z}) = p_\theta(\boldsymbol{z}|x^{(i)})$$

$\ln p_\theta(x^{(i)})$ is upper bound and $\ln p_\theta(x^{(i)}) = \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z}))$ if $q(\boldsymbol{z}) = p_\theta(\boldsymbol{z}|x^{(i)})$ (KL-divergence is zero)

- Maximize w.r.t. θ

Alternative to show that $q(\boldsymbol{z}) = p_\theta(\boldsymbol{z}|x^{(i)})$:

$$\max_q \mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z}))$$

$$\max_q \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p_\theta(x^{(i)}, \boldsymbol{z}) + H(q(\boldsymbol{z})) \quad \text{s.t.} \left\{ \begin{array}{l} q(\boldsymbol{z}) \geq 0 \\ \sum_{\boldsymbol{z}} q(\boldsymbol{z}) = 1 \end{array} \right.$$

How to solve:

Stationarity of Lagrangian

Solution:

$$q(\boldsymbol{z}) = \frac{p_\theta(x^{(i)}, \boldsymbol{z})}{\sum_{\boldsymbol{z}} p_\theta(x^{(i)}, \boldsymbol{z})} = p_\theta(\boldsymbol{z}|x^{(i)}) = r_i$$

In the Gaussian case:

$$\begin{aligned}
\mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z})) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{p_\theta(x^{(i)}, \boldsymbol{z})}{q(\boldsymbol{z})} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{\prod_{k=1}^{K} \pi_k^{z_{ik}} \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)^{z_{ik}}}{q(\boldsymbol{z})} \\
&= \sum_{\boldsymbol{z},k} q(\boldsymbol{z}) \ln \pi_k^{z_{ik}} \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)^{z_{ik}} + H(q(\boldsymbol{z})) \\
&= \sum_{k} r_{ik} \ln \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k) - \sum_{k} r_{ik} \ln r_{ik}
\end{aligned}$$

Why is this easier to optimize than the original program?

In the general case:

$$p_\theta(x^{(i)}, \boldsymbol{z}) = \frac{1}{Z(\theta)} \exp F(x^{(i)}, \boldsymbol{z}, \theta) \qquad Z(\theta)\text{: partition function}$$

$$\begin{aligned} -\mathcal{L}(p_\theta(x^{(i)}, \boldsymbol{z}), q(\boldsymbol{z})) &= -\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{p_\theta(x^{(i)}, \boldsymbol{z})}{q(\boldsymbol{z})} \\ &= \ln Z(\theta) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) F(x^{(i)}, \boldsymbol{z}, \theta) - H(q(\boldsymbol{z})) \end{aligned}$$

Keep that in mind

Concave-convex procedure (CCCP):

Model:

$$p_\theta(x^{(i)}, \boldsymbol{z}) = \frac{1}{Z(\theta)} \exp F(x^{(i)}, \boldsymbol{z}, \theta)$$

Maximum Likelihood (marginalizing over latent space):

$$\min_\theta - \ln \sum_{\boldsymbol{z}} \frac{\exp F(x^{(i)}, \boldsymbol{z}, \theta)}{Z(\theta)}$$

$$\min_\theta \quad \underbrace{\ln Z(\theta)}_{\text{convex if } F \text{ linear in } \theta} \quad \underbrace{- \ln \sum_{\boldsymbol{z}} \exp F(x^{(i)}, \boldsymbol{z}, \theta)}_{\text{convex if } F \text{ linear in } \theta}$$

Concave-convex procedure (CCCP):

- Initialize θ
- Repeat:
  - Decompose concave part into 'convex + concave' at current θ
  - Solve convex program

$$\min_\theta \quad \underbrace{\ln Z(\theta)}_{\text{convex if } F \text{ linear in } \theta} \quad \underbrace{- \ln \sum_{\boldsymbol{z}} \exp F(x^{(i)}, \boldsymbol{z}, \theta)}_{\text{convex if } F \text{ linear in } \theta}$$

How to decompose: (one possibility)

$$\begin{aligned}
\ln \sum_{\boldsymbol{z}} \exp F(x^{(i)}, \boldsymbol{z}, \theta) &= \ln \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{\exp F(x^{(i)}, \boldsymbol{z}, \theta)}{q(\boldsymbol{z})} \qquad \text{(Jensen's)} \\
&= \max_{q(\boldsymbol{z})} \left(\sum_{\boldsymbol{z}} q(\boldsymbol{z}) F(x^{(i)}, \boldsymbol{z}, \theta) + H(q(\boldsymbol{z})) \right)
\end{aligned}$$

Concave-convex procedure (CCCP): Summary

$$\min_{\theta} \quad \underbrace{\ln Z(\theta)}_{\text{convex if } F \text{ linear in } \theta} \quad \underbrace{- \ln \sum_{\boldsymbol{z}} \exp F(x^{(i)}, \boldsymbol{z}, \theta)}_{\text{convex if } F \text{ linear in } \theta}$$

Decomposition:

$$\ln \sum_{\boldsymbol{z}} \exp F(x^{(i)}, \boldsymbol{z}, \theta) = \max_{q(\boldsymbol{z})} \left(\sum_{\boldsymbol{z}} q(\boldsymbol{z}) F(x^{(i)}, \boldsymbol{z}, \theta) + H(q(\boldsymbol{z})) \right)$$

Results in:

$$\min_{\theta, q} \ln Z(\theta) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) F(x^{(i)}, \boldsymbol{z}, \theta) - H(q(\boldsymbol{z}))$$

Quiz:

- Jensen's inequality?
- General idea of CCCP?
- Variational form of the partition function?

Important topics of this lecture

- Generalizing EM
- Getting to know its relationship with CCCP
- Seeing the variational form of the partition function
- Observing its similarity to inference

What's next

- Practicing those concepts