CPSC 486/586: Probabilistic Machine Learning 2/8/2023

Lecture 8

Lecturer: Andre Wibisono *Scribe: Jacob Harrison*

Background: Variational Inference vs Expectation Propagation

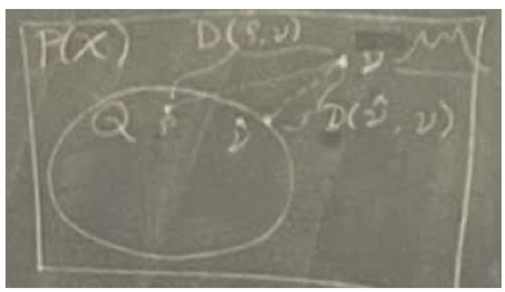
Let:

$$\mathcal{X} = \text{state space}$$
$$P(\mathcal{X}) = \text{space of probability distributions } \rho \text{ over } \mathcal{X}$$
$$Q(\subseteq P(\mathcal{X})) = \text{space of nice (computable) distributions}$$

Given a target $\nu \in P(\mathcal{X})$, find:

$$\hat{\nu} = \arg\min_{\rho \in Q} D(\rho, \nu)$$

Which can be visually summarized as:



Today, we are going to consider two approaches to calculating $D(\rho, \nu)$, including:

1. $D(\rho, \nu) = KL(\rho||\nu) = -H(\rho) + \mathbb{E}_\rho[-\log \nu]$
2. $D(\rho, \nu) = KL(\nu||\rho) = -H(\nu) + \mathbb{E}_\nu[-\log \rho]$

Here, our first and second approaches respectively are **Variational Inference** and **Expectation Propagation**, which will be explained in more detail later in this lecture.

Examples of nice Q Distribution Spaces

In order to effectively and efficiently carry out inference, it is important to choose a nice $Q \subseteq P(\mathcal{X})$ distribution space for $\mathcal{X} = \mathbb{R}^d$. Here, we will discuss commonly utilized Q spaces.

1. Point Mass Distribution Space:
$$Q = \{\delta_x : x \in X\}$$

2. Gaussian Distribution Space:
$$Q = \{\mathcal{N}(m, C) : m \in \mathbb{R}^d, C > 0 \in \mathbb{R}^{dxd}\}$$

3. Exponential Family Distribution Space:
$$Q = \{q_\Theta(x) = exp(\langle \theta, T(x) \rangle - A(\theta)) : \theta \in \Theta\}$$

4. Mixed Gaussian Distribution Space:
$$Q = \left\{ \sum_{i=1}^{n} p_i \mathcal{N}(m_i, C_i) : p \in \Delta_n, m_i \in \mathbb{R}^d, C_i \in \mathbb{R}^{dxd} \right\}$$
(Note: Δ_n is the simplex in n-dimensions: $\Delta_n = \{p \in \mathbb{R}^n \;:\; p_i \geq 0, \forall i \in [n], \sum_{i=1}^{n} p_i = 1\}$.)

5. Mean-field Distribution space (where each q_i is a probability distribution):
$$Q = \left\{ \prod_{i=1}^{d} q_i(x_i) : q_i \in P(\mathbb{R}) \right\}$$

6. Two-layer Neural Network Distribution Space (with non-linear σ function and linear $Az + b$ function).
$$Q = \{\varnothing_2(\varnothing_1(z)), z \sim \mathcal{N}(0, I) : \varnothing_1(z) = \sigma(A_1 z + b_1), \varnothing_2(z) = \sigma(A_2 z + b_2)\}$$

7. Reparametrization Trick (allows putting all complexity in F with just simple error term z, and therefore can get any distribution $\nu \in P(X)$)
$$Q = \{F(z), z \sim \mathcal{N}(0, I) | F : \mathbb{R}^d \to \mathbb{R}^d\}$$

Expectation Propagation

In expectation propagation, we find the following:
$$\hat{\nu} = \arg\min_{\rho \in Q} \{KL(\nu || \rho) = -H(\nu) - \mathbb{E}_\nu[\log \rho]\}$$

Since, $H(\nu)$ is not dependent on ρ, can simplify to:
$$\arg\min_{\rho \in Q} - \mathbb{E}_\nu[\log \rho]$$

$$\arg\max_{\rho \in Q} \mathbb{E}_\nu[\log \rho]$$

This is the MLE problem. Recall: Given $X_1, ..., X_n \sim \nu$ iid on $\mathbb{R}^d$, we can estimate $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_x$, which has the property that $\lim_{n\to\infty} \hat{\nu}_n = \nu$. Additionally, let $Q = \{q_\theta : \theta \in \Theta\}$.
Thus, our MLE becomes:

$$\begin{aligned}
&\arg\max_{\theta\in\Theta} q_\theta(x_1, ..., x_n) \\
&= \arg\max_{\theta\in\Theta} \prod_{i=1}^n q_\theta(x_i) \\
&= \arg\max_{\theta\in\Theta} exp(\sum_{i=1}^n \log q_\theta(x_i)) \\
&= \arg\max_{\theta\in\Theta} \frac{1}{n}(\sum_{i=1}^n \log q_\theta(x_i)) \\
&= \mathbb{E}_{\hat{\nu}_n}[\log q_\theta]
\end{aligned}$$

which as $n \to \infty$ approaches:

$$\mathbb{E}_\nu[\log q_\theta]$$

Lemma 1. *If $Q = \{q_\theta(x) = \exp(\langle\theta, T(x)\rangle - A(\theta)) : \theta \in \Theta\}$, then:*

$$q_{\theta^*} = \arg\min_{q_\theta \in Q} KL(\nu || q_\theta)$$

is such that:

$$\mathbb{E}_{q_{\theta^*}}[T(X)] = \mathbb{E}_\nu[T(X)]$$

This is referred to as moment matching and can be proved as follows:

$$\begin{aligned}
\theta^* &= \arg\max_{\theta\in\Theta} \mathbb{E}_\nu[\log q_\theta] \\
&= \arg\max_{\theta\in\Theta} \mathbb{E}_\nu[\langle\theta, T(x)\rangle] - A(\theta) \\
&= \arg\min_{\theta\in\Theta}(-\langle\theta, \mathbb{E}_\nu[T(x)]\rangle + A(\theta))
\end{aligned}$$

and if $-\langle\theta, \mathbb{E}_\nu[T(x)]\rangle + A(\theta) = F(\theta)$:

$$\begin{aligned}
\nabla F(\theta^*) = 0 &\iff \nabla A(\theta^*) - \mathbb{E}_\nu[T(X)] = 0 \\
&\iff \nabla A(\theta^*) = \mathbb{E}_\nu[T(X)] \\
&\iff \mathbb{E}_{q_{\theta^*}}[T(X)] = \mathbb{E}_\nu[T(X)]
\end{aligned}$$

Gaussian Example of Expectation Propagation

Let $Q = \{\rho = \mathcal{N}(m, C) : m \in \mathbb{R}^d, C > 0 \in \mathbb{R}^{dxd}\}$.

$$
\begin{aligned}
F(\rho) = F(m, C) = KL(\nu || \rho) &= -H(\nu) - \mathbb{E}_\nu[\log \rho] \\
&= \mathbb{E}_\nu \left[\frac{1}{2} ||x - m||^2_{C^{-1}} + \frac{1}{2} \log \det(2\pi C) \right] - H(\nu) \\
&= \frac{1}{2} || \mathbb{E}_\nu[X] - m||^2_{C^{-1}} + \frac{1}{2} \mathsf{Tr}(\mathsf{Cov}_\nu(X) C^{-1}) + \frac{1}{2} \log \det(C) + const
\end{aligned}
$$

Now, plugging in optimal $m^* = \mathbb{E}_\nu[X]$, will show that $C^* = C_\nu$

$$
F(C) = F(m^*, C) = \frac{1}{2} \mathsf{Tr}(\mathsf{Cov}_\nu(X) C^{-1}) + \frac{1}{2} \log \det(C) + const
$$

For $d = 1$: $C > 0$,

$$
F(C) = \frac{1}{2} \frac{C_\nu}{C} + \frac{1}{2} \log C
$$

Setting $\lambda = \frac{1}{C}$:

$$
F(\lambda) = \frac{1}{2} C_\nu \lambda - \frac{1}{2} \log \lambda
$$

Minimizing F, we find

$$
F'(\lambda^*) = \frac{1}{2} C_\nu - \frac{1}{2\lambda} = 0 \iff \lambda = \frac{1}{C_\nu} \iff C^* = C_\nu
$$

Now, for $d \geq 1$, set $\Lambda = C^{-1}$, and minimize:

$$
F(\Lambda) = \frac{1}{2} \mathsf{Tr}(C_\nu \Lambda) - \frac{1}{2} \log \det \Lambda
$$

Lemma:

$$
\nabla_\Lambda \mathsf{Tr}(C_\nu \Lambda) = C_\nu
$$

$$
\nabla_\Lambda \log \det \Lambda = \Lambda^{-1}
$$

Therefore:

$$
\nabla F(\Lambda^*) = \frac{1}{2} C_\nu - \frac{1}{2} (\Lambda^*)^{-1} = 0 \iff \Lambda^* = C_\nu^{-1} \iff C^* = C_\nu
$$