CPSC 486/586: Probabilistic Machine Learning 02/01/2023

Lecture 6

Lecturer: Andre Wibisono *Scribe: Argyris Giannisis Manes*

Approximating Probability Distributions

Suppose that we wish to approximate a distribution $\nu(\theta)$. This can be done in a variety of ways. For instance, we can compute some statistics, like the mean $\mu = \mathbb{E}_\nu(\theta)$, covariance $\Sigma = \mathsf{Cov}_\nu(\theta) = \mathbb{E}_\nu[(\theta-\mu)(\theta-\mu)^T]$, or mode $\theta^* = \arg\max_{\theta\in\Theta} \nu(\theta)$. Or we could collect samples $\theta_i \sim \nu$, $i = 1, 2, \ldots, n$, and then approximate ν using delta functions, i.e, $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$. Note that for $\mathcal{N}(\mu, \Sigma)$, the mean of the distribution is also the mode. But for a skewed distribution, mean and mode can be different.

Example: Approximate Posterior in Logistic Regression

We have some real parameter θ. Given some covariate $x \in \mathbb{R}$, then we observe label $y \in \{0, 1\}$ with $y|x, \theta \sim \mathrm{Ber}(\sigma(x\theta)) = \mathrm{Ber}(\frac{1}{1+e^{-x\theta}})$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. For a Bernoulli distribution, we can write

$$p(y|x,\theta) = \sigma(x\theta)^y(1-\sigma(x\theta))^{1-y} = \Big(\frac{1}{1+\exp\{-x\theta\}}\Big)^y\Big(\frac{e^{-x\theta}}{1+e^{-x\theta}}\Big)^{1-y}$$
$$= e^{x\theta y}\cdot\frac{e^{-x\theta}}{1+e^{-x\theta}} = \frac{e^{x\theta y}}{1+e^{x\theta}} = \exp\Big(yx\theta - \log\Big(1+e^{x\theta}\Big)\Big)$$

So $p(y|x,\theta)$ is an exponential family with $T(y) = y$, $w = x\theta$, $A(w) = \log(1+e^w)$.

Now, suppose that our prior distribution is $p_0 = \mathcal{N}(0, 1)$, and we observe the data points $(x_1, y_1), \ldots, (x_n, y_n)$. We want to calculate our posterior distribution $p_n(\theta) = p(\theta|x_1, y_1, \ldots, x_n, y_n)$. We have that from Baye's Rule,

$$p_n(\theta) \propto p_0(\theta)\prod_{i=1}^n p(y_i|\theta, x_i) \implies$$
$$p_n(\theta) \propto \exp\Bigg\{-\frac{\theta^2}{2} - \sum_{i=1}^n\Big(\log\Big(1+e^{\theta x_i}\Big) - y_i x_i\theta\Big)\Bigg\}$$
$$\propto \exp\Bigg\{-\frac{\theta^2}{2} + \sum_i x_i y_i\theta - \sum_i \log\Big(1+e^{\theta x_i}\Big)\Bigg\}$$

$$\propto \exp\left\{-\frac{1}{2}(\theta - \sum_i^n x_i y_i)^2 - \sum_i \log\left(1 + e^{\theta x}\right)\right\}$$

This is of the form $\exp(-f_n(\theta))$, where f_n is a convex function (it is going to look like a skewed quadratic). We check that $A(w)$ is convex:

$$A'(w) = \frac{e^w}{1+e^w} = p(w)$$

$$A''(w) = \frac{e^w}{1+e^w} - \frac{e^{2w}}{(1+e^w)^2} = p(w) - p^2(w) = p(1-p)$$

Since $p \in (0,1)$, we see that $A''(w) \geq 0$.

So for $A(w) = \log(1+e^w)$, we have $A'(w) = \mathbb{E}(Y)$, with Y drawn from the Bernulli distribution $Y \sim \text{Ber}(\sigma(w))$, and $A''(w) = \mathsf{Var}(Y)$.

This is true for any exponential family. Additionally, a distribution of the form $p_n \propto e^{-f_n}$, with f_n concave, is called log-concave.

Approximate Distribution $\nu \sim e^{-f(\theta)}$

We look at two ways to approximate a distribution of this form.

Approximate by a single point

We want to approximate ν with a single point, i.e, $\nu \sim \delta_{\theta^*}$. For example, we could consider $\theta^* = \arg\max_\theta \nu(\theta) = \arg\min_\theta f(\theta)$. This is a MAP Estimator, and in the case of $\nu = p_n(\theta)$ we have

$$\theta^*_{MAP} = \arg\max_\theta p_n(\theta)$$

Question: If I want to approximate by a point mass, what $\hat{\theta}$ gives best constant approximation to ν? The answer is given by

$$\hat{\theta} = \arg\min_\theta \mathbb{E}(|\theta - \hat{\theta}|^2) = \mathbb{E}(\theta)$$

Approximate with Gaussian: Laplace Approximation

Suppose that instead we want to approximate using a Normal Distribution. We can use the MAP estimate $\theta^* = \arg\max_\theta \nu(\theta)$ as our mean. But what should we use for the variance? It's hard to compute the variance of the posterior, or other quantities like the Fisher Information. This leasds us to the **Laplace Approximation**.

Note that our θ^* is such that $\theta^* = \arg\min_\theta f(\theta)$, where $\nu \sim e^{-f(\theta)}$. We can approximate f with its expansion

$$f(\theta) = f(\theta^*) + \langle \nabla f(\theta^*), \theta - \theta^* \rangle + \frac{1}{2}(\theta - \theta^*)^T \nabla^2 f(\theta^*)(\theta - \theta^*)$$

but the gradient is zero at minimizer θ^*, so

$$f(\theta) \sim \frac{1}{2}||\theta - \theta^*||^2_{\nabla^2 f(\theta^*)}$$

And since we have that $\nu \sim e^{-f}$, we conclude

$$\hat{\nu} \sim \mathcal{N}(\theta^*, (\nabla^2 f(\theta^*))^{-1})$$

which is the Laplace Approximation.

Example: Laplace for Logistic Regression

We apply this to the logistic regression example from earlier. We recall that

$$\nu(\theta) = p_n(\theta | x_1, y_1, \ldots, x_n, y_n) = \exp(-f_n(\theta))$$

$$f_n(\theta) = \frac{1}{2}(\theta - \sum_i x_i y_i)^2 + \sum_i \log\left(1 + e^{x_i \theta}\right)$$

We now use the MAP Estimate. We have that

$$\theta^* = \arg\min_\theta f_n(\theta) \implies \nabla f_n(\theta^*) = 0$$

$$\nabla f_n(\theta) = \theta - \sum_i x_i y_i + \sum_i \frac{e^{x_i\theta}}{1 + e^{x_i\theta}} x_i = \theta - \sum_i x_i (y_i - \frac{e^{x_i\theta}}{1 + e^{x_i\theta}}) = 0$$

But this is a hard fixed point equation to solve. We can optimize via algorithms like gradient descent etc.

Once you have θ^*, you can compute the Hessian. We have (in one dimension)

$$\nabla^2 f_n(\theta) = f''(\theta) = 1 + \sum_i \left(\frac{x_i e^{x_i\theta}}{1 + e^{x_i\theta}} - \frac{x_i e^{2x_i\theta}}{(1 + e^{x_i\theta})^2} \right) x_i = 1 + \sum_{i=1}^n \frac{e^{x_i\theta}}{(1 + e^{x_i\theta})^2} x_i^2$$

Question: How good is this? Is there a way to bound quantities like

$$||\mathbb{E}_{\hat{\nu}(\theta)} - \mathbb{E}_\nu(\theta)||, \ ||\mathsf{Cov}_{\hat{\nu}}(\theta) - \mathsf{Cov}_\nu(\theta)||$$

Theorem 1. *It can be shown that if $n \geq d^3$ (where d is the dimension of θ) and ν differentiable, we have that*

$$\sqrt{n}||\mathbb{E}_{\hat{\nu}}(\theta) - \mathbb{E}_{\nu}(\theta)|| \leq \sqrt{\frac{d^3}{n}}$$

$$n||\mathsf{Cov}_{\hat{\nu}}(\theta) - \mathsf{Cov}_{\nu}(\theta)|| \leq \frac{C(d)}{n}$$

for some constant C that only depends on d.