6.882: Bayesian Modeling and Inference **Lecture Date: February 6, 2018**

Introduction to Bayesian Methods

Lecturer: Prof. Tamara Broderick *Scribe: Menghua Wu, Jessy Lin*

1 Administrivia

- Office hours are from 4-4:30p after lecture.
- We can email `6882staff2018@gmail.com`. DO NOT email the professor at any other email address, especially for assignments.
- This is an advanced graduate class, so we will digest papers together.
- Class participation counts. Each class will have pairs presenting/leading discussion and pairs scribing. Prof. Broderick is open to discussion with each week's presenters beforehand.
- We also complete a project in some aspect related to the class, in groups of 1-3 (unless there are compelling reasons).
- There are no exams!
- There are weekly reflections about the papers, half a page due the night before.

There is a reading due Thursday, but no reflection required. We are reading the "classic" paper, Blei et al. (2003). Concurrently, Pritchard et al. (2000) described this model in biological situations. Both have over 20,000 citations, so these methods are very popular.[1]

2 Bayesian methodology

Now let's dive straight in.

Theorem 1 (Bayes Theorem)**.**

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

where X, Θ are random variables and x, θ are their realizations.

Bayesian methodology interprets:

- x as observations (data set),
- θ as latent parameters (unknown quantity).
- $p(x \mid \theta)$ as the likelihood of observing x, given θ,

[1] Read sections 10.1 and 10.2 of Bishop (2006) if you're not familiar with variational Bayesian inference.

- $p(\theta)$ as the prior, and
- $p(\theta \mid x)$ as the posterior.

We often simplify Bayes Theorem to

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta)$$

where we omit the normalizing factor $p(x)$.

2.1 Benefits of Bayesian inference

Bayesian inference is widely used in the latest scientific discoveries and engineering feats. Here are several reasons that these methods are so popular.

1. It provides a unified framework for many of the tasks we care about in machine learning and statistics.
   - We can provide **point estimates**. For example, via the mean/median/mode of the posterior distribution. What is the expected profit of some investment?
   - We can make **predictions**, e.g. using the *posterior predictive distribution*, $p(x_{\text{new}} \mid x_{\text{old}})$.
   - We can quantify **uncertainty**.
   - We can **compare models**. $p(x)$ is the *model evidence*.
2. We can incorporate prior information.
3. Bayesian models are modular. We study modularity in the context of graphical models.
4. These methods are flexible in other ways. Later we will study nonparametric Bayesian methods.
5. Bayesian models are robust to overfitting. The prior acts as a penalty.
6. These models can lend themselves to interpretability. This property is important because some machine learning models are "black boxes." Researchers can reason about the Bayesian models for themselves, as well as convince the public that these models are trustworthy and safe.
7. Bayesian estimators have useful frequentist properties.

2.2 Challenges of Bayesian inference

1. Nature doesn't provide a model. We must choose $p(x \mid \theta)$. All machine learning models make some assumptions, and we would like to know how these assumptions affect our analysis.

   > "All models are wrong, but some are useful."—George Box

2. We also choose the prior $p(\theta)$. There are several schools of thought here. Objective Bayes focuses on quantifying the lack of knowledge, while subjective Bayes focuses on information we do know.
3. We need to calculate the posterior $p(\theta \mid x)$ using Bayes rule. Often, the normalizing constant is hard to calculate in higher dimensions, so there are closed form distributions only in a few select cases. Generally, we must approximate the posterior, but these approximation techniques may be slow.
4. A challenge is user time, which can include derivation time and time required to tune hyperparameters. Ideally, we would want posterior calculation to be automatic and reliable with black box inference.[2]

[2]Black box inference is different from black box machine learning. The former only assumes that calculating the posterior is a robust black box.

5. We need to check the analysis or evaluation.
    - Is the approximation accurate or reliable?
    - Can we trust the result or model? Are there a priori guarantees or a posteriori checks?
    - Is the model robust—that is, if we change assumptions or data, how much does the model change?

2.3 De Finetti's Theorem

Suppose we have a data sequence $X_1, X_2, \ldots, X_n, \ldots, X_N$.

Definition 2. $X_1, X_2, \ldots, X_N$ is **N-exchangeable** if for any permutation σ of the first N natural numbers $[N] := \{1, \ldots, N\}$,

$$(X_1, X_2, \ldots, X_N) \stackrel{d}{=} (X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(N)}).$$

That is, any two permutations of $X_1, X_2, \ldots, X_N$ have the same joint distribution.

Remark. Any sequence of N i.i.d. random variables is an N-exchangeable sequence, but the converse is not true.

Example. Suppose that we have a sequence $X_1, X_2, \ldots, X_N$ and some nontrivial $X_0 \perp\!\!\!\perp X_1, X_2, \ldots, X_N$. Then $(X_0 + X_n)_n$ is N-exchangeable but not i.i.d (because of X_0).

Example. Consider the "bag of words" assumption in document analysis. We might treat each word as exchangeable (order doesn't matter), but not independent (words in the same document tend to be semantically similar). Both assumptions are wrong, but exchangeability is a bit more realistic.

Definition 3. Let $X_1, X_2, \ldots$ be an infinite data generator. $X_1, X_2, \ldots$ is **infinitely exchangeable** if for any N and permutation σ of $[N]$, $(X_1, X_2, \ldots) \stackrel{d}{=} (X_{\sigma(1)}, X_{\sigma(2)}, \ldots)$.

That is, every finite subsequence is exchangeable. This leads up to de Finetti's theorem.[3]

Theorem 4 (de Finetti's theorem, roughly)**.** *$X_1, X_2, \ldots$ are infinitely exchangeable if and only if for all N and some distribution P,*

$$p(X_1, \ldots, X_N) = \int_\theta \left[\prod_{n=1}^{N} p(X_n \mid \theta) \right] P(d\theta)$$

That is, if a sequence is infinitely exchangeable, then there must exist some parameter θ and distribution P on θ, such that all data is i.i.d. given θ (Figure 1).

De Finetti's theorem motivates several concepts.

- The use of parameters.
- Likelihoods: distributions for the data conditioned on parameters.
- Priors, or distributions over parmeters.
- We do not limit the dimensionality of the parameter, which motivates nonparametric Bayesian methods.

[3] For a more general theorem, see Hewitt and Savage (1955).

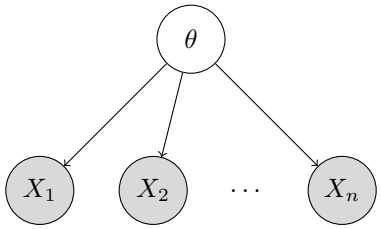

Figure 1: Graphical model representation of de Finetti's theorem.

At this point, we realize that Bayesian inference requires three components: a model, a method of inference, and a method of evaluation.

Today we introduce inference. We want to approximate the posterior since there does not always exist a closed form solution for $p(\theta \mid x)$. We're mainly interested in calculating useful functionals of $p(\theta \mid x)$, such as the mean, variance, or covariance.

One approach is to approximate $q^*(\theta) \approx p(\theta \mid x)$ and say that

$$\mathrm{E}_{q^*(\theta)}\left[\theta\right] \approx \mathrm{E}_{p(\theta|x)}\left[\theta\right]. \tag{1}$$

Specific methods include

- Markov chain Monte Carlo,
- Laplace approximation, and
- variational Bayes.

We start with variational Bayes. Suppose our exact posterior $p(\theta \mid x)$ is gnarly, but we do have "nice" distributions (can compute functionals). The idea is to take the closest "nice" distribution, where we will formalize "close" and "nice" (Figure 2).

Let

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot \mid x)), \tag{2}$$

where $p(\cdot|x)$ is a normalized function of θ. Here, f should be positive definite. That is, $f(q_1, q_2) \geq 0$, where $f(q_1, q_2) = 0$ if and only if $q_1 = q_2$.

In variational Bayes, we choose f to be Kullback-Leibler divergence in a particular direction:

$$f(q_1, q_2) = \mathrm{KL}\left(q_1 || q_2\right) = \int q_1(\theta) \log \frac{q_1(\theta)}{q_2(\theta)} d\theta. \tag{3}$$

Note that Kullback-Leibler divergence is *not* symmetric! It also does not satisfy the triangle inequality, so this is *not* a metric; it is a divergence. However, it does satisfy our desired properties, and it *is* positive definite.

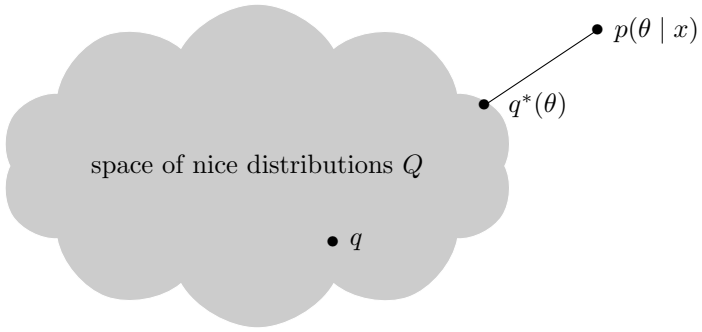

Figure 2: Intuitive representation of variational Bayes. We find q^*, which is the closest "nice" distribution from our desired p, where closeness is measured by KL divergence.

Substituting KL divergence into equation 2,

$$\begin{aligned}
q^* &= \arg\min_{q \in Q} \mathrm{KL}\left(q(\cdot) || \, p(\cdot \mid x) \right) \\
&= \arg\min_{q \in Q} \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta \\
&= \arg\min_{q \in Q} \int q(\theta) \log \frac{q(\theta) p(x)}{p(x \mid \theta) p(\theta)} d\theta \quad \text{by Bayes theorem} \\
&= \arg\min_{q \in Q} \int q(\theta) \log p(x) d\theta + \int q(\theta) \log \frac{q(\theta)}{p(x \mid \theta) p(\theta)} d\theta \\
&= \arg\min_{q \in Q} \log p(x) - \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta
\end{aligned}$$

The $\log p(x)$ factor is constant across all q's, so we may ignore it. The second term is known as the "evidence lower bound" (ELBO), and we see that

$$q^* = \arg\max_{q \in Q} ELBO(q)$$

We will continue with ELBO next lecture. Review sections 10.1 and 10.2 from Bishop (2006) for more details on variational inference and the derivations we did in class today.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, pages 945–959.