November 8, 2017 Lecture 18, Nov. 6, 2017

Bayesian Inference

Lecturer: Justin Domke *Scribe: Lakshmi Vikraman*

1 Summary

Last class we finished the discussion on Kernel methods. This class we will talk about Bayesian Inference. The readings for this topic are up on Moodle.

2 Bayesian Inference

Philosophically, Bayesian Inference is very different from other machine learning techniques covered so far. Though it is mathematically simple, it is difficult to grasp conceptually. We will go through a series of examples to gain a good understanding of the fundamental concepts.

2.1 Example

Let us start with an example. We have two monkeys(M), Alfred(a) and Betty(b). They live in a parallel universe. In addition there are two kinds of blocks (B), green(g) and yellow(y). Alfred likes the green block and Betty likes the yellow block. There is a wizard who picks one of the monkeys and gives them the power to send a block to our universe. He rolls a four sided die and picks Alfred if the outcome is 1 and Betty if the outcome is 2-4. Both monkeys will send their favorite colored block with probability 0.8. Intuitively, based on this model we will see more yellow colored blocks.

Question 1: If we get a green block, what is the probability that the magical monkey is Alfred?
Here we have two pieces of information. The wizard picks Betty most of the time and she prefers yellow blocks but we recieved a green block which is Alfred's favorite. We will use Bayes' equation to weigh these two pieces to arrive at a solution. More formally, we can specify this model as follows.

$$
\begin{aligned}
&P(M = a) = 0.25 \\
&P(M = b) = 0.75 \\
&P(B = g|M = a) = 0.8 \\
&P(B = y|M = a) = 0.2 \\
&P(B = g|M = b) = 0.2 \\
&P(B = y|M = b) = 0.8
\end{aligned}
$$

We want to calculate the following:

$$\begin{aligned} P(M = a|B = g) &= \frac{P(M = a)P(B = g|M = a)}{P(B = g)} \\ &= \frac{0.25 \times 0.8}{P(B = g)} \end{aligned}$$

$$\begin{aligned} P(M = b|B = g) &= \frac{P(M = b)P(B = g|M = b)}{P(B = b)} \\ &= \frac{0.75 \times 0.2}{P(B = g)} \end{aligned}$$

The normalization constant in the denominator can be expanded as shown below.

$$\begin{aligned} P(M = a|B = g) &= \frac{0.25 \times 0.8}{0.25 \times 0.8 + 0.75 \times 0.2} \\ &= \frac{4}{7} \end{aligned}$$

We get a value greater than $\frac{1}{2}$. This is because of the relative sizes of the prior and the likelihood containing information about the blocks.

Question 2: What is the probability that the next block is also green?
Initially when we started out, we only knew the prior, but after seeing a block, the posterior probability has been updated with this information.

$$\begin{aligned} P(B' = g|B = g) &= P(M = a|B = g)P(B' = g|M = a) + P(M = b|B = g)P(B' = g|M = b) \\ &= \frac{4}{7} \times \frac{8}{10} + \frac{3}{7} \times \frac{2}{10} \\ &= \frac{19}{35} \end{aligned}$$

When we get more and more information we can infer a distribution over the monkeys.

Lets us summarize what we did so far.

- Stated a "prior" over the different models. Here the prior is set based on the whims of the wizard.

$$\begin{aligned} P(M = a) &= \frac{1}{4} \\ P(M = b) &= \frac{3}{4} \end{aligned}$$

- Stated a "likelihood" over seeing the data if the model were true.

$$\text{Eg}: \ P(B = g|M = a) = 0.8$$

  We treat models like we treat data i.e. a probabilistic distribution over the models.

- Calculate "posterior" over models given the data using Bayes' rule.

$$\text{Eg}: \ P(M = a|B = g) = \frac{4}{7}$$

- Use posterior to make predictions on future data. Unlike the earlier machine learning algorithms, where we pick a model and use it to make the predictions, here we never pick a model. To reiterate **WE NEVER CHOOSE ONE MODEL !!!**

What do we need to do to scale this up to real world problems?

- We need more complex priors over more complex models. For eg, in a linear model, we need a prior over β. Historically, this has led to intense debates in the scientific community, since priors tend to be biased.

- We need more complex models/likelihoods. Instead of having just two monkeys with two possible outcomes, we can have models with a vector β and infinite number of outcomes.

- We have way more computational problems. For instance, it is much harder to construct integrals of the parameters β and integrate out the possible βs instead of picking one. These methods are typically much slower than the alternate methods seen so far.

Another point to note is that we have to specify the model correctly with very strong assumptions. Once the model is specified, we have a relatively simple mathematical framework to work with. This deviates away from the "user-specified" problems which we discussed earlier in the course.

2.2 Comparison with the earlier methods

This is a radical departure from the previously defined models where we did Empirical Risk Minimization. With ERM we do the following:

- Pick a set of predictors F.
- Pick a loss function L.
- Pick a regularizer.
- Minimize average loss on the dataset.

Comparison between the two:

- Given enough data, we always find the optimal solution using ERM. But in case of Bayesian Inference, if the model error is high, it performs much worse.
- Bayesian Inference performs well on problems where we have good domain knowledge which can be used to define good models.
- Bayesian Inference works best for "big models" while ERM works best for "big data".

2.3 Pros and Cons of Bayesian Inference

Cons

- We need much stronger assumptions.
- It is much more computationally expensive.

Pros

- We get the "optimal" predictor assuming that the assumptions are true.

2.4 Example continued

Let us expand the earlier example. Instead of observing just one block, we are going to observe a sequence of blocks across multiple days. Let B be the sequence of observed blocks and $B = ggy$.

$$\begin{aligned} P(M = a|B = ggy) &= \frac{1}{P(B = ggy)} P(M = a) p(B = ggy|M = a) \\ &= \frac{1}{P(B = ggy)} \times 0.25 \times 0.8 \times 0.8 \times 0.2 \end{aligned}$$

$$P(M = b|B = ggy) = \frac{1}{P(B = ggy)} \times 0.75 \times 0.2 \times 0.2 \times 0.8$$

$$\begin{aligned} P(M = a|B = ggy) &= \frac{0.25 \times 0.8 \times 0.8 \times 0.2}{0.25 \times 0.8 \times 0.8 \times 0.2 + 0.75 \times 0.2 \times 0.2 \times 0.8} \\ &= \frac{1}{13} \end{aligned}$$

We will continue with this example in the next class.