November 19, 2017 Lecture 19, Nov. 8, 2017

Bayesian Inference

Lecturer: Justin Domke *Scribe: Annamalai Natarajan*

1 Summary

Last time we discussed the problem of inferring which whimsical monkey was throwing blocks given a single block. In this lecture, we use Bayesian inference to assign probabilities to whimsical monkeys given a sequence of blocks and other factors such as weather.

2 Act II

Consider the sequence of blocks 'g,y,y' (green, yellow, yellow) and probabilities as given by,

$$P(M = a) = 0.25 \tag{2.1}$$
$$P(M = b) = 0.75 \tag{2.2}$$
$$P(B = g|M = a) = 0.8 \tag{2.3}$$
$$P(B = y|M = a) = 0.2 \tag{2.4}$$
$$P(B = g|M = b) = 0.2 \tag{2.5}$$
$$P(B = y|M = b) = 0.8 \tag{2.6}$$

where, a is Alfred, b is Betty, names of the two monkeys, and there are two possible colors of blocks, g for green and y for yellow. Given a sequence of blocks 'g,y,y' we would like to infer which monkey most likely threw those blocks. First, we compute the probability of Alfred as,

$$P(M = a|B = g, y, y) = \frac{1}{P(B = g, y, y)} \times P(M = a) \times P(B = g, y, y|M = a) \tag{2.7}$$
$$= \frac{1}{P(B = g, y, y)} \times P(M = a) \times P(B = g|M = a) \times P(B = y|M = a) \times P(B = y|M = a) \tag{2.8}$$
$$= \frac{1}{P(B = g, y, y)} \times 0.25 \times 0.8 \times 0.2 \times 0.2 \tag{2.9}$$

Next, we compute the probability of Betty as,

$$P(M = b|B = g, y, y) = \frac{1}{P(B = g, y, y)} \times P(M = b) \times P(B = g, y, y|M = b) \tag{2.10}$$
$$= \frac{1}{P(B = g, y, y)} \times 0.75 \times 0.2 \times 0.8 \times 0.8 \tag{2.11}$$

There is an unknown, $\frac{1}{P(B=g,y,y)}$, in probabilities of Alfred and Betty but we do know that these two probabilities should sum to 1. Hence we divide each of those probabilities by the sum as,

$$P(M=a|B=g,y,y) = \frac{P(M=a|B=g,y,y)}{P(M=a|B=g,y,y)+P(M=b|B=g,y,y)} \tag{2.12}$$

$$= \frac{0.25 \times 0.8 \times 0.2 \times 0.2}{0.25 \times 0.8 \times 0.2 \times 0.2 + 0.75 \times 0.2 \times 0.8 \times 0.8} \tag{2.13}$$

$$= \frac{1}{13} \tag{2.14}$$

$$P(M=b|B=g,y,y) = \frac{12}{13} \tag{2.15}$$

By dividing by the sum the unknown in numerator and denominator cancel out. Given this sequence of blocks 'g,y,y' it is most probable that Betty was throwing the blocks.

3 Act III

Consider the sequence of blocks 'g,y,y' (green, yellow, yellow) and weather 'r,r,c' (rain, cloudy, cloudy) and probabilities as given by,

$$P(B=g|M=a,W=c) = 0.8 \tag{3.1}$$

$$P(B=y|M=a,W=r) = 0.2 \tag{3.2}$$

$$P(B=g|M=b,W=c) = 0.2 \tag{3.3}$$

$$P(B=y|M=b,W=r) = 0.8 \tag{3.4}$$

where, W is a random variable representing weather and can take on two values r for rain and c for cloudy. In this act, we assume that weather is independent of monkeys *i.e.* $W \perp\!\!\!\perp M$. Again, we are interested in computing the probability of Alfred given additional evidence on weather as,

$$P(M=a|B=g,y,y,W=r,r,c) = \frac{P(B=g,y,y,W=r,r,c|M=a) \times P(M=a)}{P(B=g,y,y,W=r,r,c)} \tag{3.5}$$

$$= \frac{P(B=g,y,y|W=r,r,c,M=a) \times P(W=r,r,c|M=a) \times P(M=a)}{P(B=g,y,y,W=r,r,c)} \tag{3.6}$$

$$= \frac{P(B=g,y,y|W=r,r,c,M=a) \times P(W=r,r,c) \times P(M=a)}{P(B=g,y,y,W=r,r,c)} \tag{3.7}$$

The $P(W=r,r,c|M=a)$ is just $P(W=r,r,c)$ due to the independence between monkeys and weather. We compute the probability of Betty given additional evidence on weather as,

$$P(M=b|B=g,y,y,W=r,r,c) = \frac{P(B=g,y,y,W=r,r,c|M=b) \times P(M=b)}{P(B=g,y,y,W=r,r,c)} \tag{3.8}$$

$$= \frac{P(B=g,y,y|W=r,r,c,M=b) \times P(W=r,r,c) \times P(M=b)}{P(B=g,y,y,W=r,r,c)} \tag{3.9}$$

We have an unknown, $\frac{P(W=r,r,c)}{P(B=g,y,y,W=r,r,c)}$, which we cancel out by dividing by the sum to get probability of Alfred and Betty as,

$$P(M = a|B = g, y, y, W = r, r, c) =$$

$$\frac{P(M = a|B = g, y, y, W = r, r, c)}{P(M = a|B = g, y, y, W = r, r, c) + P(M = b|B = g, y, y, W = r, r, c)} \tag{3.10}$$

$$\tag{3.11}$$

$$P(M = a|B = g, y, y, W = r, r, c) = \frac{0.75 \times 0.8 \times 0.2 \times 0.8}{0.25 \times 0.2 \times 0.8 \times 0.2 + 0.75 \times 0.8 \times 0.2 \times 0.8} \tag{3.12}$$

$$= \frac{1}{13} \tag{3.13}$$

$$P(M = b|B = g, y, y, W = r, r, c) =$$

$$\frac{P(M = b|B = g, y, y, W = r, r, c)}{P(M = a|B = g, y, y, W = r, r, c) + P(M = b|B = g, y, y, W = r, r, c)} \tag{3.14}$$

$$\tag{3.15}$$

$$P(M = b|B = g, y, y, W = r, r, c) = \frac{0.25 \times 0.2 \times 0.8 \times 0.2}{0.25 \times 0.2 \times 0.8 \times 0.2 + 0.75 \times 0.8 \times 0.2 \times 0.8} \tag{3.16}$$

$$= \frac{12}{13} \tag{3.17}$$

Given this sequence of blocks 'g,y,y' and weather pattern 'r,r,c' it is most probable that Betty was throwing the blocks..

4 Act IV

The task in this act is to predict next block in the sequence $B' =?$ given a sequence B and $W' = r$. Specifically, we are interested in computing $P(B' = g|B = g, y, y, W = r, r, c, W' = r)$. We compute this probability as,

$$P(B' = g|B = g, y, y, W = r, r, c, W' = r) =$$

$$\begin{aligned} &P(B' = g|W' = r, M = a) \times P(M = a|B = g, y, y, W = r, r, c) \\ +&P(B' = g|W' = r, M = b) \times P(M = b|B = g, y, y, W = r, r, c) \end{aligned} \tag{4.1}$$

Since we are not given any information on the monkey that threw the blocks we have marginalized out this variable by summing over two possible values.

$$P(B' = g|B = g, y, y, W = r, r, c, W' = r) = 0.2 \times \frac{1}{13} + 0.8 \times \frac{12}{13} \tag{4.2}$$

$$= \frac{49}{65} \tag{4.3}$$

5 MAP versus Bayesian Inference

1. MAP stands for maximum a posteriori probability

2. MAP is a point estimate (mode of the posterior distribution) which is computed using both likelihood and prior

3. Bayesian inference is based on integrating out the posterior distribution

4. Bayesian inference propagates uncertainty to the quantity that needs to be computed and then integrates the posterior to make a final prediction

6 Bayesian Inference - Summary

- Traditional machine learning (ML) has inputs, outputs and a model that maps inputs to outputs

- In Bayesian inference, we treat random variables as inputs, outputs and model. For example in Act IV of the whimsical monkey problem the inputs are a sequence of blocks (B) and weather (W); the output is the new block in the sequence B'; the monkeys are treated as the model that maps the input to the outputs. In addition we have a prior over the monkeys as set by the wizard.

- Given this equivalence both traditional ML and Bayesian inference makes a prediction for B' (output) given B, W (train set) and W' (test input)

- But unlike traditional machine learning, Bayesian inference need not handle overfitting and will not need to choose an appropriate loss function

- Both these approaches are modeling $P(output|input)$ which is a much smaller space to model (*e.g., for a binary classification problem it is a space of two possible classes) in comparison to* $P(input)$ *which scales with the number of data examples*