Bayesian Machine Learning, MLE, MAP - I

Nipun Batra

February 7, 2020

IIT Gandhinagar

- Allows us to incorporate knowledge into the model, *irrespective* of what the data has to say.

Bayesian Machine Learning

- Allows us to incorporate knowledge into the model, *irrespective* of what the data has to say.
- An example: Time-series meteorological data - temporal patterns in temperature.

Bayesian Machine Learning

- Allows us to incorporate knowledge into the model, *irrespective* of what the data has to say.
- An example: Time-series meteorological data - temporal patterns in temperature.
- Particularly useful when we do not have a large amount of data - use what *we* know about the model than depend on the data.

Bayesian Machine Learning

- Allows us to incorporate knowledge into the model, *irrespective* of what the data has to say.
- An example: Time-series meteorological data - temporal patterns in temperature.
- Particularly useful when we do not have a large amount of data - use what *we* know about the model than depend on the data.
- Also allows us to predict with confidence quantified typically using variance.

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Notation: Let θ denote the parameters of the model and let $\mathcal{D}$ denote observed data. From Bayes Rule, we have

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

- In the above equation $P(\theta|\mathcal{D})$ is called the posterior, $P(\mathcal{D}|\theta)$ is called the likelihood, $P(\theta)$ is called the prior and $P(\mathcal{D})$ is called the evidence.

Likelihood, Prior and Posterior

- Likelihood $P(\mathcal{D}|\theta)$ quantifies how the current model parameters describe the data. It is a function of θ. Higher the value of $P(\mathcal{D}|\theta)$, the better the model describes the data.
- Prior $P(\theta)$ is the knowledge we incorporate into the model, *irrespective* of what the data has to say. As an example, if we have n model parameters, $\theta \sim \mathcal{N}(0, I_n)$ could be the knowledge we are incorporating into the model.
- Posterior $P(\theta|\mathcal{D})$ is the probability that we assign to the parameters after observing the data. Posterior takes into account prior knowledge unlike likelihood.
- Posterior $\propto$ Likelihood $\times$ Prior

Bayesian Learning is well suited for online learning

- In online learning, data points arrive one by one. We can index this using timestamps. So we have one data point for each timestamp.
- Initially no data: We only have $P(\theta)$, which is prior knowledge which we have about the model parameters, *without* observing any data.
- Suppose we observe $\mathcal{D}_1$ at timestamp 1. Now we have new information. This knowledge is encoded as $P(\theta|\mathcal{D}_1)$.
- Now, $\mathcal{D}_2$ arrives at timestamp 2. Now we have $P(\theta|\mathcal{D}_1)$, acting as the prior knowledge before we observe $\mathcal{D}_2$.
- Similarly, for timestamp n, we will have $P(\theta|\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \ldots \mathcal{D}_{n-1})$ acting as the prior knowledge before we observe $\mathcal{D}_n$.

Bayesian Learning is well suited for online learning

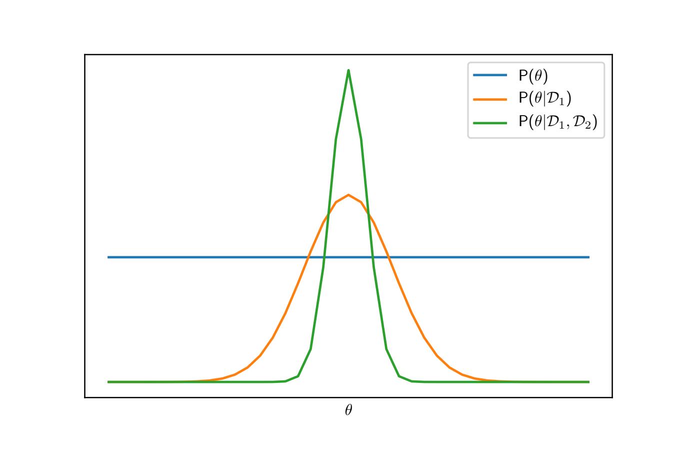

Figure 1: Online Learning: Variation of Prior as more data points arrive.

Coin flipping

- Assume we do a coin flip multiple times and we get the following observation: {H, H, H, H, H, H, T, T, T, T}: 6 Heads and 4 Tails
- What is $P(Head)$?
- Is your answer: 6/10. Why?

Coin flipping: Maximum Likelihood Estimate (MLE)

- We have $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_N\}$ for N observations where each $\mathcal{D}_i \in \{H, T\}$
- Assume we have n_H heads and n_T tails, $n_H + n_T = N$
- Let us have $P(H) = \theta, P(T) = 1 - \theta$
- We have Likelihood, $L(\theta) = P(\mathcal{D}|\theta) = P(\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N|\theta)$
- Since observations are i.i.d., $L(\theta) = P(\mathcal{D}_1|\theta).P(\mathcal{D}_2|\theta)...P(\mathcal{D}_N|\theta)$

Coin flipping: Maximum Likelihood Estimate (MLE)

- $$P(\mathcal{D}_i|\theta) = \begin{cases} \theta, & \text{for } \mathcal{D}_i = H \\ 1-\theta, & \text{for } \mathcal{D}_i = T \end{cases}$$
- Thus, $L(\theta) = \theta^{n_H} \times (1-\theta)^{n_T}$
- Log-Likelihood, $LL(\theta) = n_H log\theta + (n_T)(log(1-\theta))$
- $\frac{\partial LL(\theta)}{\partial\theta} = \frac{n_H}{\theta} - \frac{n_T}{1-\theta}$
- For optima, set derivative of LL to zero.
- $\frac{n_H}{\theta} - \frac{n_T}{1-\theta} = 0$

$$\theta = \frac{n_H}{n_H + n_T}$$

Question: Is this maxima or minima?

$$\frac{\partial^2 LL(\theta)}{\partial\theta^2} = \frac{-n_H}{\theta^2} + \frac{-n_T}{(1-\theta)^2} \in \mathbb{R}_-$$

Thus, the solution is a maxima.

- Note that for the example above θ is 0.6 after maximizing the likelihood. As already mentioned, likelihood tries to explain the data using the model. The model we had above had a single parameter θ and for $\theta = 0.6$, the model *best* explains the given data, as computed using the maximum likelihood.
- Any issues with maximum likelihood estimate or MLE?

Maximum A Posteriori estimate (MAP)

- **MLE does not handle prior knowledge**: What if we know that our coin is biased towards head?
- **MLE can overfit**: What is the probability of heads when we have observed 6 heads and 0 tails?

Maximum A Posteriori estimate (MAP)

Goal: Maximize the Posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{D}) \tag{1}$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta) \tag{2}$$

Prior distributions

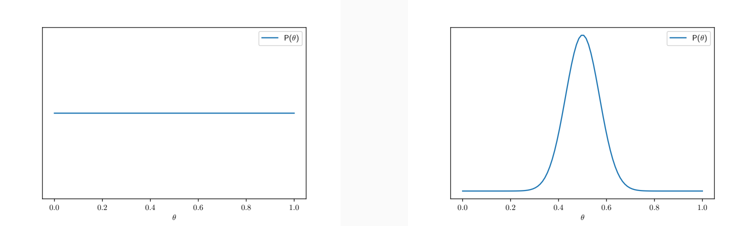

Figure 2: Uniform and Non Uniform Prior.

Beta Distribution

- It is a continuous probability distribution defined on $[0, 1]$, which has two parameters a and b.
- $Beta(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$. Note the similarity with the binomial distribution.
- $\Gamma(n) = (n-1)!$ when n is a natural number.
- $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$

Beta Distribution Examples

- $Beta(\theta|1,1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = 1$. This is the uniform distribution on [0,1].
- $Beta(\theta|2,2) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)}\theta(1-\theta) = 6\theta(1-\theta)$.

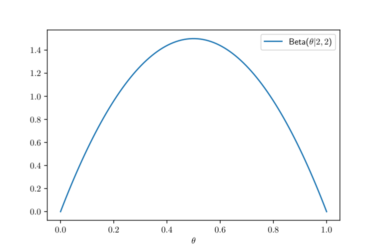

Figure 3: Beta$(\theta|2,2)$

- Note: $Beta(\theta|a,1)$ indicates higher probability of heads than tails.

Coin toss: MAP estimate

- $\mathcal{D} = n_H, n_T$
- $P(\theta) = Beta(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta^{b-1})$.
- $\hat{\theta}_{MAP} = \arg\max_\theta P(\mathcal{D}|\theta)P(\theta)$
- $\implies \hat{\theta}_{MAP} = \arg\max_\theta \theta^{n_H}(1-\theta)^{n_T}\theta^{a-1}(1-\theta)^{b-1} \times k$
- Equivalently, $\hat{\theta}_{MAP} = \arg\max_\theta \theta^{n_H+a-1}(1-\theta)^{n_T+b-1}$
- $\therefore \hat{\theta}_{MAP} = \frac{n_H+a-1}{n_H+n_T+a+b-2}$

Conjugate Prior

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

- $P(\theta)$ is conjugate to $P(\mathcal{D}|\theta)$ if $P(\theta|\mathcal{D})$ and $P(\theta)$ are from the same distribution family.
- Example: Bernoulli likelihood has gamma as conjugate.

Relationship between MLE and MAP

- When is $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$?
- Answer: When prior $P(\theta)$ is uniform, maximizing the likelihood is the same as maximizing the posterior distribution.