Bayesian Linear Regression

Nipun Batra

July 2, 2019

IIT Gandhinagar

MLE, MAP, Bayesian

Bayes Rule - 1

- $P(A|B)P(B) = P(B|A)P(A)$
- Let us consider an example from Wikipedia:
    - A particular drug is 99% sensitive and 99% specific
    - i.e. test will produce 99% true positive results for drug users and 99% true negative results for non-drug users
    - 0.5% of people are users of the drug
    - Question: What is the probability that a randomly selected individual with a positive test is a drug user?

Bayes Rule - 2

- Test will produce 99% true positive results for drug users and 99% true negative results for non-drug users $\implies$
  - $P(Test = +|User = Drug) = 0.99$, or, $P(+|User) = 0.99$
  - and $P(-|\overline{User}) = 0.99$
- 0.5% of people are users of the drug $\implies P(User) = 0.005$
- Question: What is the probability that a randomly selected individual with a positive test is a drug user? $\implies P(User|+) =?$
- $P(User|+) = \frac{P(+|User)P(User)}{P(+)} =$
- $\frac{P(\text{+}|\text{User})P(\text{User})}{P(\text{+}|\text{User})P(\text{User})+P(\text{+}|\overline{User})P(\overline{User})} = \frac{0.99\times0.005}{0.99\times0.005+0.01\times0.995} \approx .332$

Another example on Bayes rule

Bayes Rule for Machine Learning

- $P(A|B)P(B) = P(B|A)P(A)$
- Let us consider for a machine learning problem:
    - A = Parameters (θ)
    - B = Data ($\mathcal{D}$)
- We can rewrite the Bayes rule as:
    - $P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$
    - Posterior:
    - Prior:
    - Likelihood
    -

Likelihood

- Likelihood is a function of θ
- Given a coin flip and 5 H and 1 T, what is more likely: P(H) = 0.5 or P(H) = 1

Bayesian Learning is well suited for online settings

content...

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 maxima, set derivative of LL to zero
- $\frac{n_H}{\theta} + \frac{n_T}{1-\theta} = 0$

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

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

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}}\, P(\theta|\mathcal{D})$$
$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}}\, P(\mathcal{D}|\theta)P(\theta)$$

Prior distributions

Beta Distribution

Beta Distribution

Coin toss: MAP estimate