Bayesian Machine Learning, MLE, MAP - I

Nipun Batra

February 20, 2021

IIT Gandhinagar

Bayesian Machine Learning

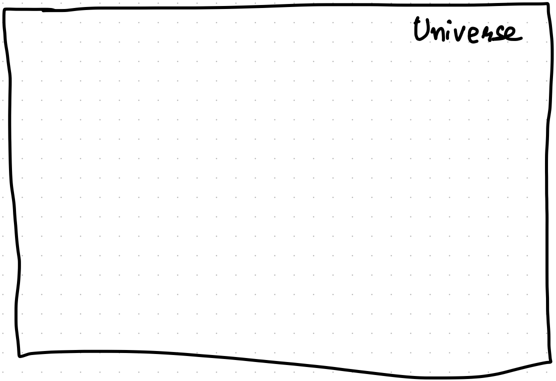
- Allows us to incorporate prior knowledge into the model, *irrespective* of what the data has to say.

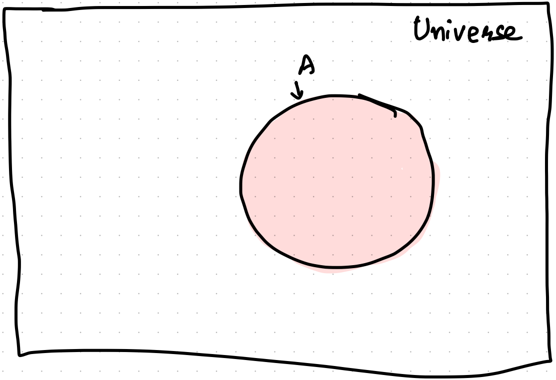
Bayesian Machine Learning

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

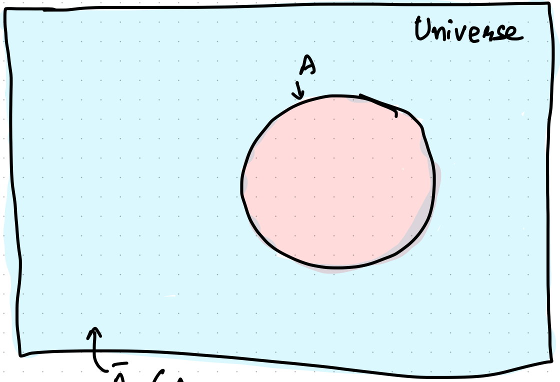
- Allows us to incorporate prior knowledge into the model, *irrespective* of what the data has to say.
- Particularly useful when we do not have a large amount of data - use what *we* know about the model than depend on the data.
- Also allows us to predict with confidence quantified typically using variance.


Universe



Universe :
All people

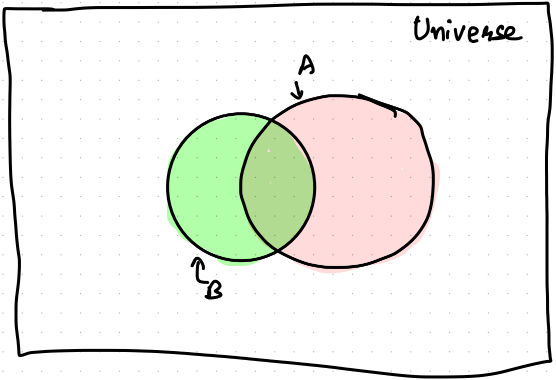
A :
People having
disease



Universe:
All people

A:
People having disease

$\bar{A}$ (A complement or Not A or $\neg A$)
People not having disease

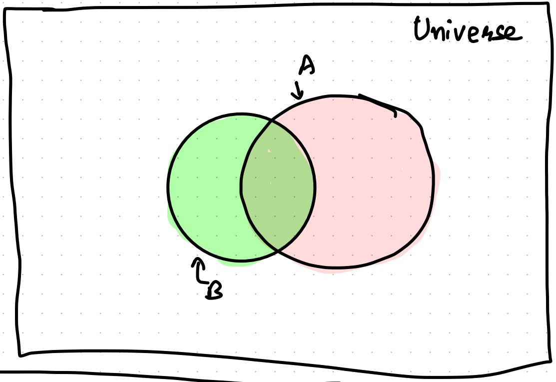


Universe:
All people

A:
People having disease

B:
People tested + for disease

$\neg B$ or $\overline{B}$: People tested -ve for disease



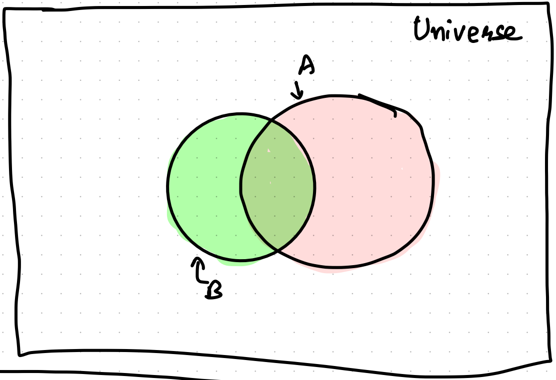
Universe: All people

A: People having disease

B: People tested + for disease

$\neg B$ or $\overline{B}$: People tested -ve for disease

$$P(\text{Disease}) = \frac{|A|}{|\text{Universe}|}$$



Universe:
All people

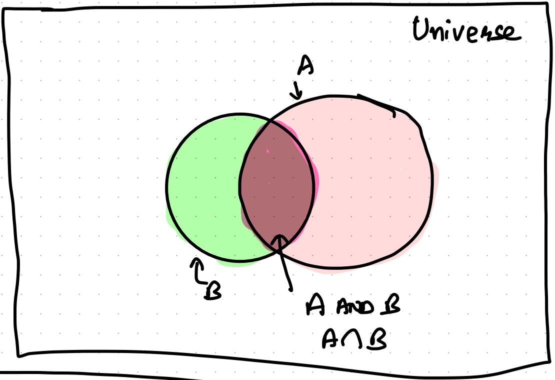
A:
People having disease

B:
People tested + for disease

$$P(\text{Disease}) = \frac{|A|}{|\text{Universe}|} = P(A)$$

$$P(\text{Test}) = \frac{|B|}{|\text{Universe}|} = P(B)$$

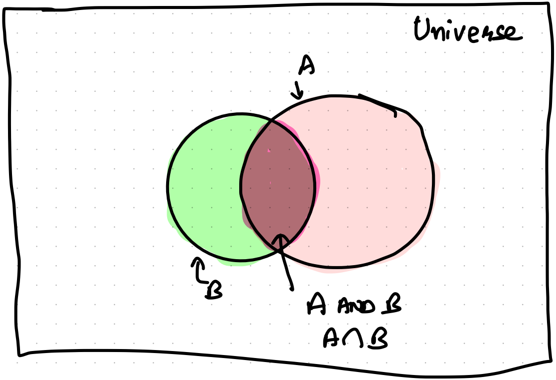
$\neg B$ or $\overline{B}$: People tested -ve for disease



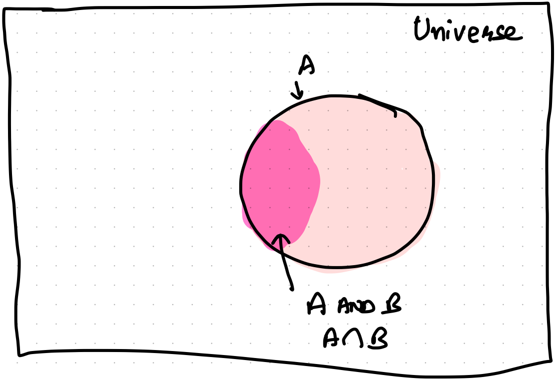
$$P(\text{Disease}) = \frac{|A|}{|\text{Universe}|} = P(A)$$

$$P(\text{Test}) = \frac{|B|}{|\text{Universe}|} = P(B)$$

$$P(\underset{(=T)}{\text{Disease}} \text{ AND } \underset{=(T)}{\text{TEST}}) = P(A \text{ AND } B) = \frac{|A \cap B|}{|\text{Universe}|}$$



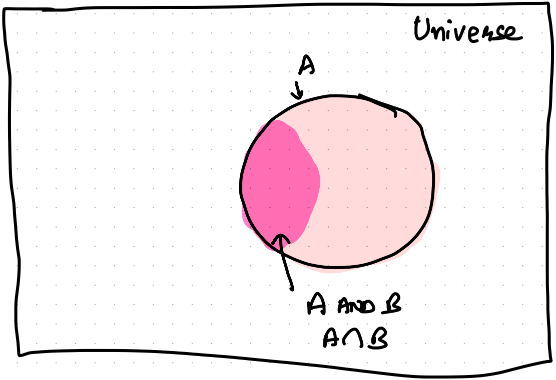
P(Test(=T) GIVEN DISEASE(=T)



P(Test(=T) GIVEN DISEASE(=T))

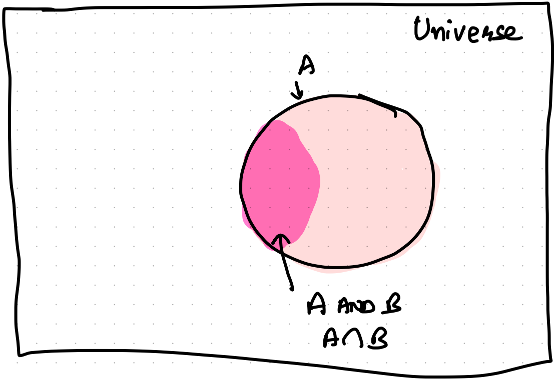
$$= P(B \mid A)$$

$$= \frac{|A \cap B|}{|A|} = \frac{\frac{|A \cap B|}{|\text{Universe}|}}{\frac{|A|}{|\text{Universe}|}} = \frac{P(AB)}{P(A)}$$



$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$



$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

$$\boxed{P(A \mid B)\,P(B) = P(B \mid A)\,P(A)}$$

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Example: You tested positive for a disease. But, the test is only 99% accurate.

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Example: You tested positive for a disease. But, the test is only 99% accurate.
- $P(Test = +ve|Disease = True) = 0.99$

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Example: You tested positive for a disease. But, the test is only 99% accurate.
- $P(Test = +ve|Disease = True) = 0.99$
- $P(Test = -ve|Disease = False) = 0.99$

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Example: You tested positive for a disease. But, the test is only 99% accurate.
- $P(Test = +ve|Disease = True) = 0.99$
- $P(Test = -ve|Disease = False) = 0.99$
- Also, the disease is a rare one. Only one in 10,000 has it.

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Example: You tested positive for a disease. But, the test is only 99% accurate.
- $P(Test = +ve|Disease = True) = 0.99$
- $P(Test = -ve|Disease = False) = 0.99$
- Also, the disease is a rare one. Only one in 10,000 has it.
- Given the result of test is positive, what is the probability that someone has the disease?

Bayes Rule

- $P(T|D) = 0.99$
- $P(\bar{T}|\bar{D}) = 0.99$
- $P(T|\bar{D}) = 0.01$
- $P(\bar{T}|D) = 0.01$
- $P(D) = 10^{-4}$
- $P(\bar{D}) = 1 - 10^{-4}$

Given the above, calculate $P(D|T)$.

Problem

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} \quad (1)$$

Problem

$$
\begin{aligned}
P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\
&= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}
\end{aligned}
\tag{2}
$$

$$
= \frac{(.99)\left(10^{-4}\right)}{(.99)\left(10^{-4}\right) + (.01)\left(1 - 10^{-4}\right)}
$$

Problem

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} \\ = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} \tag{3}$$

$$= \frac{(.99)\left(10^{-4}\right)}{(.99)\left(10^{-4}\right) + (.01)\left(1 - 10^{-4}\right)} = 0.09 << 0.99$$

Bayes Rule

- Notation: Let θ denote the parameters of the model and let $\mathcal{D}$ denote observed data. From Bayes Rule, we have

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

Bayes Rule

- 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.

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.

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.

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.

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.

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 $P(\theta|\mathcal{D}_1)$.

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 $\mathcal{D}_2$.

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 $P(\theta|\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \dots \mathcal{D}_{n-1})$ acting as the prior knowledge before we observe $\mathcal{D}_n$.

Bayesian Learning is well suited for online learning

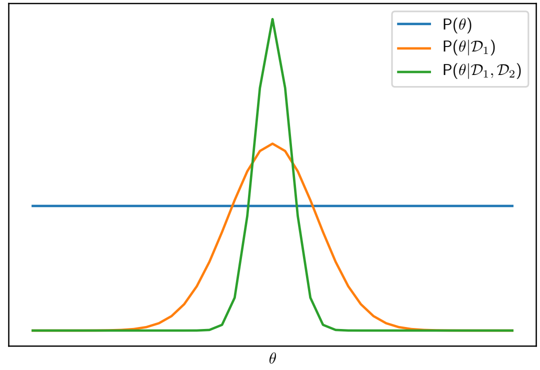


Figure 1: Online Learning: Variation of Prior as more data points arrive.

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: $4/10 = 0.4$. But why?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: $4/10 = 0.4$. But why?
- Answer 1: Probability defined as a measure of long running frequencies

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: $4/10 = 0.4$. But why?
- Answer 1: Probability defined as a measure of long running frequencies
- Answer 2: What is likelihood of seeing the above sequence when the p(Head)$=\theta$?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: $4/10 = 0.4$. But why?
- Answer 1: Probability defined as a measure of long running frequencies
- Answer 2: What is likelihood of seeing the above sequence when the p(Head)$=\theta$?
- Idea find MLE estimate for θ

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$
- Log-likelihood = $\mathcal{LL}(\theta) = n_h \log(\theta) + n_t \log(1-\theta)$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$
- Log-likelihood $= \mathcal{LL}(\theta) = n_h \log(\theta) + n_t \log(1-\theta)$
- $\frac{\partial \mathcal{LL}(\theta)}{\partial \theta} = 0 \implies \frac{n_h}{\theta} + \frac{n_t}{1-\theta} = 0 \implies \theta_{MLE} = \frac{n_h}{n_h + n_t}$

Question: Is this maxima or minima?

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.

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.

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

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{D}) \tag{4}$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta) \tag{5}$$