GENERAL ASSEMBLY

MAXIMUM LIKELIHOOD ESTIMATORS

Joseph Nelson, Data Science Immersive

AGENDA

- Review of Bayes Theorem
- What is Maximum Likelihood Estimation?
- Calculating MLE
- MLE

BAYES THEOREM EXAMPLE

- Imagine you have one fair coin, and one double-sided heads coin. Build a tree the describes the possible outcomes we may expect.
- What is the probability that you have the fair coin?

BAYES THEOREM EXAMPLE

- Imagine you have one fair coin, and one double-sided heads coin. Build a tree the describes the possible outcomes we may expect.
- What is the probability that you have the fair coin?

- Now, after you've built the tree, you find that the coin flipped was actually heads. What is the probability that the coin is the fair coin?

BAYES THEOREM EXAMPLE

- Imagine you have one fair coin, and one double-sided heads coin. Build a tree the describes the possible outcomes we may expect.
- What is the probability that you have the fair coin?

- Now, after you’ve built the tree, you find that the coin flipped was actually heads. What is the probability that the coin is the fair coin?

- You repeat this exercise. Draw the tree.

BAYES THEOREM EXAMPLE

- Imagine you have one fair coin, and one double-sided heads coin. Build a tree the describes the possible outcomes we may expect.
- What is the probability that you have the fair coin?

- Now, after you've built the tree, you find that the coin flipped was actually heads. What is the probability that the coin is the fair coin?

- You repeat this exercise. Draw the tree.

- You find that the coin flipped was heads again. What is the probability that the coin is a fair coin?

BAYES THEOREM

- Probability of a fair coin given a heads: probability of a fair coin divided by all the chances of having a heads

$$\Pr(\mathrm{A}|\mathrm{X}) = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X})}$$

MAXIMUM LIKELIHOOD ESTIMATION

- Imagine we have a Bernoilli Distribution:
- $p^5(1-p)^4$
- The function would have to "peak" at some point; therefore, in some way, the function would have to look somewhat like a inverted parabola
- This peak would have to be a 'global' peak, i.e. it can have multiple peaks, but only one can be the 'largest’
- Once we find the peak, the value/level of the peak is not what we're actually interested in (the Y-value of the function). What we're actually interested in is to guess which X-value of the function (the independent variable), needs to be inputted to get a particular Y.

MAXIMUM LIKELIHOOD ESTIMATION

- Imagine we have a Bernoilli Distribution:
- $p^5(1-p)^4$
- The function would have to "peak" at some point; therefore, in some way, the function would have to look somewhat like a inverted parabola
- This peak would have to be a 'global' peak, i.e. it can have multiple peaks, but only one can be the 'largest’
- Once we find the peak, the value/level of the peak is not what we're actually interested in (the Y-value of the function). What we're actually interested in is to guess which X-value of the function (the independent variable), needs to be inputted to get a particular Y.

MAXIMUM LIKELIHOOD ESTIMATION

- Code: Graph

- Imagine we have a Bernoilli Distribution:
- $p^5(1-p)^4$
- The function would have to "peak" at some point; therefore, in some way, the function would have to look somewhat like a inverted parabola
- This peak would have to be a 'global' peak, i.e. it can have multiple peaks, but only one can be the 'largest’
- Once we find the peak, the value/level of the peak is not what we're actually interested in (the Y-value of the function). What we're actually interested in is to guess which X-value of the function (the independent variable), needs to be inputted to get a particular Y.

MAXIMUM LIKELIHOOD ESTIMATION

- Code: Graph
- Differentiate: maximize the function (FOC) $p^5(1-p)^4 = 0$

MAXIMUM LIKELIHOOD ESTIMATION

- Code: Graph

- Differentiate: maximize the function (FOC) $p^5(1-p)^4 = 0$
- Apply the product rule: $5p^4(1-p)^4 - 4(1-p)^3p^5 = 0$

MAXIMUM LIKELIHOOD ESTIMATION

- Code: Graph

- Differentiate: maximize the function (FOC) $p^5(1-p)^4 = 0$
- Apply the product rule: $5p^4(1-p)^4 - 4(1-p)^3p^5 = 0$
- "Balance" the equation: $5p^4(1-p)^4 = 4(1-p)^3p^5$

MAXIMUM LIKELIHOOD ESTIMATION

- Code: Graph

- Differentiate: maximize the function (FOC) $p^5(1-p)^4 = 0$
- Apply the product rule: $5p^4(1-p)^4 - 4(1-p)^3p^5 = 0$
- "Balance" the equation: $5p^4(1-p)^4 = 4(1-p)^3p^5$
- Algebra: $p = .55556$

- Did it work? Check the graph…

WHAT DID WE JUST DO?

- Class?

WHAT DID WE JUST DO?

- Class?

- We determined what x (input) values provides the maximum chance of some scenario.
- Let's do an example

WHAT DID WE JUST DO?

- Let's say we have a coin that comes up heads with some probability θ

- We see two heads come up. Our likelihood then becomes $P(D|\theta)=\theta^2$
- How do we maximize this function?

WHAT DID WE JUST DO?

- Let's say we have a coin that comes up heads with some probability θ

- We see two heads come up. Our likelihood then becomes $P(D|\theta)=\theta^2$
- How do we maximize this function?
- $\theta=1$

WHAT DID WE JUST DO?

- Let's say we have a coin that comes up heads with some probability θ

- We see two heads come up. Our likelihood then becomes $P(D|\theta)=\theta^2$
- How do we maximize this function?
- $\theta=1$
- So, our MLE is that the coin always comes up heads, and so we predict future coins will all come up head

WHAT DID WE JUST DO?

- Let's say we have a coin that comes up heads with some probability θ

- We see two heads come up. Our likelihood then becomes $P(D|\theta)=\theta^2$
- How do we maximize this function?
- $\theta=1$
- So, our MLE is that the coin always comes up heads, and so we predict future coins will all come up head

BAYES VS MLE: DOCTOR

- Imagine you are a doctor. You have a patient who shows an odd set of symptoms. You look in your doctor book and decide the disease could be either a common cold or lupus. Your doctor book tells you that if a patient has lupus then the probability that he will show these symptoms is 90%. It also states that if the patient has a common cold then the probability that he will show these symptoms is only 10%. Which disease is more likely?
- (Hint: does rarity of disease matter?)

BAYES VS MLE: DOCTOR

- Imagine you are a doctor. You have a patient who shows an odd set of symptoms. You look in your doctor book and decide the disease could be either a common cold or lupus. Your doctor book tells you that if a patient has lupus then the probability that he will show these symptoms is 90%. It also states that if the patient has a common cold then the probability that he will show these symptoms is only 10%. Which disease is more likely?

MLE: A BETTER PATH – POSTERIOR ESTIMATION

- Maximum A-Posterior Estimation (MAP)
- Estimate the maximum likelihood of the posterior rather than the prior
- $\theta_{MAP} = \text{argmax}_{\theta} P(\theta|D)$
- $\theta_{MLE} = \text{argmax}_{\theta} P(D|\theta)$

MLE: A BETTER PATH – POSTERIOR ESTIMATION

- Maximum A-Posterior Estimation (MAP)
- Estimate the maximum likelihood of the posterior rather than the prior
- $\theta_{MAP} = \text{argmax}_{\theta}\ P(\theta|D)$
- $\theta_{MLE} = \text{argmax}_{\theta}\ P(D|\theta)$

- Code Example

MLE: A BETTER PATH – POSTERIOR ESTIMATION

- Maximum A-Posterior Estimation (MAP)
- Estimate the maximum likelihood of the posterior rather than the prior
- $\theta_{MAP} = \text{argmax}_{\theta} P(\theta|D)$
- $\theta_{MLE} = \text{argmax}_{\theta} P(D|\theta)$

- Advantages:
- Easy and interpretable
- Avoids overfitting
- Tends to follow the same asymptotic distribution

MLE: A BETTER PATH – POSTERIOR ESTIMATION

- Maximum A-Posterior Estimation (MAP)
- Estimate the maximum likelihood of the posterior rather than the prior
- $\theta_{MAP} = \text{argmax}_{\theta}\, P(\theta|D)$
- $\theta_{MLE} = \text{argmax}_{\theta}\, P(D|\theta)$

- Disdvantages:
- Must assume a prior on θ
- No representation of uncertainty in θ