Day-64, Feb 2, 2025 (Magh 20, 2081)

1. Bayesian Statistics
2. Updating Beliefs
3. Frequentist vs Bayesian
4. Relationship between MAP, MLE and Regularization

We keep prior information update the prior to posterior belief as we do the probability updation.

This scenario explains how Bayesian statisticians update their beliefs after observing data, using a coin toss as an example. Here's a breakdown of the process and how the statisticians' priors influence their posterior beliefs:

1. The Priors:
First Bayesian (Narrow prior around 0.5): This Bayesian starts with a strong belief that the coin is fair (i.e., heads has a 50% probability). The prior is very narrow, showing high confidence in this belief.
Second Bayesian (Wider prior around 0.5): This Bayesian also believes the coin is fair but is open to the possibility that the coin could be biased, so their prior is wider, representing a belief in a broader range of probabilities.
Third Bayesian (Non-informative prior): This Bayesian has no strong beliefs about the coin's fairness and assigns equal probability to all possible outcomes, making this a non-informative prior.

2. Observation (1 Head from Coin Toss):
First Bayesian: Despite observing heads, the first Bayesian's belief doesn't change much, as they are very confident that the coin is fair.
Second Bayesian: This Bayesian updates their belief slightly. Since they had a broader prior, their belief shifts slightly towards a higher probability of heads (but not dramatically).
Third Bayesian: The third Bayesian's belief shifts significantly after just one observation. Starting with no prior information, their belief is now more heavily weighted towards heads, showing how sensitive they are to the data.

3. After Multiple Tosses (8 Heads, 2 Tails):
First Bayesian: The belief remains almost unchanged, still strongly centered around 0.5, reflecting their high confidence in a fair coin.
Second Bayesian: The second Bayesian's belief shifts to around 0.65, showing a moderate bias towards heads after more data.
Third Bayesian: The third Bayesian now believes with high certainty that the probability of heads is around 0.8, reflecting a strong belief based on the data.

4. MAP Estimation (Maximum A Posteriori):
First Bayesian: The MAP estimate for heads is 0.501, very close to the original belief that the coin is fair.
Second Bayesian: The MAP estimate is 0.607, showing a moderate bias towards heads but still not extreme.
Third Bayesian: The MAP estimate is 0.8, reflecting a strong belief that the coin is biased towards heads after observing the data.

5. Conclusion:
Effect of Priors: Even though all three Bayesian statisticians observe the same data (8 heads and 2 tails), their priors (initial beliefs) influence how strongly they update their beliefs. The first Bayesian's conservative prior leads to minimal change, while the third Bayesian's non-informative prior leads to a more dramatic shift.
Bayesian vs Frequentist: When a Bayesian uses a non-informative prior, their MAP estimate aligns with a frequentist approach, which estimates the probability based purely oBayesianserved data. This highlights how priors uniquely influence Bayesian statistics, making them distinct from frequentist methods.
In short, this example illustrates how Bayesian statistics incorporate prior beliefs, and how those beliefs are updated with data, producing different outcomes depending on the initial assumptions.

Three Bayesians are trying to figure out the probability of a coin landing heads. Each has a different belief (prior):
First Bayesian: Thinks the coin is definitely fair (prior centered around 0.5).
Second Bayesian: Thinks the coin is mostly fair but could be slightly biased (wider prior around 0.5).
Third Bayesian: Has no initial belief, so they assume all possibilities are equally likely (non-informative prior).

After observing one coin toss that lands heads:
First Bayesian: Belief doesn’t change much.
Second Bayesian: Updates belief slightly towards heads.
Third Bayesian: Belief shifts strongly towards heads (since they had no prior information).

After 8 heads and 2 tails, the MAP (maximum a posteriori) estimates:
First Bayesian: Still believes the coin is almost fair (0.501 probability of heads).
Second Bayesian: Believes the coin is slightly biased towards heads (0.607 probability).
Third Bayesian: Believes the coin is mostly biased towards heads (0.8 probability).

The key takeaway: The prior (initial belief) affects how strongly the Bayesian updates their beliefs based on the data. The third Bayesian's belief changes the most because they started with no prior, while the first Bayesian's belief hardly changes because they were already very confident the coin was fair.

Bayes' Theorem

P(A|B) → ?
P(B|A) → likelihood
P(B) → Mujh

1. **Bayes' Theorem**:
   - It helps calculate the **posterior** (updated probability) of an event **A** based on new evidence **B**.
   - The formula is:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

     - **P(A|B)**: Posterior — the probability of **A** given **B** (what we're solving for).
     - **P(B|A)**: Likelihood — the probability of **B** given **A**.
     - **P(A)**: Prior — initial probability of **A** before evidence **B**.
     - **P(B)**: Marginal likelihood — the overall probability of **B**.

Example - Job Offer:

A: Whether you'll be offered the job.
B: Evidence that you were asked for a follow-up phone call.
The goal is to update your belief about the job offer based on this new evidence.
Example - Fair or Biased Coin:

You have two possible coins: fair (P(heads) = 0.5) or biased (P(heads) = 0.8).
Start with priors: You believe there's a 75% chance the coin is fair and 25% chance it's biased.
Flip the coin: You get heads (evidence).
Bayes' Theorem helps update the belief that the coin is fair based on the flip outcome.
Posterior Calculation:

The posterior is calculated by multiplying the likelihood (how likely heads is for a fair coin) by the prior (how likely the coin is fair).
Then, divide by the total probability of getting heads (considering both fair and biased coins).
Generalized Formula:

Bayes' theorem can be generalized for both discrete and continuous random variables.
If both x and y are discrete, the formula uses probability mass functions (PMFs).
If either or both are continuous, use probability density functions (PDFs).
Machine Learning:

In machine learning, we often update the prior belief about a model parameter (denoted as θ) using evidence.
The main idea is that Bayes' theorem allows you to update your beliefs about an event by incorporating new data or evidence.

Bayes' Theorem is the foundational formula for calculating probabilities.
Naive Bayes is an algorithm that applies Bayes' Theorem but assumes feature independence (the "naive" part) to simplify the calculation.

Full worked Example on Bayesian Statistics:

Bayesian statistics, specifically how to update prior beliefs with new data through Bayes' theorem.

Here's a breakdown of key points covered in the video:

Setting Up the Problem:

The coin flip problem is used to illustrate how Bayes' theorem works in practice. The goal is to estimate the probability, θ, that the coin lands heads.
The data collected consists of 10 flips, with 8 heads and 2 tails.
The prior belief (probability of θ) is initially uniform, meaning every possible value between 0 and 1 is equally likely.
Likelihood Calculation:

Using the collected data, the likelihood of observing the specific outcomes (8 heads and 2 tails) is computed for different values of θ.
The likelihood is expressed as the product of individual Bernoulli probabilities: θ^8 * (1 - θ)^2.
Bayes' Theorem Application:

Bayes' theorem is used to update the prior belief with the likelihood, resulting in a posterior belief about the value of θ.
The posterior is proportional to the product of the likelihood and the prior, with the denominator being a normalizing constant (often ignored in practical calculations).
Posterior Distribution:

The result of the update is a Beta distribution, which represents the updated belief about θ after considering the observed data.
The shape of this distribution is much more informative than the original flat (uniform) prior.

Maximum A Posteriori (MAP) Estimate:
The MAP estimate is the value of θ that maximizes the posterior distribution. In this case, it is 0.8, which is the most likely value of θ given the data.

Frequentist vs Bayesian:
A frequentist would simply count the proportion of heads in the data (8 out of 10 flips, yielding θ = 0.8).
A Bayesian approach considers prior beliefs and updates them with new data, which, for an uninformative prior, coincidentally aligns with the frequentist estimate.

Impact of Informative Priors:
If an informative prior is used (based on earlier data), it influences the posterior distribution and can lead to different conclusions than a frequentist approach.
The video shows how incorporating a new set of 10 flips with 6 heads and 4 tails changes the belief, and how Bayesian updating continues.

Final Observations:
As more data is collected, the Bayesian posterior becomes more refined, and the influence of the prior is gradually "watered down."
MLE and MAP estimates converge as more data is collected, making the prior's influence less important over time.

The lecture demonstrates how Bayesian statistics works through a coin flip example, showing how we update our beliefs as we gather more data. Here are the key points:

Basic Setup

1. They use a coin with unknown probability θ (theta) of getting heads
2. Initial 10 flips result in 8 heads and 2 tails
3. They start with a uniform prior (all probabilities equally likely)

The Bayesian Process
- Used Bayes' theorem with a probability density function for θ and probability mass function for the data
- Likelihood of the data given θ was calculated as $\theta^8(1-\theta)^2$
- The posterior distribution turned out to be a beta distribution
- The MAP (Maximum A Posteriori) estimate was 0.8, matching the frequentist estimate of 8/10

Second Round of Data
- Next 10 flips showed 6 heads and 4 tails
- Previous posterior became the new prior
- New MAP estimate was 0.7
- Total across all 20 flips: 14 heads (14/20 = 0.7)

Key Insights:
- With uninformative priors, MAP gives same results as Maximum Likelihood Estimation (MLE)
- The order of incorporating data doesn't matter (all at once or in chunks)

Bayesian methods are most useful when you have:
- Limited data access
- Strong prior beliefs

The main risk is that incorrect priors can lead to wrong conclusions, especially with limited data

Summary of Bayesian Statistics Video

This video explains Bayesian statistics through a coin-flipping example, where prior beliefs about a coin's probability of landing heads (θ) are updated as new data is collected.

Key Concepts:

1. **Bayesian Framework:**
   - The probability of heads (θ) is treated as a random variable.
   - Data (X) consists of independent Bernoulli trials (1 for heads, 0 for tails).
   - Bayes' theorem is used to update beliefs about θ based on observed data.
2. **Step-by-Step Bayesian Updating:**
   - **Initial Experiment:** 10 flips → 8 heads, 2 tails.
   - **Likelihood Calculation:** The probability of this sequence given θ is $\theta^8(1-\theta)^2$.
   - **Prior Selection:** A uniform prior (all θ values equally likely).
   - **Posterior Distribution:** Results in a Beta distribution, concentrating around $\theta \approx 0.8$.
3. **Incorporating More Data:**
   - **New Experiment:** Another 10 flips → 6 heads, 4 tails.
   - **Updated Posterior:** Bayesian updating leads to a new posterior, peaking at $\theta = 0.7$.
4. **Key Takeaways:**
   - **MAP Estimate (Maximum A Posteriori):** The most probable value of θ, which aligns with frequentist MLE when priors are uniform.
   - **Sequential Updating:** Whether data is processed in chunks or all at once, the final posterior remains the same.
   - **Impact of Priors:** With more data, priors become less influential, converging toward frequentist estimates.
   - **Strengths & Weaknesses:**
     - **Bayesian:** Useful with limited data or strong prior knowledge.
     - **Frequentist:** Works well when large amounts of data are available.

Conclusion:
Bayesian statistics continuously update beliefs using new data, making it powerful when data is scarce or prior knowledge is available. However, incorrect priors can lead to misleading conclusions. Ultimately, choosing between Bayesian and frequentist approaches depends on the problem and available data.

1. **Starting with Uncertainty**: You start with a coin whose probability of landing heads (θ) is unknown. You collect data (10 coin flips) to update your belief about θ.
2. **Bayes' Theorem**: The situation is framed using Bayes' theorem, where:
   - **X** is the data collected (the 10 coin flips).
   - θ is the unknown probability of heads.
   - The goal is to learn the probability density function (PDF) of θ.
3. **Likelihood Function**: The likelihood of observing the data, given θ, is computed using the Bernoulli distribution (since each flip is independent). For 8 heads and 2 tails, the likelihood is θ^8 * (1 - θ)^2.
4. **Priors**: You assume a uniform prior (no prior belief about θ), meaning every value of θ between 0 and 1 is equally likely.
5. **Posterior Distribution**: Using Bayes' theorem, you update your beliefs to get the posterior distribution, which is proportional to θ^8 * (1 - θ)^2. This distribution is a Beta distribution, which summarizes your updated belief about the probability of heads.
6. **Maximum a Posteriori (MAP) Estimate**: The MAP estimate is the value of θ that maximizes the posterior distribution. In this case, it turns out to be 0.8 (i.e., the coin has an 80% chance of landing heads).
7. **Updating with New Data**: If you flip the coin 10 more times (getting 6 heads and 4 tails), you update your beliefs again. The new prior becomes the posterior from the previous round, and you repeat the process.
8. **Frequentist vs. Bayesian Approach**: While a frequentist would directly use the data (6 heads out of 10 flips) to estimate θ = 0.6, the Bayesian approach incorporates prior beliefs, resulting in a different estimate before and after updating with new data.

Key Insight: Even though data is added in two chunks (10 flips at a time), the final posterior beliefs are the same as if all data were used at once. This shows the consistency of the Bayesian updating process.

Key Insight: Even though data is added in two chunks (10 flips at a time), the final posterior beliefs are the same as if all data were used at once. This shows the consistency of the Bayesian updating process.

Conclusion:

Bayesian statistics helps to update prior beliefs with new data.
With uninformative priors, the MAP estimate and Maximum Likelihood Estimation (MLE) are the same.
Over time, as more data is collected, the influence of priors diminishes, and the estimates converge.
This process emphasizes how Bayesian statistics allows for continuous updating and refinement of beliefs with accumulating evidence

Relationship Between Regularization, MLE and MAP

How They All Connect

MLE gives us the best fit based on the observed data alone.
MAP adds a prior to prefer simpler models.
Regularization is the way we implement MAP in regression by adding a penalty for complexity.
By combining these ideas, we can train better models that generalize well to new data! 🚀

Detailed Explanation of Regularization, Maximum Likelihood, and Bayesian Inference in Model Selection

This lesson explores the relationship between **Maximum Likelihood Estimation (MLE), Maximum a Posteriori (MAP) Estimation, and Regularization** in machine learning through a **Bayesian approach**.

---

1. Understanding Maximum Likelihood Estimation (MLE)

MLE is a method for estimating the parameters of a statistical model by maximizing the probability of the observed data given the model.

Example:

Suppose we have some data points, and we are trying to fit a model. Different models generate data with different probabilities:

- **Model 1:** $P(\text{data} \mid M_1)$
- **Model 2:** $P(\text{data} \mid M_2)$
- **Model 3:** $P(\text{data} \mid M_3)$

We select the model that **maximizes** $P(\text{data} \mid M)$, meaning the model that best explains the data. This is the essence of **MLE**:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} P(\text{data} \mid \theta)$$

where θ represents the model parameters.

---

2. Bayesian Inference and Maximum A Posteriori (MAP) Estimation

MLE does not account for model complexity—it simply selects the model that best explains the

2. Bayesian Inference and Maximum A Posteriori (MAP) Estimation

MLE does not account for model complexity—it simply selects the model that best explains the
data. However, **Bayesian inference** introduces a **prior probability** for each model, favoring simpler models.

Bayes' Theorem in Model Selection

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model}) P(\text{model})}{P(\text{data})}$$

- $P(\text{data} \mid \text{model})$: **Likelihood** (used in MLE)
- $P(\text{model})$: **Prior** (expresses preference for simpler models)
- $P(\text{model} \mid \text{data})$: **Posterior probability** (used in MAP estimation)

MAP Estimation:

Instead of just maximizing the likelihood like MLE, MAP maximizes the **posterior probability**:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\text{data} \mid \theta) P(\theta)$$

The prior $P(\theta)$ penalizes **complex models**, favoring simpler models with **smaller parameter values**.

---

3. Connection Between MAP and Regularization

Regularization techniques in machine learning (like Ridge and Lasso regression) **naturally arise** from MAP estimation.

Transforming the Probability into a Loss Function

1. The likelihood of data given a model follows a Gaussian distribution:

$$P(\text{data} \mid \text{model}) = e^{-\frac{1}{2}\sum d_i^2}$$

Transforming the Probability into a Loss Function

1. The likelihood of data given a model follows a Gaussian distribution:

$$P(\text{data} \mid \text{model}) = e^{-\frac{1}{2}\sum d_i^2}$$

where d_i represents the differences between predicted and actual values (errors).

2. The prior $P(\text{model})$ follows a normal distribution over parameters:

$$P(\text{model}) = e^{-\frac{1}{2}\sum a_i^2}$$

where a_i are the model's coefficients.

3. To simplify, take the **logarithm**:

$$\log P(\text{data} \mid \text{model}) = -\frac{1}{2}\sum d_i^2$$

$$\log P(\text{model}) = -\frac{1}{2}\sum a_i^2$$

4. **Maximizing Posterior Probability is Equivalent to Minimizing the Regularized Loss Function:**

$$\text{Loss} = \sum d_i^2 + \lambda \sum a_i^2$$

- **First term:** Sum of squared errors (MLE)
- **Second term:** Regularization (prevents overfitting)
- **Lambda λ:** Controls the balance between accuracy and simplicity

This shows that **MAP estimation using a Gaussian prior naturally leads to L2 regularization (Ridge regression).**

---

4. Summary of Key Insights

1. **MLE selects the model that best explains the data** (without considering complexity).
2. **MAP extends MLE by incorporating prior knowledge**, favoring simpler models.
3. **Taking the logarithm transforms the problem into a regularized loss function**:

4. Summary of Key Insights

1. **MLE selects the model that best explains the data** (without considering complexity).
2. **MAP extends MLE by incorporating prior knowledge**, favoring simpler models.
3. **Taking the logarithm transforms the problem into a regularized loss function**:

- **MAP estimation leads to Ridge regression** (L2 regularization).
- Other priors (e.g., Laplace) lead to **Lasso regression** (L1 regularization).

4. **Regularization helps prevent overfitting by penalizing large coefficients**, ensuring a balance between complexity and accuracy.

---

5. Next Steps in the Course

- **Exploratory Data Analysis (EDA) Lab:**
  - Working with the **World Happiness dataset.**
  - Using **linear regression** to predict happiness scores.
  - Testing different features and their impact on predictions.
- **Graded Quiz:**
  - Assessing understanding of Bayesian inference, MLE, MAP, and regularization.

Conclusion

This lesson demonstrates how **Bayesian probability, maximum likelihood, and regularization** are deeply interconnected. **Regularization emerges naturally from a Bayesian framework**, showing that the **best models balance data fit and complexity** to avoid overfitting.

Explaining Maximum Likelihood Estimation (MLE), Maximum A Posteriori Estimation (MAP), and Regularization

This lesson covers three key topics in machine learning:

1. **Maximum Likelihood Estimation (MLE)**
2. **Maximum A Posteriori Estimation (MAP)**
3. **Regularization**

Each of these methods is used to **train models** and **prevent overfitting**, but they approach the problem differently. Let's break it down step by step.

1️⃣ Maximum Likelihood Estimation (MLE)

MLE is a fundamental technique in machine learning. The idea is:

- We assume our **data was generated by some model** with unknown parameters.
- We estimate the parameters by **maximizing the probability (likelihood) of the observed data** given the model.

🔷 **Example**: Suppose we are fitting a straight line $y = ax + b$ to a set of points.

- MLE finds the values of a and b that maximize the likelihood of the observed data.
- If we assume **Gaussian noise**, this means minimizing the **sum of squared errors**:

$$\sum (y_i - \hat{y}_i)^2$$

where y_i are actual data points and $\hat{y}_i$ are the model predictions.

Message ChatGPT

- If we assume **Gaussian noise**, this means minimizing the **sum of squared errors**:

$$\sum (y_i - \hat{y}_i)^2$$

where y_i are actual data points and $\hat{y}_i$ are the model predictions.

Problem with MLE

MLE **only focuses on the data** and does not consider model complexity.

- If we use **too many parameters**, MLE **overfits**, meaning the model fits noise instead of the real pattern.
- If we use **too few parameters**, MLE **underfits**, meaning the model is too simple.

2️⃣ Maximum A Posteriori Estimation (MAP)

MAP **improves MLE** by introducing **prior knowledge** about the model.

- Instead of only maximizing the likelihood $P(\text{Data}|\text{Model})$, we also consider the probability of the model itself:

$$P(\text{Model}|\text{Data}) = P(\text{Data}|\text{Model}) \times P(\text{Model})$$

  - **MLE only maximizes** $P(\text{Data}|\text{Model})$
  - **MAP maximizes** $P(\text{Model}|\text{Data})$ by adding $P(\text{Model})$, which acts as a regularizer.

🔷 **Example**: Instead of choosing any value for a, b, MAP **prefers smaller values** to avoid overfitting.

Key Difference from MLE

Message ChatGPT

Relationship Between MLE, MAP, and Regularization

MLE, MAP, and Regularization are all connected through the **Bayesian framework** and the goal of **finding the best model parameters** while balancing model complexity and data fit.

🔗 1. MLE and MAP: The Bayesian Connection

- **MLE (Maximum Likelihood Estimation)** only focuses on the **likelihood** $P(\text{Data}|\text{Model})$, meaning it finds the parameters that maximize how well the model explains the observed data.
- **MAP (Maximum A Posteriori Estimation)** extends MLE by incorporating **prior knowledge** through Bayes' theorem:

$$P(\text{Model}|\text{Data}) \propto P(\text{Data}|\text{Model}) \times P(\text{Model})$$

- **MLE maximizes** $P(\text{Data}|\text{Model})$ (**data fit**).
- **MAP maximizes** $P(\text{Data}|\text{Model}) \times P(\text{Model})$ (**data fit + prior preference**).
- The prior $P(\text{Model})$ **favors simpler models**, reducing overfitting.

🔗 2. MAP and Regularization: Practical Implementation

- MAP **naturally leads to regularization in regression** when we assume a **Gaussian prior** on parameters.
- The prior $P(\text{Model})$ is often modeled as a **Gaussian distribution**:

$$P(a) \propto e^{-\frac{1}{2}a^2}$$

  - This introduces an **L2 penalty** on large values of a.
  - The MAP estimate then **minimizes**:

$$\sum (y_i - \hat{y}_i)^2 + \lambda \sum a_i^2$$

    which is exactly **L2 regularization (Ridge Regression)**!
- If we assume a **Laplace prior** instead of a Gaussian:

$$P(a) \propto e^{-|a|}$$

  - This leads to **L1 regularization (Lasso Regression)**, which encourages sparsity.

🔗 3. MLE vs. MAP vs. Regularization in Practice

Method	Objective	Formula	Effect
MLE (Maximum Likelihood)	Find parameters that best fit the data	(\max P(\text{Data} | \text{Model}))	
MAP (Maximum A	Find parameters that fit the [illegible] while considering prior beliefs	(\max P(\text{Data} | \text{Model}) \times P(\text{Model}))	

Message ChatGPT

🔗 3. MLE vs. MAP vs. Regularization in Practice

Method	Objective	Formula	Effect
MLE (Maximum Likelihood)	Find parameters that best fit the data	(\max P(\text{Data}	\text{Model}))
MAP (Maximum A Posteriori)	Find parameters that fit the data while considering prior beliefs	(\max P(\text{Data}	\text{Model}) \times P(\text{Model}))
Regularization	Practical implementation of MAP in regression	$\min(\text{Loss} + \lambda \sum a_i^2)$	Prevents overfitting by penalizing large coefficients

🎯 Conclusion

- **MLE** is a purely data-driven approach.
- **MAP** improves MLE by introducing prior beliefs, preventing overfitting.
- **Regularization** is how we apply MAP principles in machine learning models like linear regression.

Thus, **Regularization = MAP with a prior assumption on model parameters!** 🚀

Message ChatGPT

Sources:

1) Coursera Probability & Statistics for Data Science and Machine Learning.