Frequentist vs. Bayesian Estimation

CSE 4309 – Machine Learning
Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington

Estimating Probabilities

- In order to use probabilities, we need to estimate them.
- For example:
  - What is the **prior** probability p(snows | January), that it snows on a January day in Arlington, Texas?
  - **Prior** means that we do not take any current observations into account (like weather in neighboring areas, weather the previous day, etc).
- How can we compute that?

Probability of Snow in January

- What is the **prior** probability p(snows | January), that it snows on a January day in Arlington, Texas?
- To compute that, we can go through historical data, and measure:
  - N: number of January days for which we have weather records for Arlington.
  - S: number of January days (out of the N days above) for which the records indicate that it snowed.

$$\text{p(snows | January)} = \frac{S}{N}$$

Frequentist Approach

- The method we just used for estimating the probability of snow in January is called **frequentist**.
- In the frequentist approach, probabilities are estimated based on observed frequencies in available data.
- The frequentist approach is simple, and widely used.
- However, there are pitfalls.
- Can you think of any?

Frequentist Pitfalls

- In our "snow in January" example, suppose that:
  - The historical record contains data for only two January days.
  - It did not snow either day.
- Then, what is p(snows | January) according to the frequentist approach?
- p(snows | January = 0)
- This means that your system predicts that there is **zero chance of snowing on a January day**.
- Anything wrong with that?

Frequentist Pitfalls

- The frequentist approach can fail miserably, by wrongly predicting 0% or 100% probabilities, based on limited data.
- If an artificial intelligence system predicts 0% chance of something happening, and that something does happen, we do not consider the system either intelligent or successful.
- Example:
  - Suppose that we need to do a very expensive operation, and that the operation will fail if it snows.
  - We ask our AI system (which we blindly trust) what the probability of snow is.
  - The AI system (following the frequentist approach, and limited data) answers that the probability is 0.
  - We perform the operation, it snows, we fail, we swear never to use AI again.

More Data Helps

- The previous pitfall can be (mostly) avoided if we have lots of data on the historical record.
- If we have data for the last 100 years, then we have data for 3100 January days.
- If the true probability of snow is 10%, it is very unlikely that the frequentist approach will give an estimate that is less than 8% or more than 12%.
- How unlikely? You will find out on your next homework.

The Dangers of Absolute Certainty

- Suppose that we have data for all January days from the last 100 years.
- Suppose it never snowed on those days.
- Our frequentist-based AI system again predicts a probability of snow that is 0%?
- Should we risk our lifes, or lots of money, or the fate of humanity, on such a prediction?
- What is the chance that the prediction is wrong?
- Are we being too skeptical if we expect a 100-year pattern to break? If we expect a 1000-year pattern to break?

The Sunrise Problem

- A similar problem to "snow in January" is the sunrise problem.
  - It is sufficiently well known to have its own Wikipedia article: https://en.wikipedia.org/wiki/Sunrise_problem
- Simply stated: you are an ancient (but statistically savvy) human, and you ask the question: what is the probability that the sun will rise tomorrow?
- You have no idea that the Earth is a planet, that the sun is a star, and so on.
- You just know that the sun has risen every day as far as you can remember.

The Sunrise Problem – Frequentist Solution

- The sun has risen every day you can remember.
- Therefore, the sun has risen 100% of the times in your observations.
- Therefore, the probability that the sun will rise tomorrow is 100%.

The Dangers of Absolute Certainty (2)

- For both the "snow in January" problem (version where it has not snowed for 100 years), and for the sunrise problem, the frequentist approach gives an answer with 100% certainty for the outcome.
  - 0% chance it will snow.
  - 100% chance the sun will rise tomorrow.
- Intuitively, this outcome is correct in one case, incorrect in another case.
  - 100 years of not snowing do not guarantee that it will never snow.
  - The sun will rise again tomorrow, no doubt about that.

When Can We Trust Frequentist Conclusions?

- Short answer: we can never trust probability estimations absolutely, but more data helps.
  - There is always a chance that our observations did not reflect the true distribution.
  - The more data we have observed, the more confident we can be that our estimate is close to the true distribution.
- Especially when it comes to predictions of absolute certainty (like 0% chance, or 100% chance), we must be very careful when those predictions are based just on past observations.

Bayesian Estimation

- p(sunrise) = θ.
- We do not know θ.
- The first step in doing **Bayesian Estimation** of a probability distribution, is to assign a **prior** to the parameters that we want to estimate.
- In the sunrise problem, what are the parameters that we want to estimate?

Bayesian Estimation

- p(sunrise) = θ.
- We do not know θ.
- The first step in doing **Bayesian Estimation** of a probability distribution, is to assign a **prior** to the parameters that we want to estimate.
- In the sunrise problem, the only parameter is θ.
- That is why we will indicate p(sunrise) as p_θ(sunrise).
- So, before we look at observations, we must define a p(θ).
- In other words, for each possible θ, we need to define the probability that the probability of sunrise is θ.

Bayesian Estimation

- p_θ(sunrise) = θ.
- Before we look at observations, we must define a **prior** p(θ).
- In other words, for each possible θ, we need to define the probability that the probability of sunrise is θ.
- Unfortunately, there is no automatic way to choose the right prior, or to prove that a certain prior is the right one.
  - We just need to pick one and live with it.

Bayesian Estimation

- p_θ(sunrise) = θ.
- Before we look at observations, we must define a **prior** p(θ).
- For example, suppose that, before we look at any observations, we assume that all values of θ to be equally likely.
- Then, how should we define p(θ) to reflect that assumption?
- p(θ) is uniform for values between 0 and 1, and zero elsewhere.
- In other words:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

Bayesian Estimation

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Why did we assign zero density for $\theta < 0$ and $\theta > 1$?

Bayesian Estimation

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Why did we assign zero density for θ < 0 and θ > 1?
- Because θ itself is a probability value, so it can only take values between 0 and 1.
- Why did we assign density 1 for θ between 0 and 1? Why not assign density 2, for example?

Bayesian Estimation

- $p_\theta(\text{sunrise}) = \theta$.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Why did we assign zero density for $\theta < 0$ and $\theta > 1$?
- Because θ itself is a probability value, so it can only take values between 0 and 1.
- Why did we assign density 1 for θ between 0 and 1? Why not assign density 2, for example?
- Because density 1 is the only value that makes p(θ) integrate to 1.

Bayesian Estimation, Before Day 1

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- According to this prior, what is the probability p(sunrise) for the first day? (Before we have made any observations).

$$p(sunrise) = \int_0^1 p_\theta(sunrise) p(\theta) d\theta$$

$$= ???$$

Bayesian Estimation , Before Day 1

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- According to this prior, what is the probability p(sunrise) for the first day? (Before we have made any observations).

$$p(sunrise) = \int_0^1 p_\theta(sunrise) p(\theta) d\theta$$

$$= \int_0^1 \theta * 1 * d\theta = 0.5$$

Bayesian Estimation , Before Day 1

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- According to this prior, probability p(sunrise) for the first day is 0.5.
- With Bayesian estimation, we can define p(sunrise) **even before we get any observations**.
  - Why? Because we use the prior p(θ), which we pick ourselves manually.
- What would be p(sunshine) according to the frequentist approach?

Bayesian Estimation , Before Day 1

- $p_\theta(\text{sunrise}) = \theta$.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- According to this prior, probability p(sunrise) for the first day is 0.5.
- With Bayesian estimation, we can define p(sunrise) **even before we get any observations**.
  - Why? Because we use the prior p(θ), which we pick ourselves manually.
- What would be p(sunshine) according to the frequentist approach? Undefined, until we get at least one observation.

Bayesian Estimation, Day 1

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Now, suppose that we observe the sun rise the first day.
  - Let's denote that observation as s_1.
- What is p(θ | s_1)? How do we compute that?

Bayesian Estimation, Day 1

- p_θ(sunrise) = θ.
- We decide to define the prior p(θ) as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Now, suppose that we observe the sun rise the first day.
  - Let's denote that observation as s_1.
- What is p(θ | s_1)? How do we compute that?
- Using Bayes rule:

$$p(\theta|s_1) = \frac{p(s_1|\theta)p(\theta)}{p(s_1)} = \frac{? * ?}{?}$$

Bayesian Estimation, Day 1

- $p_\theta(\text{sunrise}) = \theta$.
- We decide to define the prior $p(\theta)$ as:

$$p(\theta) = \begin{cases} 0, & \text{if } \theta < 0 \\ 1, & \text{if } \theta \epsilon [0,1] \\ 0, & \text{if } \theta > 1 \end{cases}$$

- Now, suppose that we observe the sun rise the first day.
  - Let's denote that observation as s_1.
- What is $p(\theta \mid s_1)$? How do we compute that?
- Using Bayes rule:

$$p(\theta|s_1) = \frac{p(s_1|\theta)p(\theta)}{p(s_1)} = \frac{\theta * 1}{0.5} = 2\theta$$

Bayesian Estimation, Day 1

- Let's denote by s_2 the observation that the sun rises the second day.
- What is $p(s_2 \mid s_1)$? How do we compute that?

$$p(s_2 \mid s_1) = \int_0^1 p_\theta(s_2 \mid s_1)\, p(\theta \mid s_1) d\theta$$

$$= \int_0^1 ? * \; p(\theta \mid s_1) d\theta$$

Bayesian Estimation, Day 1

- Let's denote by s_2 the observation that the sun rises the second day.
- What is $p(s_2 \mid s_1)$? How do we compute that?

$$p(s_2 \mid s_1) = \int_0^1 p_\theta(s_2 \mid s_1)\, p(\theta \mid s_1) d\theta$$

Note: s_2 is conditionally independent of s_1 given θ.

$$= \int_0^1 p_\theta(s_2)\, p(\theta \mid s_1) d\theta$$

$$= \int_0^1 ?$$

Bayesian Estimation, Day 1

- Let's denote by s_2 the observation that the sun rises the second day.
- What is $p(s_2 \mid s_1)$? How do we compute that?

$$p(s_2 \mid s_1) = \int_0^1 p_\theta(s_2 \mid s_1)\, p(\theta \mid s_1) d\theta$$

$$= \int_0^1 p_\theta(s_2)\, p(\theta \mid s_1) d\theta$$

$$= \int_0^1 \theta \, * 2\theta * d\theta = \frac{2}{3}$$

Frequentist Vs. Bayesian Estimation, After One Observation

- As we saw, using Bayesian estimation and assuming uniform prior for θ, after observing the first sunrise, the probability that the sun will rise again the second day is 2/3.
- If we follow the frequentist approach, after observing the first sunrise, what is the probability that the sun will rise again the second day?

Frequentist Vs. Bayesian Estimation, After One Observation

- As we saw, using Bayesian estimation and assuming uniform prior for θ, after observing the first sunrise, the probability that the sun will rise again the second day is 2/3.
- If we follow the frequentist approach, after observing the first sunrise, what is the probability that the sun will rise again the second day?
  - 1, or 100%.
- Which approach seems more intelligent to you?

Frequentist Vs. Bayesian Estimation, After One Observation

- As we saw, using Bayesian estimation and assuming uniform prior for θ, after observing the first sunrise, the probability that the sun will rise again the second day is 2/3.
- If we follow the frequentist approach, after observing the first sunrise, what is the probability that the sun will rise again the second day?
  - 1, or 100%.
- The Bayesian approach is more conservative, and more "intelligent".
  - The Bayesian approach captures the fact that we cannot be certain of the second outcome just because of the first observation.
  - The frequentist approach fails to capture that intuition.

Bayesian Estimation, Day 2

- p_θ(sunrise) = θ.
- Suppose we observe the sun rise both the first day and the second day.
- What is $p(\theta \mid s_1, s_2)$? How do we compute that?
- Using Bayes rule:

$$p(\theta|s_1, s_2) = \frac{p(s_2|\theta, s_1)\, p(\theta \mid s_1)}{p(s_2 \mid s_1)}$$

$$= \frac{? * ?}{?}$$

Bayesian Estimation, Day 2

- p_θ(sunrise) = θ.
- Suppose we observe the sun rise both the first day and the second day.
- What is $p(\theta \mid s_1, s_2)$? How do we compute that?
- Using Bayes rule:

$$p(\theta|s_1, s_2) = \frac{p(s_2|\theta, s_1)\, p(\theta \mid s_1)}{p(s_2 \mid s_1)}$$

$$= \frac{\theta * 2\theta}{\frac{2}{3}} = 3\theta^2$$

Bayesian Estimation, Day 2

- Let's denote by s_3 the observation that the sun rises the third day.
- What is $p(s_3 \mid s_1, s_2)$? How do we compute that?

$$p(s_3 \mid s_1, s_2) = \int_0^1 p_\theta(s_3 \mid s_1, s_2)\, p(\theta \mid s_1, s_2)\, d\theta$$

$$= \int_0^1 ? * p(\theta \mid s_1, s_2)\, d\theta$$

Bayesian Estimation, Day 2

- Let's denote by s_3 the observation that the sun rises the third day.
- What is $p(s_3 \mid s_1, s_2)$? How do we compute that?

$$p(s_3 \mid s_1, s_2) = \int_0^1 p_\theta(s_3 \mid s_1, s_2)\, p(\theta \mid s_1, s_2)\, d\theta$$

Note: s_3 is conditionally independent of s_1 and s_2 given θ.

$$= \int_0^1 p_\theta(s_3)\, p(\theta \mid s_1, s_2)\, d\theta$$

$$= \int_0^1 ? * ? * d\theta$$

Bayesian Estimation, Day 2

- Let's denote by s_3 the observation that the sun rises the third day.
- What is $p(s_3 \mid s_1, s_2)$? How do we compute that?

$$p(s_3 \mid s_1, s_2) = \int_0^1 p_\theta(s_3 \mid s_1, s_2)\, p(\theta \mid s_1, s_2)\, d\theta$$

$$= \int_0^1 p_\theta(s_3)\, p(\theta \mid s_1, s_2)\, d\theta$$

$$= \int_0^1 \theta \, * 3\theta^2 * d\theta = \frac{3}{4}$$

Bayesian Estimation, Day N

- Suppose that have seen the sun rise for the first N days.
- Then, according to Bayesian estimation (and a uniform prior on θ), what the probability that the sun will rise on day N+1?
- If we do the math, it turns out to be $\frac{N+1}{N+2}$.
- Thus, the Bayesian approach will never be 100% certain that the sun will rise again, regardless of how many days we have seen the sun rise in the past.
- Similarly, for the "snow in January" problem. If we have a record of N January days, and it never snowed on those days, the Bayesian estimate is that the probability of snow on the next January day is $\frac{1}{N+2}$.

Frequentist Versus Bayesian Estimation, Recap.

- The frequentist approach is more simple to use.
  - However, the frequentist approach can be unreasonably certain after only one observation, or a few observations, and predict probabilities of 0% or 100%, when such predictions do not make sense.
- The Bayesian approach is more conservative.
  - It can estimate a non-zero probability for outcomes that have never been observed before.
  - It can correctly capture the fact that we cannot give probabilities of 0% or 100% based on just a few observations.
  - It converges to the frequentist approach as we get more observations.
- However, to follow the Bayesian approach:
  - We must define a prior on the parameters we want to estimate.
  - Oftentimes there is no scientific justification for picking a specific prior. We just pick a prior out of the blue.

Frequentist Versus Bayesian Estimation, Recap.

- Philosophers and statisticians break their heads on implications of different priors for various philosophical problems.
- Here is an example on Wikipedia (not relevant for our course, but showing how different choices of priors can lead to different conclusions).

https://en.wikipedia.org/wiki/Doomsday_argument