Chapter 5, Part 1: The Bayesian Paradigm

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

Bayesian Probability is All About Belief

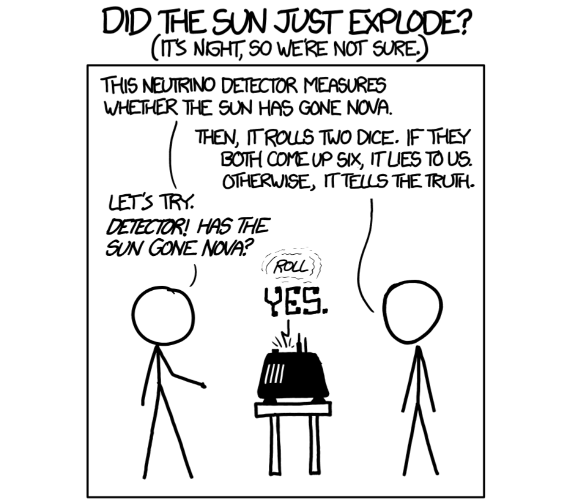
Frequentist Probability

The frequentist interpretation of probability is that it is the **average proportion of the time an event will occur if a trial is repeated infinitely many times**.

Bayesian Probability

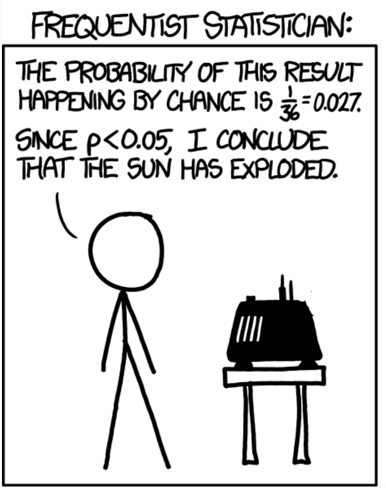
The Bayesian interpretation of probability is that it is the **subjective belief that an event will occur in the presence of incomplete information**

Bayesianism vs Frequentism



https://xkcd.com/1132/

Bayesianism vs Frequentism



https://xkcd.com/1132/

Bayesianism vs Frequentism

https://xkcd.com/1132/

Bayes' Rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Using Bayes' Rule

- Encode initial belief about parameters θ using a **prior** $p(\theta)$
- Characterize how likely different values of θ are to have given rise to observed data $\mathcal{D}$ using a **likelihood function** $p(\mathcal{D}|\theta)$
- Combine these to give **posterior**, $p(\theta|\mathcal{D})$, using **Bayes' rule**:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{1}$$

- This represents our **updated belief** about θ once the information from the data has been incorporated
- Finding the posterior is known as **Bayesian inference**
- $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ is a normalization constant known as the **marginal likelihood** or **model evidence**
- This does not depend on θ so we have

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) \tag{2}$$

Example: Positive COVID Test

We just got a positive COVID test, what is the probability we actually have COVID?

Short answer: it rather depends on why we got tested and the current prevalence of COVID

Note the numbers in this example are not remotely accurate and only used for demonstration

Example: Positive COVID Test from Randomized Testing

- Let $\theta = 1$ denote the scenario where we have COVID and say $1/100$ people in our area currently have COVID
- If we got tested at random we might thus choose to use the prior of $p(\theta = 1) = 1/100$
- Let's assume the test is 95% accurate regardless of whether we have COVID, so $p(\mathcal{D}|\theta = 1) = 0.95, p(\mathcal{D}|\theta = 0) = 0.05$.

Applying Bayes rule:

$$\begin{aligned} p(\theta = 1|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta = 1)p(\theta = 1)}{p(\mathcal{D}|\theta = 1)p(\theta = 1) + p(\mathcal{D}|\theta = 0)p(\theta = 0)} \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \\ &\approx 0.16 \end{aligned}$$

So it seems our chances of having COVID are actually quite low!

Example: Positive COVID Test with Symptoms

- Imagine we now instead go a test specifically because we were showing symptoms
- Let the proportion of such tests being positive be 0.3, so we choose the prior of $p(\theta = 1) = 0.3$

Bayes rule now yields

$$
\begin{aligned}
p(\theta = 1|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta = 1)p(\theta = 1)}{p(\mathcal{D}|\theta = 1)p(\theta = 1) + p(\mathcal{D}|\theta = 0)p(\theta = 0)} \\
&= \frac{0.95 \times 0.3}{0.95 \times 0.3 + 0.05 \times 0.7} \\
&\approx 0.89
\end{aligned}
$$

So now it is extremely likely we have COVID!

Take home: **the prior matters**

Multiple Observations: Using the Posterior as the Prior

- One of the key characteristics of Bayes' rule is that it is **self-similar** under multiple observations
- We can use the posterior after our first observation as the prior when considering the next:

$$\begin{aligned} p(\theta|\mathcal{D}_1, \mathcal{D}_2) &= \frac{p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)}{p(\mathcal{D}_2|\mathcal{D}_1)} \\ &= \frac{p(\mathcal{D}_1, \mathcal{D}_2|\theta)p(\theta)}{p(\mathcal{D}_1, \mathcal{D}_2)} \end{aligned}$$

- We can thinking of this as continuous updating of beliefs as we receive more information

Making Predictions

- Prediction in Bayesian models is done using the **posterior predictive distribution**
- This is defined by taking the expectation of a predictive model for new data, $p(\mathcal{D}^*|\theta)$, with respect to the posterior:

$$p(\mathcal{D}^*|\mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^*|\theta)]. \tag{3}$$

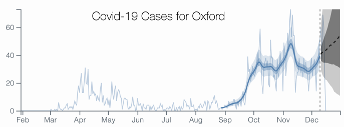
- Note here that we are making the standard assumption that the data is conditionally independent given θ (can in theory use $p(\mathcal{D}^*|\theta, \mathcal{D})$ instead)
- Prediction is often done dependent on an input point such that we actually calculate $p(y|x, \mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}[p(y|x, \theta)]$
- Note that this can be very expensive: typically requires approximations

Why Should we Take a Bayesian Approach?

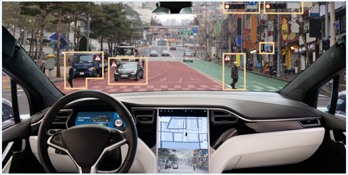
Bayesian Reasoning is the Language of Epistemic Uncertainty

Bayesian reasoning is the basis for how to make decisions with **incomplete information**

Bayesian methods allow us to construct models that return principled **uncertainty estimates**



[Source: `https://localcovid.info/`]



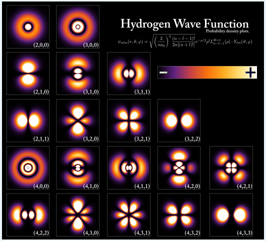
Why Should we Take a Bayesian Approach?

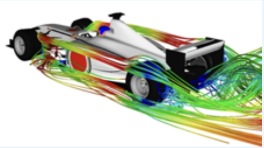
Bayesian Modeling Lets us Utilize Domain Expertise

Bayesian modeling allows us to combine information from data with that from **prior expertise**

Models make clear assumptions and are **explainable**

Bayesian models are often **interpretable**; they can be easily queried, criticized, and built on by humans





Shortfalls [Non Exhaustive]

- Bayesian inference is typically very difficult and expensive: getting around the proportionality constant in Bayes rule is surprisingly challenging
- All models are approximations of the world
  - Constructing accurate models can be very difficult
  - We will always impart incorrect assumptions on our model, particular in our likelihood function
  - For large datasets, the bias from these can usually be avoided by using a powerful discriminative method
- Bayesian reasoning only incorporates uncertainty that is within our model: it does not account for unknown unknowns
  - This can lead to overconfidence
  - Our probabilities/uncertainties are always inherently subjective
- Can struggle to deal with outliers in the data because likelihood terms are multiplicative

Recap

- Bayesian machine learning is a generative approach that allows us to incorporate **uncertainty** and information from **prior expertise**
- Bayes' rule: $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$
- Posterior predictive: $p(\mathcal{D}^*|\mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}\left[p(\mathcal{D}^*|\theta)\right]$

Further Reading

- Additional examples in the notes
- Chapter 1 of C Robert. **The Bayesian choice: from decision-theoretic foundations to computational implementation**. 2007. https://www.researchgate.net/publication/41222434_The_Bayesian_Choice_From_Decision_Theoretic_Foundations_to_Computational_Implementation.
- Michael I Jordan. Are you a Bayesian or a frequentist? Video lecture, 2009. http://videolectures.net/mlss09uk_jordan_bfway/