STAT7630: Bayesian Statistics
Lecture Slides # 2

Chapter 2 - Bayes' Rule

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: August, 2024

Outline

Bayes' Rule for Events (with Illustrative Examples)

Statistics Using Bayes' Rule

An Illustrative Example

Categorizing Online News Items

- The rise of "fake news" has highlighted the need for reliable methods to distinguish real from fake news.
- Bayesian analysis offers a powerful approach to tackling this issue.
- The goal is to classify online news items as either "fake news" or "real news."
- The true nature of an article (fake or real) is not directly observable.
- However, certain observable characteristics of the article can be noted.
- Prior knowledge may provide insight into the frequency of "fake news" articles.

The Dataset and Prior Information

- **Example:** In a dataset of 150 Facebook articles, 60 were identified as "fake news" by experts.
- Assuming this is a representative sample, it informs our prior probability of an article being fake.
- Simple filter: Assume articles are real unless strong evidence suggests otherwise.

Conditional Probability & Exclamation Points

- Data shows exclamation points are more common in fake news. Among the 60 fake news items, 16 have exclamation points in the headline:
- In contrast, only 2 out of 90 real news items have exclamation points.
- Exclamation points can serve as an indicator of whether an article is fake or real.
- We balance prior knowledge with new data to update our understanding using Bayes' Rule.
- This observable characteristic can be considered as data information.
- By combining prior knowledge with this data, we can update the probability of an article being fake.
- The combination of prior and data information yields posterior information about the probability of fake news.

Setting Up a Prior Model

- Let B represent the event that a random news item is fake news.
- Based on prior knowledge, set $P(B) = 0.4$, implying $P(B^c) = 0.6$ for real news.
- i.e., the prior model: Probability that an article is fake $P(B) = 0.4$, real $P(B^c) = 0.6$.
- This is a valid prior, as the probabilities sum to 1 and cover all possible outcomes.

Incorporating Observable Data

- Let A denote the event that a news item's title contains an exclamation point.
- From the data:

$$P(A|B) \approx \frac{16}{60} = 0.2667, \quad P(A|B^c) \approx \frac{2}{90} = 0.0222$$

- $P(A|B)$ is the probability of an exclamation point given the article is fake news.
- Recall that in general, if $P(A|B)$ equals the unconditional probability $P(A)$, events A and B are independent.

Understanding Likelihood

- If event A ("exclamation point") is observed, we can use this to assess the likelihood of event B ("fake news").
- The likelihood function L is defined as $L(B|A) = P(A|B)$ for discrete/categorical cases.
- Note: The likelihood function is not a probability function (e.g., $L(B|A) + L(B^c|A) = 0.2889$, not 1) (Qu: Can you show this in general?).
- The likelihood helps determine how compatible the observed data is with a hypothetical scenario.

Marginal and Joint Probabilities

- The likelihood function is not a valid probability distribution, but marginal probability $P(A)$ can be used as a normalizing constant.
- Joint probability $P(A \cap B)$ represents the probability of both events A and B occurring.
- Example:

$$P(A \cap B) = P(A|B)P(B) = 0.2667 \times 0.4 = 0.1067$$

- Similarly,

$$P(A \cap B^c) = P(A|B^c)P(B^c) = 0.0222 \times 0.6 = 0.0133$$

Why is $P(A)$ a Normalizing Constant?

- **Bayes' Rule:**

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

- **Normalization:**
  - The numerator $P(B) \times P(A|B) = P(A \cap B)$ gives an unnormalized probability (i.e. $P(A \cap B) + P(A \cap B^c) = P(A)$ for events B and B^c).
  - Dividing by $P(A)$ scales the posterior $P(B|A)$ so that the total probability across all B (restricted to A) sums to 1.
- **Conclusion:** $P(A)$ ensures $P(B|A)$ is a valid probability distribution (i.e. $P(A \cap B)/P(A) + P(A \cap B^c)/P(A) = P(A)/P(A) = 1$ for events B and B^c), making it the **normalizing constant** in Bayes' Rule.

Law of Total Probability

- The total probability that an article has an exclamation point is the sum of:
  - The probability that the article has an exclamation point and is fake.
  - The probability that the article has an exclamation point and is real.
- Thus,

$$P(A) = P(A \cap B) + P(A \cap B^c) = 0.1067 + 0.0133 = 0.12$$

- This is an example of the Law of Total Probability (LTP).

Bayes' Rule & Posterior Probability

- The key question: Given that an article's title has an exclamation point, what is the probability it is fake news?
- This is calculated as $P(B|A)$.
- Bayes' Rule for events is expressed as:

$$\begin{aligned} P(B|A) &= \frac{P(B) \times P(A|B)}{P(B) \times P(A|B) + P(B^c) \times P(A|B^c)} \\ &= \frac{P(B) \times L(B|A)}{P(B) \times L(B|A) + P(B^c) \times L(B^c|A)} \end{aligned}$$

- In words: Posterior = (Prior) $\times$ (Likelihood) / (Normalizing constant).

Application of Bayes' Rule: News Example

- Applying Bayes' Rule:

$$P(B|A) = \frac{(0.4) \times (0.2667)}{0.12} = 0.889$$

- Given an article with an exclamation point in the title, the probability it is fake news is 0.889.
- Prior to observing the exclamation point, the probability was 0.4.
- The observed data has updated our estimate.

Simulation and Model Validation

- Simulate 10,000 articles to validate the model and understand the distribution of fake vs. real articles.
- Simulation reflects the prior model and likelihood of exclamation point usage.
- Results: Approximate posterior probability of an article being fake when it uses exclamation points is around 88.7%.

Another Bayes' Rule Example: 1975 UK Referendum

- Context: 1975 UK national referendum on remaining in the EEC.
- Suppose 52% of voters supported the Labour Party, and 48% the Conservative Party. (i.e. voters are assumed to belong to either the Labour Party or the Conservative Party.)
- 55% of Labour voters supported remaining in the EEC, while 85% of Conservative voters supported it.
- What is the probability that a person voting "Yes" to remaining in the EEC is a Labour voter?

$$P(L|Y) = \frac{P(Y|L) \times P(L)}{P(Y)}$$

Example Continued

- Note that:

$$P(Y) = P(Y \cap L) + P(Y \cap L^c) = P(Y|L)P(L) + P(Y|L^c)P(L^c)$$

- So:

$$P(L|Y) = \frac{(0.55) \times (0.52)}{(0.55) \times (0.52) + (0.85) \times (0.48)} = 0.41$$

Bayes' Rule for Multiple Events

- Let $\mathbf{D}$ represent observed data, and A, B, and C be mutually exclusive (and exhaustive) events.
- We can express $P(\mathbf{D})$ as:

$$\begin{aligned} P(\mathbf{D}) &= P(\mathbf{D} \cap A) + P(\mathbf{D} \cap B) + P(\mathbf{D} \cap C) \\ &= P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C) \end{aligned}$$

- By Bayes' Rule:

$$P(A|\mathbf{D}) = \frac{P(\mathbf{D}|A)P(A)}{P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C)}$$

- $P(B|\mathbf{D})$ and $P(C|\mathbf{D})$ are similar.

Generalizing Bayes' Rule

- Denoting k events $A, B, C, \ldots,$ as $\theta_1, \theta_2, \theta_3, \ldots, \theta_k$, we generalize as:

$$P(\theta_i|\mathbf{D}) = \frac{P(\theta_i)P(\mathbf{D}|\theta_i)}{\sum_{j=1}^{k} P(\theta_j)P(\mathbf{D}|\theta_j)}$$

- The denominator equals $P(\mathbf{D})$, the marginal distribution of the data.
- For continuous θ, the sum may be replaced by an integral.

Example: General Social Survey

- In the 1996 General Social Survey, for males (age 30+):
  - 11% of those in the lowest income quartile were college graduates.
  - 19% of those in the second-lowest income quartile were college graduates.
  - 31% of those in the third-lowest income quartile were college graduates.
  - 53% of those in the highest income quartile were college graduates.

Example: General Social Survey

- What is the probability that a college graduate falls in the lowest income quartile?

$$\begin{aligned}
P(Q_1 \mid G) &= \frac{P(G \mid Q_1)P(Q_1)}{\sum_{j=1}^{4} P(G \mid Q_j)P(Q_j)} \\
&= \frac{(.11)(.25)}{(.11)(.25) + (.19)(.25) + (.31)(.25) + (.53)(.25)} \\
&= 0.09
\end{aligned}$$

- **Exercise:**
- Find $P(Q_2|G)$, $P(Q_3|G)$, and $P(Q_4|G)$ as well.
- How does this conditional distribution differ from the unconditional distribution $\{P(Q_1), P(Q_2), P(Q_3), P(Q_4)\}$?

Bayes' Rule Applied to Regional Dialects

- Example: Use of the term “pop” to infer the region of a speaker in the U.S.
- Prior information: Regional population distribution.
- Likelihood: Probability of using “pop” in different regions.
- Posterior: Updated probability of the speaker's region after hearing “pop.”

Outline

Bayes' Rule for Events (with Illustrative Examples)

Statistics Using Bayes' Rule

Inference About Parameters

- We consider inference about parameters based on observed data.
- Let θ represent an unobserved parameter of interest, and $\mathbf{D}$ represent the observed data.
- The probability model for the data, given θ, is denoted $p(\mathbf{D}|\theta)$.
- The prior knowledge about θ is denoted $p(\theta)$.
- This prior can be highly specific or quite vague.

Posterior Distribution

- We seek to make probability statements about θ, given the observed data $\mathbf{D}$: $p(\theta|\mathbf{D})$.
- By Bayes' Rule:

$$p(\theta|\mathbf{D}) = \frac{p(\theta)p(\mathbf{D}|\theta)}{p(\mathbf{D})}$$

- Note $p(\mathbf{D})$ does not depend on θ and is merely a **normalizing constant**.
- For inference about θ, we can write:

$$p(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta)$$

Summarizing the Posterior

- The **posterior distribution** $p(\theta|\mathbf{D})$ represents a compromise between prior information $p(\theta)$ and sample information $p(\mathbf{D}|\theta)$.
- Useful summaries of the posterior include:
  - Posterior mean:

$$E[\theta|\mathbf{D}] = \int \theta p(\theta|\mathbf{D})\, d\theta$$

  - Posterior variance:

$$\mathrm{Var}[\theta|\mathbf{D}] = \int (\theta - E[\theta|\mathbf{D}])^2 p(\theta|\mathbf{D})\, d\theta$$

Posterior Probability in Chess Example

- Analyze Kasparov's chances of winning against Deep Blue in 1997.
- Prior model: Kasparov's win probability π could be 0.2, 0.5, or 0.8.
- Assume the number of games Kasparov wins, Y, out of 6 games follows a Binomial(6, π) distribution.
- After observing one win out of six games, the likelihood strongly suggests $\pi = 0.2$.
- Posterior model confirms Kasparov is likely the weaker player.
- A more realistic analysis would spread the prior distribution for π over the entire interval from 0 to 1.
- We will explore such models in the next chapter.

Likelihood values in Chess Example

- After observing one win out of six games (i.e. Data is $y = 1$), the likelihood for each π is:

$$L(\pi|y = 1) = \binom{6}{1}\pi^1(1-\pi)^5$$

$$L(0.2|y = 1) = \binom{6}{1}(0.2)^1(0.8)^5 \approx 0.3932$$

$$L(0.5|y = 1) = \binom{6}{1}(0.5)^1(0.5)^5 \approx 0.0938$$

$$L(0.8|y = 1) = \binom{6}{1}(0.8)^1(0.2)^5 \approx 0.0013$$

Posterior Probabilities in Chess Example

- Posterior probabilities are proportional to $P(\pi) \times L(\pi|y = 1)$.
- Assume a uniform prior, $P(\pi = 0.2) = P(\pi = 0.5) = P(\pi = 0.8) = \frac{1}{3}$.
- Posterior for $\pi = 0.2$ is:

$$P(\pi = 0.2|y = 1) =$$
$$(P(\pi = 0.2) \times L(0.2|y = 1))/(P(\pi = 0.2) \times L(0.2|y = 1)+$$
$$P(\pi = 0.5) \times L(0.5|y = 1) + P(\pi = 0.8) \times L(0.8|y = 1)) =$$
$$\frac{0.3932 \times \frac{1}{3}}{0.3932 \times \frac{1}{3} + 0.0938 \times \frac{1}{3} + 0.0013 \times \frac{1}{3}} \approx 0.799$$

- Similarly, $P(\pi = 0.5|y = 1) \approx 0.191$, $P(\pi = 0.8|y = 1) \approx 0.010$.

Summary of Chapter 2

- Construct prior models for the variable of interest.
- Summarize data dependence via conditional probability.
- Define likelihood functions based on observed data.
- Use **Bayes' Rule to balance prior and likelihood to form the posterior model**.
- Simulation helps to validate and understand Bayesian models.