Term Paper

"How Facebook applies Probability and Statistics to empower its features"

01006719 PROBABILITY AND STATISTICS 1

Software Engineering Program

Faculty of Engineering, KMITL

By

65011277 Chanasorn Howattanakulphong

Table of Contents

Introduction :

In today's digital age, Facebook has become synonymous with social media and interconnectedness, serving as a platform that transcends geographical boundaries. What often escapes the spotlight, however, is the intricate fusion of probability and statistics that underpins Facebook's most prominent features. This article embarks on a journey through the realms of mathematical modeling and data-driven decision-making to unveil the ways in which Facebook utilizes probability and statistics to empower its platform. Our primary focus will be on the world of targeted advertising, where Facebook's expertise in probability modeling, and Bayesian bandit's paradigm converges to offer a personalized experience like no other.

But that's just the tip of the iceberg. Beyond personalized ads, Facebook applies statistical methods to address challenges that extend to platform safety. Statistical thresholds are employed to swiftly identify and combat suspicious or malicious activities, safeguarding the integrity of the platform.

As we delve deeper into the realm of statistical algorithms, we'll explore how Facebook leverages segmentation to cluster users based on their interests and behaviors. This, in turn, empowers advertisers to effectively target specific demographics, resulting in more relevant and engaging ads for users. Facebook's innovative use of statistical algorithm and time series analysis enhances accuracy and personalization.

In the context of the ever-scrolling news feed, we'll examine Facebook's application of statistical algorithms and probability distributions, like the Poisson distribution, to predict user interactions with posts within a specific timeframe. Furthermore, we'll shine a light on the role of time series analysis in dissecting user activity data to identify trends, seasonality, and irregular patterns, optimizing the timing of content delivery to users.

In conclusion, Facebook's mastery of probability and statistical techniques extends far beyond mere calculations. These mathematical foundations play a pivotal role in shaping the platform's targeted advertising strategies, enhancing user experiences, and fortifying its security measures. By continually refining and expanding its data-driven capabilities, Facebook remains at the forefront of personalized digital advertising and content delivery, ensuring that its features empower users and enrich their interactions on the platform.

Targeted Advertising:

Facebook, like many other online advertising platforms, uses a variety of techniques and algorithms for targeted advertising. Among these techniques I'm showing probability modeling and how they could apply Bayesian bandit's paradigm. Here's an explanation of how Facebook uses these methods:

1.Probability Modeling:

In probability modeling, one of the concepts Facebook uses is **conditional probability**. It is used to estimate the likelihood of a user taking a specific action in response to an ad.

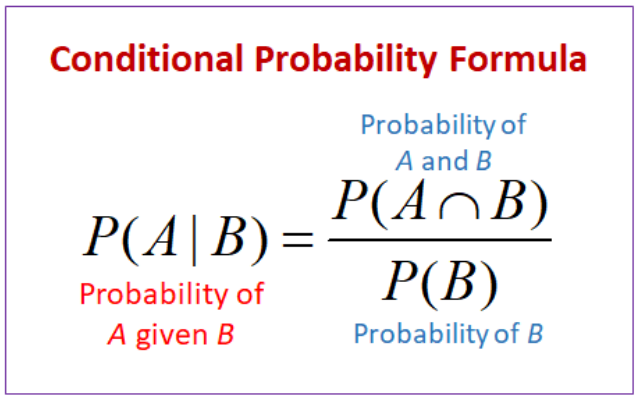

Figure1 – Conditional Probability Formula

For example, the conditional probability P(User clicks on ad | User is shown the ad) represents the probability that a user clicks on an ad given that the user has been shown that ad.

Suppose Facebook wants to estimate the probability that a user will click on a specific ad when the user is shown the ad. They are interested in finding P(User clicks on ad | User is shown the ad).

Formulas and Definitions:

Conditional Probability (P(A | B)): Conditional probability is defined as the probability of event A occurring given that event B has already occurred.

To calculate P(User clicks on ad | User is shown the ad), Facebook would need to gather data on two probabilities:

a. P(User clicks on ad and User is shown the ad): The probability that a user both clicks on the ad and is shown the ad.

b. P(User is shown the ad): The probability that a user is shown the ad. To obtain these probabilities, Facebook collects and analyzes large amounts of user data. They track how often users are shown specific ads and how often users click on those ads.

Calculation Example:

Suppose Facebook has collected data and found the following probabilities:

- P(User clicks on ad and User is shown the ad) = 0.10 (10% of the time a user both clicks on the ad and is shown the ad).
- P(User is shown the ad) = 0.30 (30% of the time a user is shown the ad).

Now, you want to calculate **P(User clicks on ad | User is shown the ad)**, which represents the likelihood of a user clicking on a specific ad when they are shown that ad.

P(User clicks on ad | User is shown the ad)

= P(User clicks on ad and User is shown the ad) / P(User is shown the ad)

= 0.10 / 0.30 = 1/3 ≈ 0.3333

So, the conditional probability that a user clicks on a specific ad when they are shown that ad is approximately 0.3333, or 33.33%.

2.The Bayesian Bandits Paradigm:

Facebook's advertising algorithm applies this concept to the selection and prioritization of ads for individual users, aiming to maximize revenue. The challenge Facebook faces is akin to a gambler deciding which slot machine to play to win the most money, with each ad being one of the slot machines.

The Bayesian Bandits paradigm is a powerful approach to decision-making under uncertainty. It combines Bayesian probability theory with the concept of "bandits," which refers to problems involving selecting options with different payouts, much like playing various slot machines.

Application in Facebook Ads

Facebook could employ a straightforward A/B testing method, randomly showing different ads to users and selecting the one with the best performance. However, this approach has several drawbacks, including the time and wasted initial spend. It also fails to address the probabilistic nature of real-life scenarios.

The Bayesian Bandits approach addresses these issues by utilizing prior information to optimize ad impressions.

Instead of relying solely on A/B testing, the algorithm begins by assuming that it doesn't know the expected performance (e.g., purchases per thousand impressions - PPM) of each ad. It displays all ads to a random selection of users and measures their purchases. When one ad outperforms another in terms of purchases, it infers that the winning ad has a higher probability of having a higher PPM. The algorithm then assigns more impressions to the winning ad, thus optimizing ad spend.

Bayesian Bandit Solution:

The Bayesian Bandit solution involves using Bayes' rule to make informed decisions.

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Figure2 – Bayes Theorem Formula

We calculate the probability of each ads winning (θ) based on the data (x) and our current beliefs (prior, p(θ)). This allows us to give more chances with higher win rates and higher confidence in those estimates.

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

Figure3 – Bayes Theorem Formula with our meaning

Bayesian Bandits often use the Beta distribution as the prior because of its convenient properties. It can represent probabilities between 0 and 1, making it suitable for applications like click-through rates. When the prior is a Beta distribution and the likelihood is a Bernoulli distribution (binary outcomes), the posterior is also a Beta distribution.

The Beta distribution has 2 parameters, a and b, which govern its shape. We refer to a specific distribution as Beta(a,b)

To update the posterior distribution with new data, we use the parameters a and b. Initially, a = 1 and b = 1, creating a minimally informative prior. When we pull an ads and observe a result (x), we update the parameters as follows: a' = a + x and b' = b + 1 - x. This results in the updated distribution Beta(a + x, b + 1 - x).

Case Example:

Imagine an advertiser is running an ad campaign with three different ads: red, blue, and green. They decide to test these ads in two scenarios, one with 20 trials and another with 100 trials. Each trial represents showing one of these ads to a user and measuring their response.

Graph 1 (20 Trials):

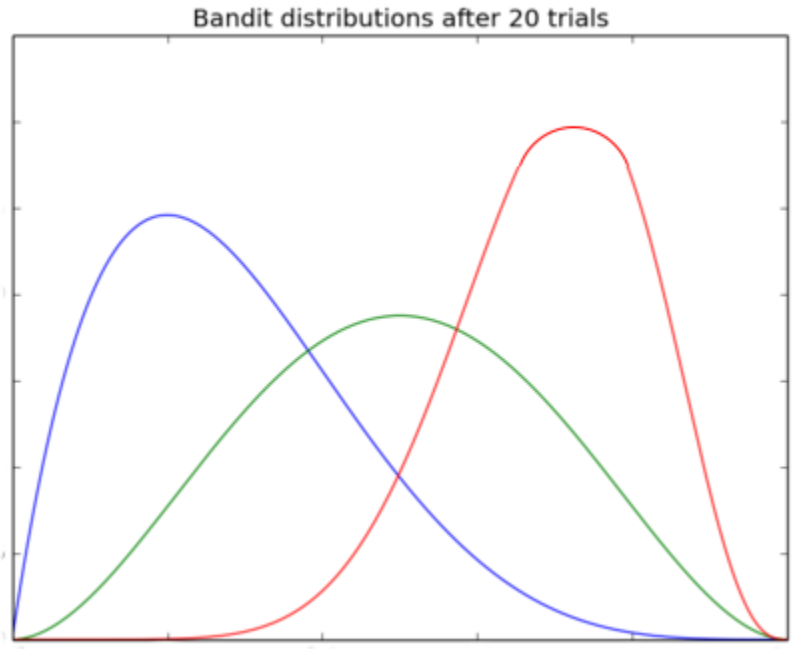

Figure4 – Bayesian Bandits Distribution 20 trials

In the first 20 trials, the results are observed. Blue initially seems to perform well then the performance gradually drops. Red, on the other hand, didn't perform as well initially but displayed a bell-shaped curve with the peak to the right. Green, while somewhat normal in distribution, doesn't have the same high probability as red.

Blue's early success might make it seem like a good choice. However, Bayesian Bandits take more than just initial performance into account. Even with

20 trials, the algorithm recognizes that users are showing more interest in red over time.

Graph 2 (100 Trials):

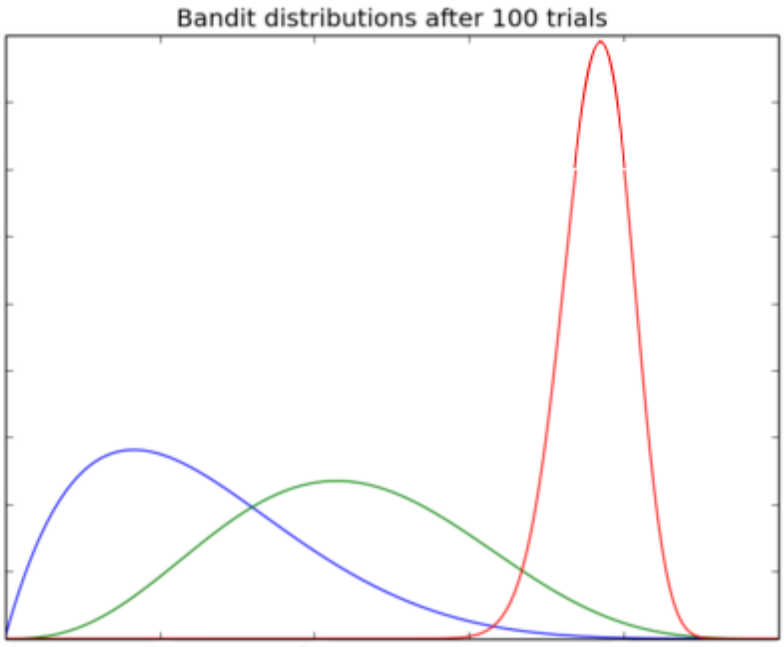

Figure5 – Bayesian Bandits Distribution 100 trials

In the second scenario with 100 trials, the same three ads are tested again. Now, the shape of the red curve becomes much sharper and taller compared to blue and green. With 100 trials, the Bayesian Bandits algorithm gains more confidence in red as a choice that users are increasingly engaging with. The sharp, tall curve for red in Graph 2 indicates that it's becoming more evident that red is the preferred choice for the users.

Platform Safety:

Imagine, Facebook is using statistical hypothesis testing and the concept of the z-score to identify accounts engaging in suspicious behavior, such as **spamming** by analyzing user behaviors, including the frequency, and timing of posts.

Facebook will need to collect data on the number of posts made by each user over a specified time period to plot the threshold of normal content posting. If the user posts more than the threshold, it would be considered suspicious and facebook might want to look on the case.

To calculate the threshold, we use the Z-score formula:

$$z = \frac{x - \mu}{\sigma}$$

Figure6 – Z-score formula

Where:

- Z is the z-score.
- X is the data point you want to measure.
- μ (mean) is the average or expected value.
- σ (standard deviation) is the measure of variability in the data.

To set a threshold for identifying unusual data points, you can rearrange the z-score formula as follows:

X = μ + (Z * σ)

In this rearranged formula,

"X" represents the threshold value, which is calculated based on the mean (μ), the z-score (Z), and the standard deviation (σ). When you use a multiplier (e.g., 2) as Z,

An example case, Detecting Unusual Friend Requests, would contain data:

Mean (μ): 3 friend requests per week.

Standard Deviation (σ): 1 friend requests per week.

Define a threshold, such as two standard deviations above the mean($\mu + 2\sigma$). In this case, the threshold would be 3 (μ) + 2 * 1 (σ) = 5 friend requests per week.

Next, Facebook calculate the Z-score of each users and compare it with the threshold value.

Provided are two users with different data:

User A:

- Number of Friend Requests (X): 22
- Mean (μ): 10
- Standard Deviation (σ): 3
- Z-Score (Z) = (22 - 10) / 3 = 4

User A's z-score is 4, which is below the threshold of 5, indicating that User A's friend request behavior is typical and does not exceed the threshold.

User B:

- Number of Friend Requests (X): 10
- Mean (μ): 3
- Standard Deviation (σ): 1
- Z-Score (Z) = (10 - 3) / 1 = 7

User B on the other hand, has z-score of 7, which is higher than the threshold of 5, indicating that User B's friend request unusual behavior is suspicious and potentially indicative since it exceeds the threshold.

*Note that unrealistic data was used for simplifying calculations. Real data will be more reasonable.

News Feed Prediction

1. Statistical Algorithms:

Facebook uses statistical algorithms to predict and rank content in users' News Feeds. The goal is to show users the most relevant and engaging content based on their preferences and interactions.

Modeling User Engagement:

Modeling user engagement in social media platforms like Facebook often involves using probability distributions to estimate the likelihood of certain user actions, such as likes, comments, shares, or other interactions with posts

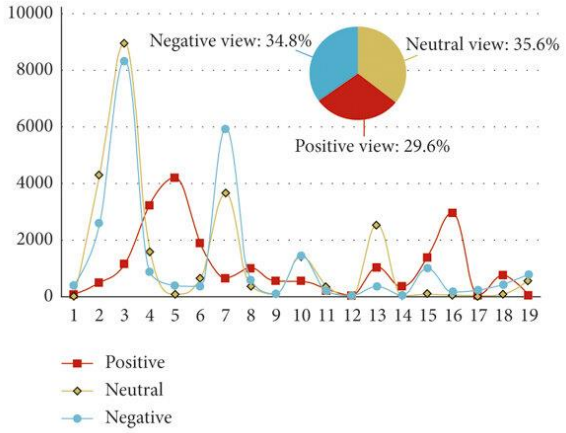

Figure7 – User interactivity distribution

The followings are basic data users could access. This means they collects so much data to work on probability modelling user engagements

- Reach (overall)
- Reach: Organic/Paid
- Impressions: Organic/Paid
- Reach: Fans/Non-Fans

- Post Clicks/Reactions, Comments & Shares
- Reactions/Comments/Shares
- Post Hides, Hides of All Posts, Reports of Spam, Unlikes of Page
- Engagement Rate

Poisson probability is used to model the likelihood of a user engaging with a post, such as liking or commenting on it. Each interaction event (e.g., a user liking a post) is considered a discrete event, and Poisson distribution is applied to model the rate at which these events occur.

Poisson distribution:

The probability mass function (PMF) of the Poisson distribution is defined as:

$$P_X(k) = \frac{e^{-\lambda} * \lambda^k}{k!}$$

Figure8 – PMF of the Poisson distribution

Where:

- P(X = k) is the probability of observing exactly k events (interactions) in a given time frame.
- λ (lambda) is the average rate of events per unit of time or space.
- e is Euler's number, approximately equal to 2.71828.
- k is the actual count of events, and k! denotes the factorial of k.

This PMF is used to calculate the probability of observing a specific number of events (k) in a Poisson process, where events occur randomly in time or space, and the average rate of events per unit of time or space is given by λ. It's a fundamental formula for modeling the distribution of rare, independent events, which is often

applied in various fields, including social media analytics for predicting user engagement.

Here's an example on how Poisson distribution works:

Facebook aims to estimate the likelihood of a user posting a comment on a post within the next 10 minutes. They have noticed that, on average, a post receives 2 comments every 30 minutes. To adapt this average to a 10-minute rate, we calculate that it's approximately 2/3 comments every 10 minutes.

Now, let's compute the probability of observing 1 comment (k = 1) in 10 minutes using the Poisson Probability Mass Function (PMF):

λ = (Average rate in the original time interval) * (New time interval)

λ = (2 comments / 30 minutes) * 10 minutes = **2/3** comments / 10 minutes.

Poisson PMF formula:

$P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$,

which translates to this when applying our values:

$P(X = 1) = (e^{(-2/3)} * (2/3)^1) / 1!$,

resulting in an approximate probability of 51.32%.

Interpretation:

The probability, approximately 51.32%, represents the likelihood of observing 1 comment in the next 10-minute interval based on the historical data and the estimated comment rate. Which means there is about a 51.32% chance of seeing 1 comment in the upcoming 10 minutes.

2.Time Series Analysis:

Facebook can use time series analysis to predict and optimize news feed content to provide a more engaging and relevant experience for its users. Time series analysis involves analyzing data points collected over time to identify patterns, trends, and make predictions.

Pattern Identification:

Facebook applies time series analysis techniques to this data to identify patterns and trends. They may use techniques like moving averages, autoregressive models, or more sophisticated machine learning algorithms to uncover temporal patterns.

Seasonal and Daily Trends:

Time series analysis can reveal weekly or daily patterns in user activity. For example, there may be a consistent peak in user engagement during weekends or specific hours during weekdays.

Predictive Models:

Facebook uses the insights gained from time series analysis to build predictive models. These models forecast when users are likely to be most active on the platform based on historical data. The forecasts could include both long-term trends and short-term variations.

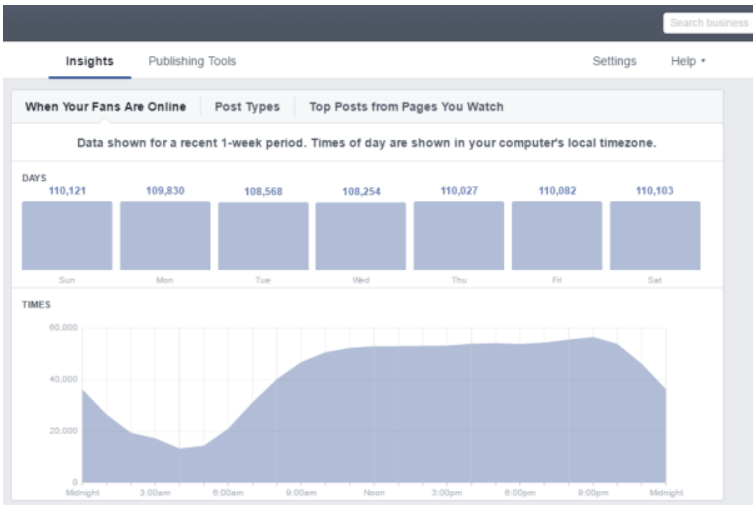

Figure9 – Insights on user online time

Combining with the comment estimation example stated earlier, but this time we have a random set of comment counts for a Facebook post within each time interval

- 12:00 AM - 12:15 AM: 5 comments
- 12:15 AM - 12:30 AM: 4 comments
- 12:30 AM - 12:45 AM: 7 comments
- 12:45 AM - 1:00 AM: 3 comments
- 1:00 AM - 1:15 AM: 6 comments
- 1:15 AM - 1:30 AM: 8 comments
- 1:30 AM - 1:45 AM: 4 comments
- 1:45 AM - 2:00 AM: 6 comments
- ...
- 11:30 PM - 11:45 PM: 9 comments
- 11:45 PM - 12:00 AM: 5 comments

In this case, we have 96 intervals in a 24-hour period (24 hours x 4 intervals per hour). These numbers represent the counts of comments on the Facebook post within each 15-minute interval.

We can now perform a Poisson distribution analysis on this data. Let's assume you want to estimate the average comment rate (λ) for each 15-minute interval. You would calculate λ for each interval based on the observed comment counts.

For example, in the 12:00 AM - 12:15 AM interval, the λ could be estimated as 5 comments. You can then use the Poisson distribution to model the expected number of comments for each interval, make predictions, and optimize engagement strategies based on the findings.

To estimate the number of comments for the next time interval, you can use the observed value of λ for that interval, and plug it into the formula with the desired value of k (the number of comments you want to estimate). The result will give you the probability of observing k comments in that interval.

For example, if you observed that the average rate of comments (λ) for the 12:00 AM - 12:15 AM interval was 5, and you want to estimate the probability of getting 6 comments in the next interval (12:15 AM - 12:30 AM), you would use the Poisson PMF formula with λ=5 and k=6:

Poisson PMF formula:

$P(X = k) = (e^{(-\lambda)} * \lambda^k) / k!$

$P(X=6)= (e^{(-5)} * 5^6) / 6! \approx$ **0.14622**

Interpretation:

The calculated probability, approximately 0.14622 (or 14.622%), represents the likelihood of observing exactly 6 comments in the next 10 minutes.

Conclusion :

In summary, Facebook's profound reliance on probability and statistical techniques extends its influence beyond mere calculations. These mathematical foundations are fundamental to the platform's success. We've witnessed their impact on personalized advertising, user engagement, and security. Facebook's use of probability modeling and Bayesian bandits results in a uniquely personalized advertising experience. Additionally, statistical algorithms and time series analysis empower advertisers to precisely target demographics and optimize user experiences. The application of probability distributions and time series analysis also refines content delivery timing. Facebook's continual refinement of data-driven capabilities keeps it at the forefront of personalized digital advertising and content delivery, shaping the future of social media and interconnectedness.

References :

1. Lada, A., Lada, A., Wang, M., & Yan, T. (2022, May 25). How machine learning powers facebook's news feed ranking algorithm. Engineering at Meta. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/
2. Platform Safety | Working to help keep our technologies safe for all people and communities. Facebook. (n.d.). https://www.facebook.com/business/platform-safety
3. Toward fairness in personalized ads. (n.d.). https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf
4. Cárdenas-Rodríguez, J. (2021, October 23). Data Science with the Penguins data set: Conditional probability. Medium. https://jdatascientist.medium.com/data-science-with-the-penguins-data-set-conditional-propability-bd998bfedd35
5. Bayesian bandit tutorial - lazy programmer. (n.d.-a). https://lazyprogrammer.me/bayesian-bandit-tutorial/
6. Lada, A., Lada, A., Wang, M., & Yan, T. (2022, May 25). How machine learning powers facebook's news feed ranking algorithm. Engineering at Meta. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/
7. Getting to know the Poisson process and the Poisson probability distribution. Time Series Analysis, Regression, and Forecasting. (2022, May 3). https://timeseriesreasoning.com/contents/poisson-process/
8. Collier, A. (2022, August 30). Find the best time and content to post on Facebook with facebook insights. Constant Contact. https://www.constantcontact.com/blog/new-facebook-insights/
9. Z-score table: Definition and its types, Z-score formula. Toppr. (2020, June 16). https://www.toppr.com/guides/maths/z-score-table/