A Personal Paper on Statistics and its Real World Applications

Matthew Balogh

Table of contents

Probability

Conditional Probability

Conditional probability is the probability of event A, given that event B has occured.

Definition 0.1. *Bayes theorem*, describing the relationship between dependent events A, and B.

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

Table 1: Parts of the *Bayes formula.* [1]

Part	Notation	Description
posterior probability	$P(A|B)$	Probability of A, given that B occured
prior probability	$P(A)$	Probability of A alone
likelihood	$P(B|A)$	Probability of B, given that A occured
marginal likelihood	$P(B)$	Probability of B alone

Definition 0.1 can be interpreted as if we want to determine the probability of an event Q in light of another event, then we can calculate it by examining the independent probabilities of the two events and the probability of the other event given that the event Q in question has occured.

> **Tip**
>
> See Example 0.1 in the Appendix section for a practical example on this topic.

Classification

Naïve Bayes

Naïve Bayes classification relies on the method that describes the probability of an event in light of additional information.

The classification is based on the following question:

> "Based on prior evidence, what is the most likely class of a new unlabeled instance?" [1]

To make use of prior evidence, the method utilizes the concept of the *conditional probability.*

Definition 0.1. *Bayes formula* for conditional probability.

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

Definition 0.2. Conditional probability with multiple predictors, describing the relationship between dependent events A and S, where S is an n-term set of independent events based on the *Naïve Bayes* assumption.

$$P(A|S) = \frac{P(A) \times P(s_1|A) \times P(s_2|A) \times \ldots \times P(s_n|A)}{P(s_1, s_2, \ldots, s_n)}$$

> Tip
>
> See Example 0.2 in the Appendix section for a practical example on this topic.

Concepts

Logarithmic scale

Logarithmic scale or simply **log scale** is used to display data that would span a broad range of values on a *linear scale*, especially when there are significant magnitude differences between the individual data points. [2]

While on a *linear scale*, each unit corresponds to the same increment on the scale, on a *logarithmic scale* each unit corresponds to a multiple of the base value and each subsequent unit is the multiplication of the previous one using the base value. [2]

Semi-log plot

A **semi-log plot** or graph has one axis on a logarithmic and one axis on a linear scale. [3]

Log-log plot

A **log-log plot** or graph has both its abscissa and ordinate in a logarithmic scale.

> **Note**
>
> A semi-log scaled plot can help in the following:
>
> - shrink the data points into a smaller area
> - draw the best-fit line if the original data follows an exponential trend

Tip

See Example 0.3 in the Appendix section for a practical example on this topic.

Appendix

Example 0.1. Imagine a bus station with its predefined schedule table. We want to determine whether the bus departing at 8 AM from the nearest station will be late if it's raining when we wake up. With this information, we might consider taking the metro, which is only a few minutes farther than the bus station.

The bus company has a database with historical records of routes, each of which indicates whether the bus reached a certain station in time or not. We also have another dataset that contains weather conditions.

Projecting the dataset to the bus route and station in question and merging with weather data, we can determine the *posterior probability* in question, that is:

$$P(A|B) = P(bus\ is\ late \mid rain\ in\ the\ morning)$$

The *prior probability*, that is $P(A) = P(bus\ is\ late)$, is the proportion of records, in which the bus had been marked as **late** to the station in any given occasion in the past.

The *likelihood*, that is $P(B|A) = P(rain\ in\ the\ morning \mid bus\ is\ late)$, is the proportion of records, in which the morning weather had been marked as **rain** considering only the records, in which the bus had been marked as **late** in any given occasion in the past.

The *marginal likelihood*, that is $P(B) = P(rain\ in\ the\ morning)$, is the proportion of records, in which the morning weather had been marked as **rain** in any given occasion in the past.

The values of the parts of the equation - based on Table 2-Table 5 - are the following:

```
                     Value
Prior probability   0.3000
Likelihood          0.6667
Marginal likelihood 0.3833
```

Expressed with the *Bayes formula*:

$$
\begin{aligned}
P(A|B) = P(bus\ is\ late \mid rain\ in\ the\ morning) \\
= \frac{\frac{18}{60} \times \frac{12}{18}}{\frac{23}{60}} \\
= 0.5217
\end{aligned}
$$

That is, the probability that the bus will be late if it rains in the morning is around **52%**.

> **Warning**
>
> This example is solely based on the small sixty-record sample of Table 2, to demonstrate the determination of different parts of the *Bayes formula*.

Example 0.2. Consider the scenario from Example 0.1 with a modification to the condition. Instead of determining whether the bus will be late if it rains in the morning, we are interested in whether the bus will be late if it rains on a winter day.

The dataset is now labeled with seasonal information based on the date of the records.

The *posterior probability* in question is denoted as:

$$P(A|S) = \frac{P(A) \times P(s_1|A) \times P(s_2|A)}{P(s_1, s_2)}$$

$$\begin{aligned} A &= \textit{bus is late} \\ S &= \textit{rain in the morning on a winter day} \\ s_1 &= \textit{rain in the morning} \\ s_2 &= \textit{in winter} \end{aligned}$$

The *prior probability*, that is $P(A) = P(\textit{bus is late})$, is the proportion of records, in which the bus had been marked as **late** to the station in any given occasion in the past.

The *likelihood* for s_1, that is $P(s_1|A) = P(\textit{rain in the morning} \mid \textit{bus is late})$, is the proportion of records, in which the morning weather had been marked as **rain** considering only the records, in which the bus had been marked as **late** in any given occasion in the past.

The *likelihood* for s_2, that is $P(s_2|A) = P(\textit{in winter} \mid \textit{bus is late})$, is the proportion of records, in which the season is **winter** considering only the records, in which the bus had been marked as **late** in any given occasion in the past.

The *marginal likelihood*,
that is $P(s_1, s_2) = P(\textit{rain in the morning on a winter day})$, is the proportion of records, in which the morning weather had been marked as **rain** and the season as **winter** in any given occasion in the past.

The values of the parts of the equation - based on Table 2-Table 4, and Table 6 - are the following:

```
                      Value
Prior probability   0.3000
Likelihood (s1)     0.6667
Likelihood (s2)     0.5000
Marginal likelihood 0.1500
```

Expressed with the *Bayes formula*:

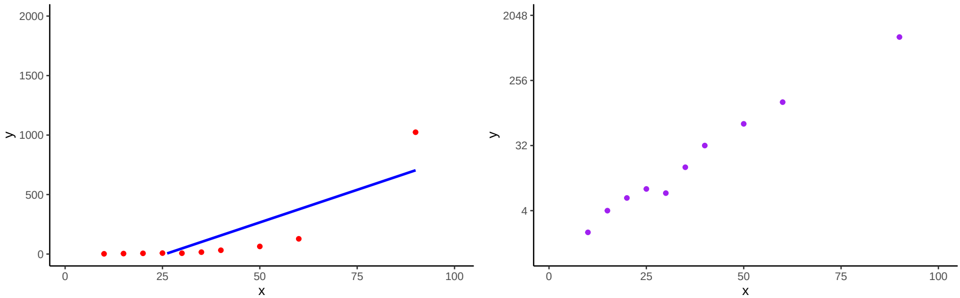
$$
\begin{aligned}
P(A|S) &= \frac{\frac{18}{60} \times \frac{12}{18} \times \frac{9}{18}}{\frac{9}{60}} \\
&= \frac{0.3 \times 0.67 \times 0.5}{0.15} \\
&= 0.6667
\end{aligned}
$$

That is, the probability that the bus will be late if it rains in the morning during winter is around **66%**.

> **Warning**
>
> This example is solely based on the small sixty-record sample of Table 2, to demonstrate the multiple predictor version of the *Bayes formula*.

Example 0.3. Finding the best-fit line (might be a curve?) - utilizing logarithmic scale.

Given a dataset which shows a raising but strange tendency depicted in Figure 1a. If one tries to fit a straight line between these data points they observe that those do not hug the linear so well. It seems that the differences between values on the ordinate get bigger as the values on the abscissa increase. In such case, switching to a logarithmic scale on the ordinate might be a good decision as presented in Figure 1b.



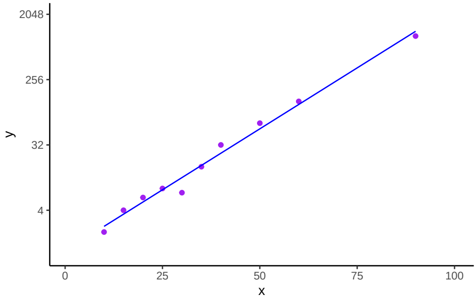
(a) Linear scales, the data shows non-linear uptrend.

(b) Semi-log scales, the data shows an uptrend close to linear.

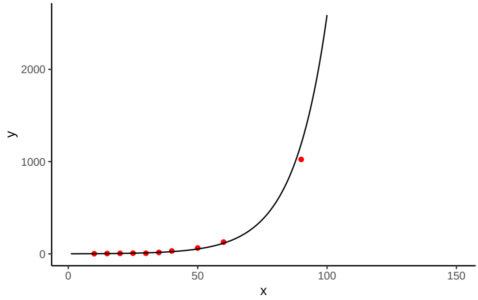
Figure 1: The same dataset plotted on two graphs with linear and semi-log scales, respectively.

The data on the semi-log (`log2`) plot shows a tendency close to linear which means that the data on the original scale follows an exponential trend. One could find the best-fit line on the semi-log plot (Figure 2a) and then transform it back to the original scales to arrive at a non-linear, exponential best-fit curve (Figure 2b) that describes the original data on the original scales.

> ⚠ **Warning**
>
> Applying `geom_smooth` on the original data indicates a different curve from the one acquired above. This might be because data points on the semi-log plot do not perfectly fit the straight line, nor the exponential

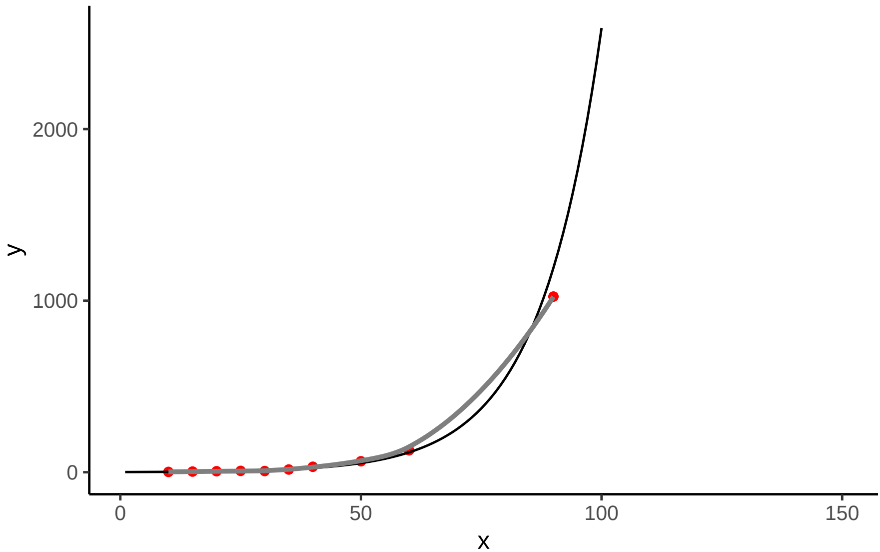


(a)



(b)

curve on the original plot.



Tables

Table 2: A sample of a historical dataset containing the bus schedule outcome along with the morning weather condition for all the first 5 operating days of each month in the past year related to the route and station in Example 0.1 and Example 0.2.

	rowid	date	schedule_outcome	morning_weather	season
1	1	2023-06-01	On schedule	No rain	Summer
2	2	2023-06-02	On schedule	No rain	Summer
3	3	2023-06-05	Late	No rain	Summer
4	4	2023-06-06	On schedule	No rain	Summer
5	5	2023-06-07	On schedule	No rain	Summer
6	6	2023-07-03	On schedule	No rain	Summer
7	7	2023-07-04	On schedule	No rain	Summer
8	8	2023-07-05	On schedule	No rain	Summer
9	9	2023-07-06	On schedule	No rain	Summer
10	10	2023-07-07	On schedule	No rain	Summer
11	11	2023-08-01	On schedule	No rain	Summer
12	12	2023-08-02	On schedule	Rain	Summer
13	13	2023-08-03	On schedule	No rain	Summer
14	14	2023-08-04	On schedule	No rain	Summer
15	15	2023-08-07	Late	No rain	Summer
16	16	2023-09-01	Late	Rain	Fall
17	17	2023-09-04	On schedule	No rain	Fall
18	18	2023-09-05	Late	Rain	Fall
19	19	2023-09-06	On schedule	No rain	Fall
20	20	2023-09-07	On schedule	No rain	Fall
21	21	2023-10-02	On schedule	No rain	Fall
22	22	2023-10-03	On schedule	Rain	Fall
23	23	2023-10-04	On schedule	Rain	Fall
24	24	2023-10-05	On schedule	No rain	Fall
25	25	2023-10-06	On schedule	No rain	Fall

(continued)

	rowid	date	schedule_outcome	morning_weather	season
26	26	2023-11-01	On schedule	Rain	Fall
27	27	2023-11-02	On schedule	Rain	Fall
28	28	2023-11-03	Late	No rain	Fall
29	29	2023-11-06	On schedule	Rain	Fall
30	30	2023-11-07	Late	Rain	Fall
31	31	2023-12-01	Late	Rain	Winter
32	32	2023-12-04	Late	Rain	Winter
33	33	2023-12-05	On schedule	No rain	Winter
34	34	2023-12-06	On schedule	Rain	Winter
35	35	2023-12-07	On schedule	No rain	Winter
36	36	2024-01-01	Late	No rain	Winter
37	37	2024-01-02	Late	Rain	Winter
38	38	2024-01-03	On schedule	No rain	Winter
39	39	2024-01-04	Late	Rain	Winter
40	40	2024-01-05	On schedule	Rain	Winter
41	41	2024-02-01	Late	Rain	Winter
42	42	2024-02-02	Late	Rain	Winter
43	43	2024-02-05	Late	Rain	Winter
44	44	2024-02-06	Late	No rain	Winter
45	45	2024-02-07	On schedule	No rain	Winter
46	46	2024-03-01	Late	Rain	Spring
47	47	2024-03-04	On schedule	No rain	Spring
48	48	2024-03-05	On schedule	Rain	Spring
49	49	2024-03-06	On schedule	No rain	Spring
50	50	2024-03-07	On schedule	No rain	Spring
51	51	2024-04-01	Late	No rain	Spring
52	52	2024-04-02	On schedule	No rain	Spring
53	53	2024-04-03	Late	Rain	Spring
54	54	2024-04-04	On schedule	Rain	Spring
55	55	2024-04-05	On schedule	Rain	Spring

(continued)

	rowid	date	schedule_outcome	morning_weather	season
56	56	2024-05-01	On schedule	No rain	Spring
57	57	2024-05-02	On schedule	No rain	Spring
58	58	2024-05-03	On schedule	No rain	Spring
59	59	2024-05-06	On schedule	No rain	Spring
60	60	2024-05-07	On schedule	No rain	Spring

Table 3: Filtered records of Table 2, where the bus schedule outcome is *Late*.

	rowid	date	schedule_outcome	morning_weather	season
1	3	2023-06-05	Late	No rain	Summer
2	15	2023-08-07	Late	No rain	Summer
3	16	2023-09-01	Late	Rain	Fall
4	18	2023-09-05	Late	Rain	Fall
5	28	2023-11-03	Late	No rain	Fall
6	30	2023-11-07	Late	Rain	Fall
7	31	2023-12-01	Late	Rain	Winter
8	32	2023-12-04	Late	Rain	Winter
9	36	2024-01-01	Late	No rain	Winter
10	37	2024-01-02	Late	Rain	Winter
11	39	2024-01-04	Late	Rain	Winter
12	41	2024-02-01	Late	Rain	Winter
13	42	2024-02-02	Late	Rain	Winter
14	43	2024-02-05	Late	Rain	Winter
15	44	2024-02-06	Late	No rain	Winter
16	46	2024-03-01	Late	Rain	Spring
17	51	2024-04-01	Late	No rain	Spring
18	53	2024-04-03	Late	Rain	Spring

Table 4: Filtered records of Table 2, where the bus schedule outcome is *Late* and the weather condition is *Rain*.

	rowid	date	schedule_outcome	morning_weather	season
1	16	2023-09-01	Late	Rain	Fall
2	18	2023-09-05	Late	Rain	Fall
3	30	2023-11-07	Late	Rain	Fall
4	31	2023-12-01	Late	Rain	Winter
5	32	2023-12-04	Late	Rain	Winter
6	37	2024-01-02	Late	Rain	Winter
7	39	2024-01-04	Late	Rain	Winter
8	41	2024-02-01	Late	Rain	Winter
9	42	2024-02-02	Late	Rain	Winter
10	43	2024-02-05	Late	Rain	Winter
11	46	2024-03-01	Late	Rain	Spring
12	53	2024-04-03	Late	Rain	Spring

Table 5: Filtered records of Table 2, where the weather condition is *Rain*.

	rowid	date	schedule_outcome	morning_weather	season
1	12	2023-08-02	On schedule	Rain	Summer
2	16	2023-09-01	Late	Rain	Fall
3	18	2023-09-05	Late	Rain	Fall
4	22	2023-10-03	On schedule	Rain	Fall
5	23	2023-10-04	On schedule	Rain	Fall
6	26	2023-11-01	On schedule	Rain	Fall
7	27	2023-11-02	On schedule	Rain	Fall
8	29	2023-11-06	On schedule	Rain	Fall
9	30	2023-11-07	Late	Rain	Fall
10	31	2023-12-01	Late	Rain	Winter
11	32	2023-12-04	Late	Rain	Winter
12	34	2023-12-06	On schedule	Rain	Winter

(continued)

	rowid	date	schedule_outcome	morning_weather	season
13	37	2024-01-02	Late	Rain	Winter
14	39	2024-01-04	Late	Rain	Winter
15	40	2024-01-05	On schedule	Rain	Winter
16	41	2024-02-01	Late	Rain	Winter
17	42	2024-02-02	Late	Rain	Winter
18	43	2024-02-05	Late	Rain	Winter
19	46	2024-03-01	Late	Rain	Spring
20	48	2024-03-05	On schedule	Rain	Spring
21	53	2024-04-03	Late	Rain	Spring
22	54	2024-04-04	On schedule	Rain	Spring
23	55	2024-04-05	On schedule	Rain	Spring

Table 6: Filtered records of Table 2, where the weather condition is *Rain* and the season is *Winter*.

	rowid	date	schedule_outcome	morning_weather	season
1	31	2023-12-01	Late	Rain	Winter
2	32	2023-12-04	Late	Rain	Winter
3	34	2023-12-06	On schedule	Rain	Winter
4	37	2024-01-02	Late	Rain	Winter
5	39	2024-01-04	Late	Rain	Winter
6	40	2024-01-05	On schedule	Rain	Winter
7	41	2024-02-01	Late	Rain	Winter
8	42	2024-02-02	Late	Rain	Winter
9	43	2024-02-05	Late	Rain	Winter

References

[1] *Practical machine learning in r*. 2020.

[2] *Wikipedia*. Available: https://en.wikipedia.org/wiki/Logarithmic_scale

[3] *Wikipedia*. Available: https://en.wikipedia.org/wiki/Semi-log_plot