Bayes' Theorem

David J. H. Shih

Intended learning outcomes

- Apply Bayes' theorems in data modelling

Bayes' theorem for probability functions

If $\mathcal{B}_1, \mathcal{B}_2, \ldots$ be a partition of the sample space $\mathcal{S}$, for any $\mathcal{A} \subseteq \mathcal{S}$,

$$P\left(\mathcal{B}_j \mid \mathcal{A}\right) = \frac{P\left(\mathcal{A} \mid \mathcal{B}_j\right) P\left(\mathcal{B}_j\right)}{P\left(\mathcal{A}\right)},$$

where $P\left(\mathcal{A}\right)$ is given by the total law of probability:

$$P\left(\mathcal{A}\right) = \sum_j P\left(\mathcal{A} \mid \mathcal{B}_j\right) P\left(\mathcal{B}_j\right)$$

Proof

It follows directly from the definition of conditional probability.

Example: Medical testing

Let X be the result of a medical testing: positive $(X = +1)$ or negative $(X = -1)$.

Let Y represent whether a person has the disease $(Y = 1)$ or not $(Y = 0)$.

The test has a sensitivity of 0.9, meaning that $P(X = +1 \mid Y = 1) = 0.9$.

Further, the test has a specificity of 0.6, meaning that $P(X = -1 \mid Y = 0) = 0.6$, so the false positive rate is $P(X = +1 \mid Y = 0) = 1 - 0.6 = 0.4$.

Before taking the test, without knowing any extra knowledge, we may expect a person to have same probability of having the disease as the general population, which is equal to the disease prevalence.

If the disease prevalence is 1%, we would set $P(Y = 1) = 0.01$.

If the test result is positive, then the post-test probability of disease is

$$
\begin{aligned}
&P\left(Y=1 \mid X=+1\right) \\
&=\frac{P\left(X=+1 \mid Y=1\right) P\left(Y=1\right)}{P\left(X=+1 \mid Y=1\right) P\left(Y=1\right)+P\left(X=+1 \mid Y=0\right) P\left(Y=0\right)} \\
&=\frac{(0.9)(0.01)}{(0.9)(0.01)+(1-0.6)(1-0.01)} \approx 0.022,
\end{aligned}
$$

which is higher than probability of disease before the test. This value is also called positive predictive value.

Conversely, if the result is negative, then the probability of disease after the test is

$$
\begin{aligned}
&P(Y=1 \mid X=-1) \\
&= \frac{P(X=-1 \mid Y=1)\,P(Y=1)}{P(X=-1 \mid Y=1)\,P(Y=1)+P(X=-1 \mid Y=0)\,P(Y=0)} \\
&= \frac{(1-0.9)(0.01)}{(1-0.9)(0.01)+(0.6)(1-0.01)} \approx 0.0017,
\end{aligned}
$$

which is lower than the probability of disease before the test.

Notations

Probability function $P(\mathcal{A})$ takes a set $\mathcal{A} \subseteq \mathcal{S}$ as a input.

Induced probability function $P_X(X = x)$ takes a realized value x of the random variable X.

Probability mass or density function $f_X(x)$

Cumulative probabilify mass or density function $F_X(x)$

The term "probability distribution" means $f_X(x)$ or $F_X(x)$.

Bayesian shorthand for pmf/pdf is $p(x)$.

Bayes' theorem for probablity distributions

Given observed data x, likelihood $p\left(x \mid \theta\right)$, and prior $p\left(\theta\right)$, the posterior distribution of the parameter θ is given by

$$p\left(\theta \mid x\right) = \frac{p\left(x \mid \theta\right)\, p\left(\theta\right)}{p\left(x\right)},$$

If θ is discrete, then the model evidence $p\left(x\right)$ is given by

$$p\left(x\right) = \sum_{\theta \in \Theta} p\left(x \mid \theta\right) p\left(\theta\right).$$

If θ is continuous, then

$$p\left(x\right) = \int_{\Theta} p\left(x \mid \theta\right) p\left(\theta\right)\, d\theta.$$

Notations

Sampling distribution $p\left(X \mid \theta\right)$

Likelihood $L(\theta) = p\left(x \mid \theta\right)$

Uniform-Bernoulli model (single sample)

$$X \sim \text{Bernoulli}(\theta), \quad \theta \sim \text{Uniform}(0,1)$$

In other words,

$$p\left(x \mid \theta\right) = \theta^x (1-\theta)^{1-x}, \quad p\left(\theta\right) = 1$$

Therefore,

$$p\left(\theta \mid x\right) = \frac{p\left(x \mid \theta\right)\, p\left(\theta\right)}{p\left(x\right)} = \frac{\left(\theta^x (1-\theta)^{1-x}\right)(1)}{p\left(x\right)},$$

where

$$p\left(x\right) = \int_\Theta p\left(x \mid \theta\right) p\left(\theta\right)\, d\theta = \int_\Theta \theta^x (1-\theta)^{1-x} = \frac{1}{2}.$$

Finally,

$$p\left(\theta \mid x\right) = \frac{1}{2}\,\theta^{x}(1-\theta)^{1-x}.$$

Plot $p\left(\theta\right)$, $p\left(\theta \mid X=0\right)$, and $p\left(\theta \mid X=1\right)$.

Beta distribution

For $x \in (0, 1)$, the beta distribution is defined by

$$\text{Beta}\,(x \mid \alpha, \beta) = B(\alpha, \beta)^{-1} x^{\alpha-1}(1-x)^{\beta-1},$$

where $B(\alpha, \beta)$ is the beta function defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dt.$$

Clearly, Beta $(x \mid \alpha, \beta)$ is a distribution:

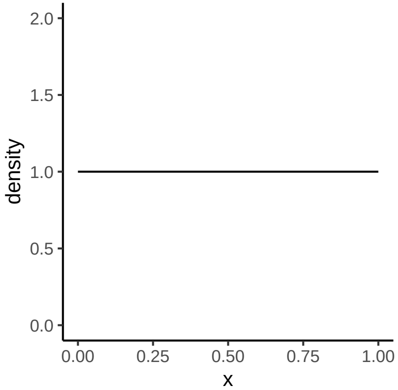
$$x \geq 0 \text{ and } x \leq 0 \quad \Rightarrow \quad \text{Beta}\,(x \mid \alpha, \beta) \geq 0$$

$$\begin{aligned}\int_0^1 \text{Beta}\,(x \mid \alpha, \beta) &= B(\alpha, \beta)^{-1} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \\ &= B(\alpha, \beta)^{-1} B(\alpha, \beta) = 1.\end{aligned}$$

Plotting

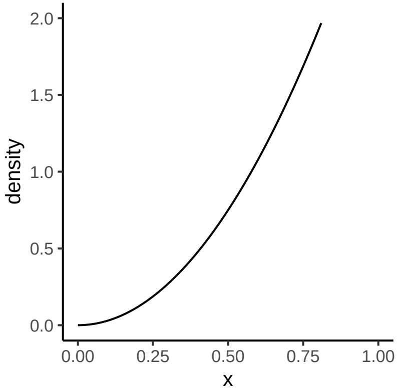
```r
library(ggplot2)

plot_dbeta <- function(alpha, beta) {
  x <- seq(0, 1, 0.01);
  p <- dbeta(x, alpha, beta);
  qplot(x, p, geom="line") + theme_classic() +
    ylab("density") + ylim(0, 2)
}
```

```r
plot_dbeta(1, 1)
```



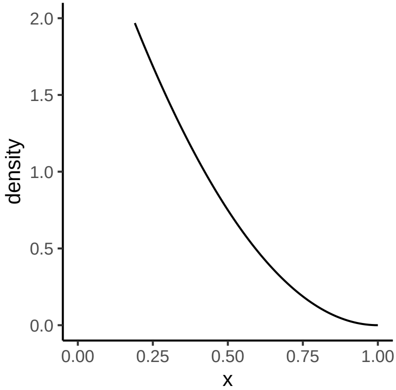
$$\text{Beta}\left(x \mid \alpha, \beta\right) = B(\alpha, \beta)^{-1} x^{\alpha-1} (1-x)^{\beta-1},$$

```r
plot_dbeta(3, 1)
```



$$\text{Beta}\left(x \mid \alpha, \beta\right) = B(\alpha, \beta)^{-1} x^{\alpha-1} (1-x)^{\beta-1},$$

```r
plot_dbeta(1, 3)
```



$$\text{Beta}\,(x \mid \alpha, \beta) = B(\alpha, \beta)^{-1} x^{\alpha-1} (1-x)^{\beta-1},$$

Beta-Bernoulli model

Given $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]$,

$$X_i \sim \text{Bernoulli}(\theta), \quad \theta \sim \text{Beta}(a, b)$$

Define $y = \sum_i x_i$.

$$\begin{aligned}
p(\boldsymbol{x} \mid \theta) &= \prod_i p(x_i \mid \theta) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y (1-\theta)^{N-y} \\
p(\theta) &= B(a,b)^{-1}\theta^{a-1}(1-\theta)^{b-1} \\
p(\boldsymbol{x}, \theta) &= B(a,b)^{-1}\theta^{a+y-1}(1-\theta)^{b+N-y-1} \\
p(\boldsymbol{x}) &= \int_\Theta p(\boldsymbol{x}, \theta)\, d\theta = B(a,b)^{-1}\int_\Theta \theta^{a+y-1}(1-\theta)^{b+N-y-1} \\
&= B(a,b)^{-1}B(a+y, b+N-y).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
p\left(\theta \mid \boldsymbol{x}\right) &= \frac{p\left(x \mid \theta\right)\, p\left(\theta\right)}{p\left(x\right)} \\
&= \frac{B(a,b)^{-1}\theta^{a+y-1}(1-\theta)^{b+N-y-1}}{B(a,b)^{-1}B(a+y,b+N-y)} \\
&= B(a+y,b+N-y)^{-1}\theta^{a+y-1}(1-\theta)^{b+N-y-1} \\
&= \mathsf{Beta}\left(\theta \mid a+y, b+N-y\right).
\end{aligned}
$$

Before and after observing data $\boldsymbol{x}$, our beliefs about θ are

$$
\begin{aligned}
\theta \;&\sim\; \mathsf{Beta}(a,b) \\
\theta \mid \boldsymbol{x} \;&\sim\; \mathsf{Beta}\left(a + \sum_i x_i, b + N - \sum_i x_i\right).
\end{aligned}
$$

Summary

"A Bayesian is skeptical about everything except that he's a Bayesian." - Geoffrey Hinton (original source unknown)

Blais 2014, chapters 4-5.

Intended learning outcomes

- Apply Bayes' theorems in data modelling