Bayesian Estimation

When estimating a parameter θ (or parameters $\boldsymbol{\theta}$) using the maximum likelihood framework, we make almost no assumptions — the only knowledge we used was membership in a set $\mathcal{T}$.

It is often the case, however, that some values of $\boldsymbol{\theta}$ are *a priori* more likely than others.

Here is an example. Suppose that we have an optical character recognition system that is working from a fuzzy image, and it trying to decide whether it is looking at an 'O' or a 'Q'. We know that the letter O appears in written English about 80 times more frequently than the letter Q[1], so it would make sense to bias the decision towards 'O'.

Of course, this model might depend on the context. The bigram OU is only about 12 times as likely as QU.

Bayes rule gives us a mathematical framework for incorporating information like this. Let's have a quick review, starting with the simplest possible situation. Let A and B be probabilistic events (either they happen or they don't). We know that [2]

$$\mathrm{P}(A, B) = \mathrm{P}(A)\,\mathrm{P}(B|A)$$

or equivalently

$$\mathrm{P}(A, B) = \mathrm{P}(B)\,\mathrm{P}(A|B)\,.$$

[1] About 7.5% of written letters are 'O', while about .095% are 'Q'

[2] Again, the notation $\mathrm{P}(A, B)$ means "the probability that both A and B occur".

Bayes rule combines these statements into

$$\mathrm{P}\left(A|B\right) = \frac{\mathrm{P}\left(B|A\right)\mathrm{P}\left(A\right)}{\mathrm{P}\left(B\right)}.$$

When we use Bayes rule for inference, that is trying to ascertain what might be happening with event A after observing that event B occurred, we trade-off how A effects B (as quantified by $\mathrm{P}\left(B|A\right)$) versus the base rate of A (as quantified by $\mathrm{P}\left(A\right)$). Integrating information about the base rate can be very powerful, as this straightforward example illustrates.

Example: The incidence rate for disease X is 15 in $100,000$. There is a test for disease X that is 95% accurate: if you have the disease, there is a 95% chance the test comes back positive, if you do not have the disease, there is a 95% chance the test comes back negative.

You test positive for disease X. What are the chances that you actually have it?

Answer:

More generally, in Bayesian estimation, the unknown parameters $\boldsymbol{\theta}$ are themselves treated as a *random variable/vector*, and prior information (i.e. a model) for these parameters in encoded in a distribution $f_\Theta(\boldsymbol{\theta})$:

$$\Theta \sim f_\Theta(\boldsymbol{\theta}).$$

We make an observation X, which is another random variable/vector that is related to $\boldsymbol{\theta}$ through the conditional distribution

$$X \sim f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}).$$

Given a particular observation $X = \boldsymbol{x}$, Bayes rule tells us how to *update* our model for Θ in light of this information:

$$f_\Theta(\boldsymbol{\theta}|X = \boldsymbol{x}) = \frac{f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}) f_\Theta(\boldsymbol{\theta})}{f_X(\boldsymbol{x})}$$

The model $f_\Theta(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ before the observation is called the *prior*; the updated model $f_\Theta(\boldsymbol{\theta}|\boldsymbol{x})$ after the observation is called the *posterior*.

This expression holds for X, Θ being either discrete random variables (and so the associated f_Θ, f_X are probability mass functions) or continuous-valued random variables (where the f_Θ, f_X are density functions). Given the modeling information above, the marginal $f_X(\boldsymbol{x})$ can be computed as

$$f_X(\boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \mathcal{T}} f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}) \ \mathrm{P}\left(\Theta = \boldsymbol{\theta}\right), \quad \Theta \text{ discrete},$$

or

$$f_X(\boldsymbol{x}) = \int_{\boldsymbol{\theta} \in \mathcal{T}} f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}) \, f_\Theta(\boldsymbol{\theta}) \ d\boldsymbol{\theta}, \quad \Theta \text{ continuous},$$

where $\mathcal{T}$ is the range (i.e. the set of all possible outcomes) of Θ.

Example: Let's return to our free throw shooting example. As before, our model is that every player has an intrinsic parameter θ, which is the probability they make a free throw on a given attempt. We capture the outcome of the ith free throw as a binary-valued random variable X_i ($X_i = 1$ for make, $X_i = 0$ for miss),

$$\mathrm{P}\left(X_i = 1\right) = \theta, \quad \mathrm{P}\left(X_i = 0\right) = 1 - \theta.$$

We encode this model in the conditional probability mass function

$$\begin{aligned} f_X(x|\Theta = \theta) &= \{1-\theta, \theta\} \quad \text{for } x = \{0, 1\}, \\ &= \theta^x (1-\theta)^{1-x}. \end{aligned}$$

When a player enters the NBA, we have no idea what their free throw percentage θ is, so we will use

$$f_\Theta(\theta) = \mathrm{Uniform}([0,1]) = 1, \text{ for } \theta \in [0,1].$$

Suppose that they make their first free throw, $X_1 = 1$. How does our model for their θ update?

$$f_\Theta(\theta|X_1 = 1) =$$

Now suppose they make their first free throw, miss their second, then make the third and fourth:

$$X_1 = 1, \quad X_2 = 0, \quad X_3 = 1, \quad X_4 = 1.$$

Assume that the outcomes of the free throws are conditionally independent given θ:

$$f_X(x_1, x_2, x_3, x_4|\theta) = f_X(x_1|\theta) f_X(x_2|\theta) f_X(x_3|\theta) f_X(x_4|\theta).$$

Now what does our model for Θ look like?

To answer this, need the following integral expression:

$$\int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We have an explicit expression for the gamma function at the integers, $\Gamma(\alpha) = (\alpha - 1)!$ for $\alpha \in \mathbb{Z}$.

This is actually a common model for many problems in machine learning: the observations are binary (meaning that their sum is *binomial*), and the prior on the parameter θ is *Beta distributed*:

$$\Theta \sim \text{Beta}(\alpha, \beta), \quad f_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$

Think of the "beta function" $B(\alpha, \beta)$ as the normalization constant we need to get the pdf to integrate to 1; it is a fact that

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

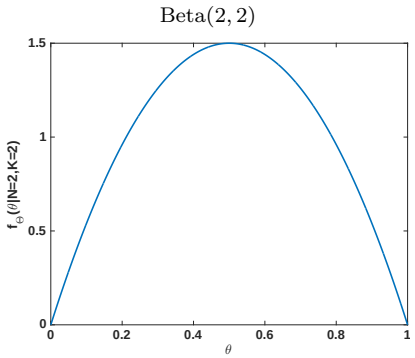
If we observe N trials $X_1 = x_1, \ldots, X_N = x_N$, and K of them are successful

$$K = \sum_{n=1}^{N} x_n,$$

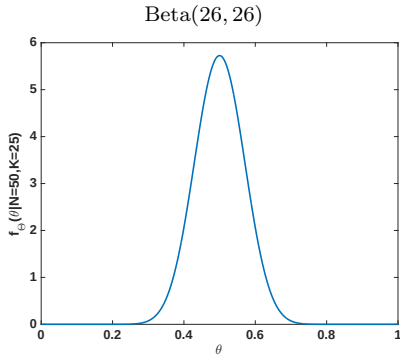
then the posterior for Θ is

$$f_\Theta(\theta | x_1, \ldots, x_N) = \text{Beta}(\alpha + K, \beta + N - K).$$

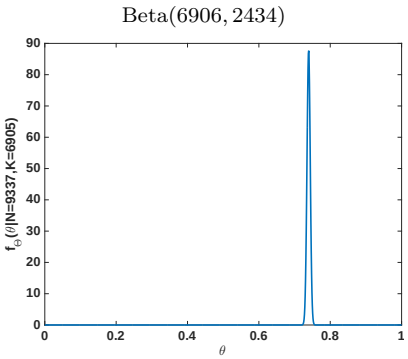
Starting out with the prior $\Theta \sim \text{Uniform}([0,1]) = \text{Beta}(1,1)$, here is the posterior after two observations, $X_1 = 1$, $X_2 = 0$:



Here is what it looks like after 50 observations, with $K = 25$ successes:



Here is what it looks like if we put in LeBron James' free throw totals for the first 15 years of his career[3] (9337 attempts, 6905 makes):



If we return to our shooting-free-throws-in-the-NBA example, we might ask if taking $f_\Theta(\theta)$ as uniform is really the best prior — I mean, do we really want to assign equal weight to $\theta = 0.05$ (which is unheard of) as to $\theta = 0.8$ (which would be only slightly above average). Indeed, the average free throw rate in the NBA is about 0.75, with a standard deviation of about $\sigma = 0.1$ (so the variance

[3] October 29, 2003 through October 30, 2018, https://www.basketball-reference.com/players/j/jamesle01.html.

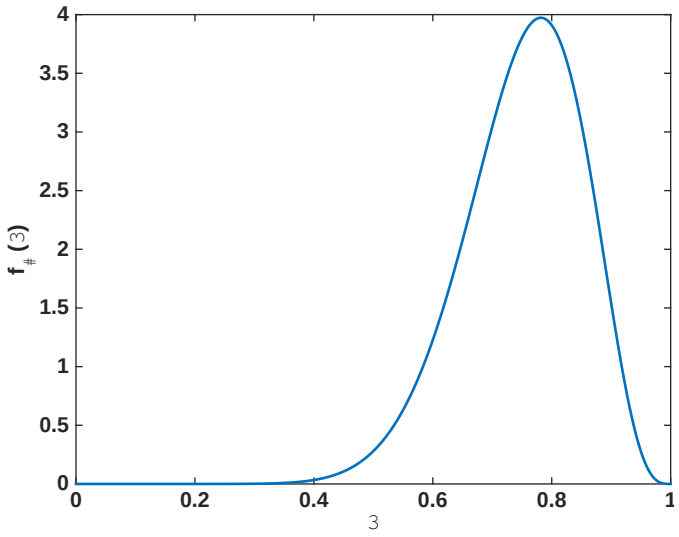
is $\sigma^2 = 0.01$). Let's find a Beta distribution that matches these moments. If $\Theta \sim \text{Beta}(\alpha, \beta)$, then

$$\mathrm{E}[\Theta] = \frac{\alpha}{\alpha + \beta},$$
$$\mathrm{var}(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

We can then solve a system of equations to find α, β such that $\mathrm{E}[\Theta] = 0.75$ and $\mathrm{var}(\Theta) = 0.01$ — the result of this is

$$\alpha \approx 13.313, \quad \beta = 4.104.$$

Here is a picture of this prior:



Bayesian Parameter Estimation

So far in this set of notes, we have only talked about how to update our model for Θ once one or more observations of X have been made. The result of is a (new) probability density function for Θ.

How do we turn the posterior density[4] $f_\Theta(\boldsymbol{\theta}|X = \boldsymbol{x})$ into an estimate of θ? Here are two popular approaches.

Conditional Mean. Set

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \text{E}[\Theta|X = x] = \int \boldsymbol{\theta} f_\Theta(\boldsymbol{\theta}|X = x)\, d\boldsymbol{\theta}$$

As we have already seen, this estimate provides the *minimum mean-squared error* $\text{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2]$.

Maximum a Posteriori (MAP). Set

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} f_\Theta(\boldsymbol{\theta}|X = \boldsymbol{x}).$$

That is, choose the value of θ that is after-the-fact (a posteriori) most likely. Note that since

$$f_\Theta(\boldsymbol{\theta}|X = \boldsymbol{x}) = \frac{f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}) f_\Theta(\boldsymbol{\theta})}{f_X(\boldsymbol{x})},$$

and $f_X(\boldsymbol{x})$ doesn't depend on $\boldsymbol{\theta}$, we can write this as

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} f_X(\boldsymbol{x}|\Theta = \boldsymbol{\theta}) f_\Theta(\boldsymbol{\theta}). \tag{1}$$

[4]For compactness, we are abbreviating all observations as $X = \boldsymbol{x}$; if there is more than one, this might be more properly written as $f_\Theta(\boldsymbol{\theta}|X_1 = \boldsymbol{x}_1, \ldots, X_N = \boldsymbol{x}_N)$.

With (1), the comparison of the MAP estimator to the MLE is clear — the MLE simply maximizes the likelihood

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} f_X(\boldsymbol{x};\boldsymbol{\theta}),$$

while the MAP estimator weights this likelihood by the prior f_Θ. (The difference in notation, $f_X(\boldsymbol{x};\boldsymbol{\theta})$ versus $f_X(\boldsymbol{x}|\Theta=\boldsymbol{\theta})$ just comes about because we are treating $\boldsymbol{\theta}$ as deterministic in one case, but random in the other. In both cases, the quantity is a likelihood function that depends on $\boldsymbol{\theta}$.)