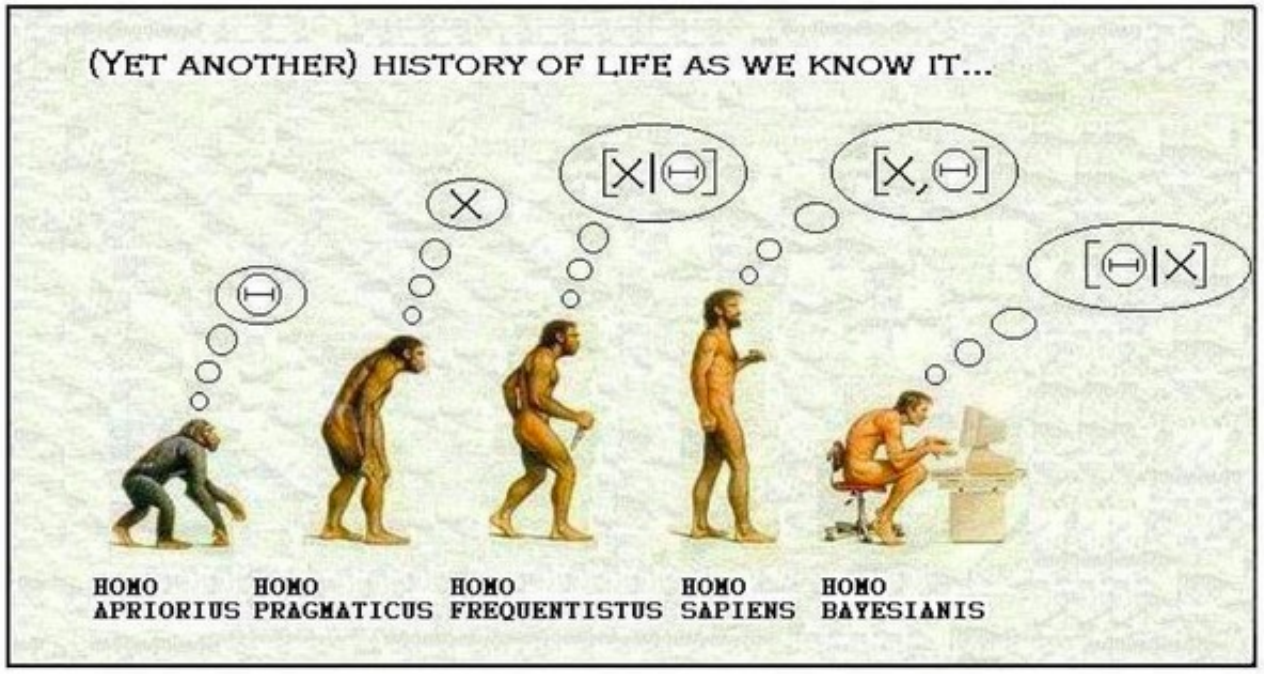
Introduction to Data Science

BAYESIAN THINKING

BRIAN D'ALESSANDRO

MOTIVATION

Frequentist vs Bayesian debates aside, it is good to have more tools at your disposal and to match the right tool to the right problem.



USING THE RIGHT TOOL FOR THE JOB

When Frequentist approaches are likely fine

- Big Data, esp., large sample sizes
- You only care about point estimates (i.e., E[Y|X], which is most prediction systems)

When Bayesian approaches may be better

- Small data samples
- Parameter/prediction uncertainty is part of the decision making process

REVIEW – BAYES' RULE

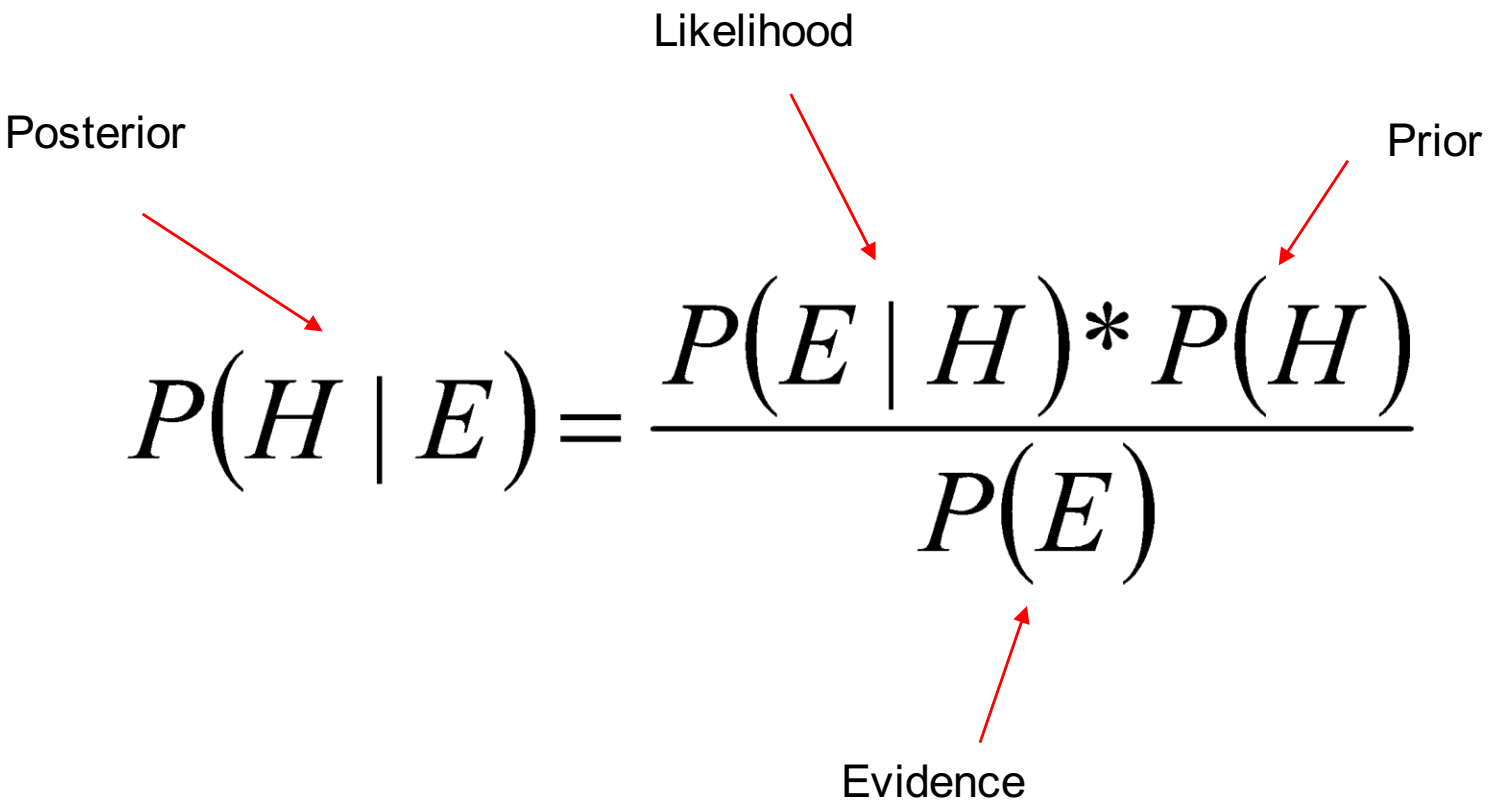
Particularly useful for reasoning about hypotheses (H) given evidence (E)

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

To see this:

$$P(H \mid E) * P(E) = P(EH) = P(E \mid H) * P(H)$$

REVIEW - BAYES TERMS



USEFUL/COMMON APPLICATIONS

- Naïve Bayes
- General probabilistic reasoning
- Regularization
- Smoothing probability estimates
- AB Tests
- Multi-Armed Bandits

EXAMPLE OF BAYES RULE FOR INFERENCE

- Someone tested for a disease D, which has a generally low base rate: P(D) = 0.04%
- A test comes back positive (PT), but the test isn't perfect:
  - Sensitivity/TPR (P(PT|D)) = 94%
  - Specificity/TNR (P(ND|!PT)) = 98%
  - FPR (P(PT|ND)) = 2%

How worried should we be?

I.e., what is the likelihood that our friend actually has the disease?

USING BAYES RULE

$$P(D \mid PT) = \frac{P(PT \mid D)P(D)}{P(PT)}$$

$$= \frac{P(PT \mid D)P(D)}{P(PT \mid D)P(D) + P(PT \mid ND)P(ND)}$$

$$= \frac{(0.94)(0.0004)}{(0.94)(0.0004) + (0.02)(0.9996)}$$

$$= 0.018$$

In Bayesian terms, we want to compute the posterior distribution of having the disease given the new evidence of the positive test result.

What is the impact of base rates P(D) on the final determination?

What is the impact of FPR and TPR of the test on the final determination?

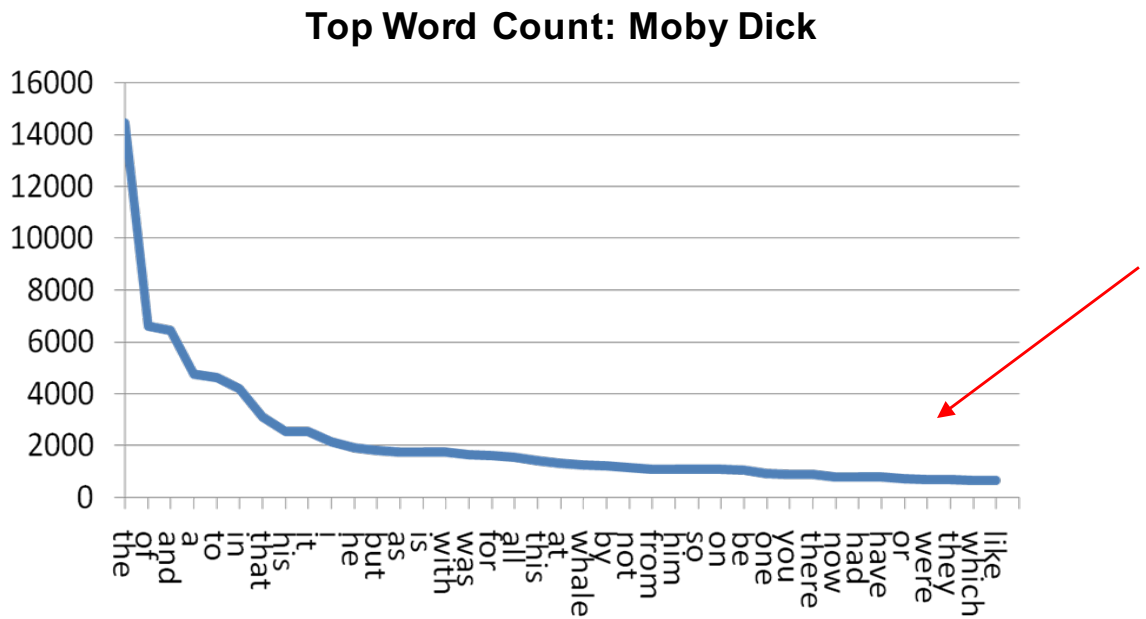
So while P(D|PT) >> P(D), a positive test result still produces a low probability of the disease.

COMPUTING PROBABILITIES

MOTIVATING EXAMPLE

Let's assume we want P(Word|Y=y) (for Naïve Bayes). The most popular words in a book/text appear often enough that we can trust our estimate of P(Word|Y=y).

But popular words are often the least discriminative, i.e., P(Word|Y=y)=P(Word)

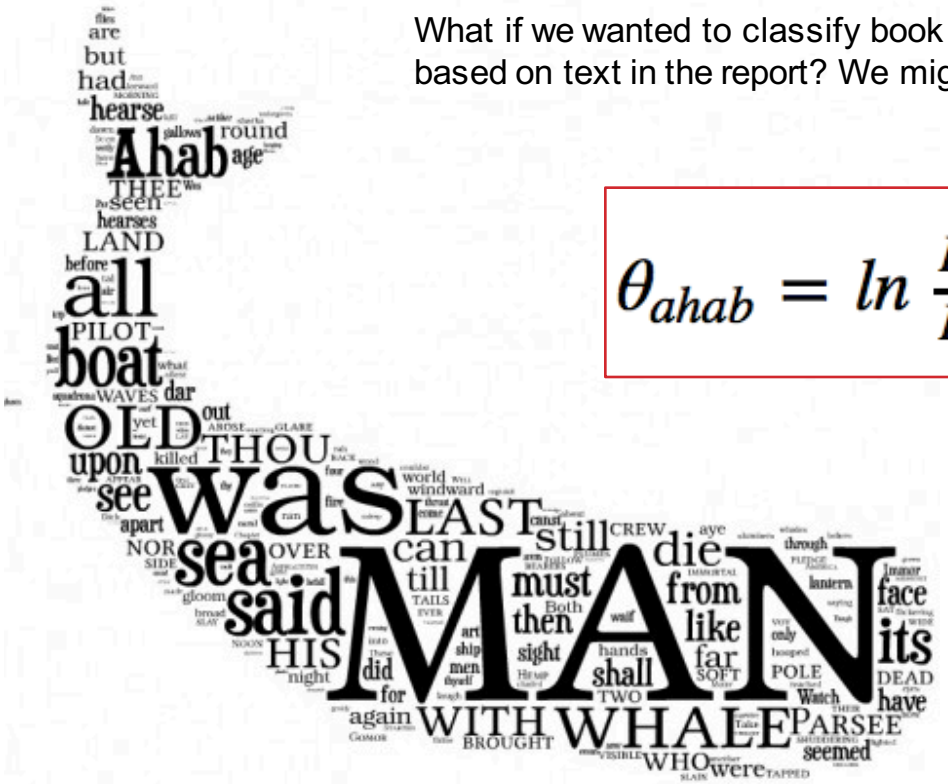


What is the best way to deal with the more sparse words?

Source: http://searchengineland.com/the-long-tail-of-search-12198

SPARSITY IS INTERESTING

What we usually want are the rare words, like “whale”, “ahab” and “harpooner”



What if we wanted to classify book reports as relating to Moby Dick based on text in the report? We might estimate parameters like:

$$\theta_{ahab} = ln \frac{P(ahab|Y=1)}{P(ahab|Y=0)}$$

We want to estimate class conditional rates for the rarer words, but how can we do this in a stable (low variance) way?

Source: http://ahistoryofnewyork.com/2013/01/moby-dick-big-read-day-117/

PROBABILITY ESTIMATION

Estimating a probability just amounts to counting.

$$p(ahab|Y = y) = \frac{cnt(ahab) \mid y}{cnt(allwords) \mid y}$$

But rare words (events) pose many problems with probability estimation:

- **Zeros**: If the number of samples is low, sometimes the word/event won't ever be observed, so the probability is zero.
- **High-variance:** the variance of a probability estimate is p(1-p)/n. If n or p is low, our estimate won't be very trustworthy. Think about if a word was observed once in a sample. If by chance another example came in with the word, the probability estimate of the word will double!

WORDS AS A MULTINOMIAL

We can think of a word in a corpus as a multinomial discrete random variable X, that can take one of any words in the vocabulary. The likelihood of a particular x_i is simply the number of times x_i appears over the total possible occurrences of any X.

$$P(x_i) = \frac{Ni}{N}$$

This has the constraint that:

$$\sum_i P(x_i) = 1$$

We can see that this is how Multinomial Naïve Bayes is set up. Now, in many cases Ni might be low due to the rarity of the word. We thus want to set up the appropriate prior to smooth our estimate of $P(x_i)$.

SMOOTHING A PROBABILITY

In the MNB, we added a little extra to the standard maximum likelihood estimate of a multinomial probability.

$$P(x_i|y) = \frac{N_{yi}+\alpha}{N_y+\alpha n}$$

What is this and where did it come from?
What is the motivation?

CONJUGATE PRIORS

The right prior distribution can lead to analytically tractable posteriors when the prior and the likelihood have the same algebraic form.

Example 1 (A Beta Prior and Binomial Likelihood)
Let $\theta \in (0,1)$ represent some unknown probability. We assume a Beta(α, β) prior so that

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < \theta < 1.$$

We also assume that $X \mid \theta \sim \text{Bin}(n,\theta)$ so that $p(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0,\ldots,n.$ The posterior then satisfies

$$\begin{aligned} p(\theta \mid x) &\propto \pi(\theta)p(x \mid \theta) \\ &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1} \end{aligned}$$

http://www.columbia.edu/~mh2078/MachineLearningORFE/MCMC_Bayes.pdf

CONJUGATE PRIORS

Again, these are chosen more for mathematical convenience than because of a true belief about the prior distribution.

Example 2 (Conjugate Prior for Mean of a Normal Distribution)
Suppose $\theta \sim \mathsf{N}(\mu_0, \gamma_0^2)$ and $p(X_i \mid \theta) = \mathsf{N}(\theta, \sigma^2)$ for $i = 1, \ldots, N$ with σ^2 is assumed known. In this case we have $\boldsymbol{\alpha}_0 = (\mu_0, \gamma_0^2)$. If $\mathbf{X} = (X_1, \ldots, X_N)$ we then have

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\theta; \boldsymbol{\alpha}_0) \\ &\propto e^{-\frac{(\theta-\mu_0)^2}{2\gamma_0^2}} \prod_{i=1}^{N} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\ &\propto \exp\left(-\frac{(\theta-\mu_1)^2}{2\gamma_1^2}\right) \end{aligned}$$

where

$$\gamma_1^{-2} := \gamma_0^{-2} + N\sigma^{-2} \quad \text{and} \quad \mu_1 := \gamma_1^2\left(\mu_0\gamma_0^{-2} + \sum_{i=1}^{n} x_i\sigma^{-2}\right).$$

Of course we recognize $p(\theta \mid \mathbf{x})$ as the $\mathsf{N}(\mu_1, \gamma_1^2)$ distribution.

http://www.columbia.edu/~mh2078/MachineLearningORFE/MCMC_Bayes.pdf

DIRICHLET PRIOR

The conjugate prior of the Multinomial distribution is the Dirichlet distribution, given by:

$$f(x;\alpha) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_i)} \prod_i x^{\alpha_i - 1}$$

Given this prior, the MAP estimate of $P(x_i)$ is then:

$$P = \frac{N_i + \alpha_i}{N + \alpha}$$

If we choose $\alpha_i = 1$, then $\alpha = |\text{vocabulary}|$. This assumes a priori that $P(X_i)$ is uniform (i.e., all words have an equal prior probability).

Reference: http://research.microsoft.com/pubs/69588/tr-95-06.pdf

BETA DISTRIBUTION

The beta distribution is the conjugate prior of the Binomial distribution. It is a special form of the Dirichlet distribution, where X has only two discrete values. The Dirichlet prior has a specific application to Naïve Bayes because X is often defined as a multinomial. In many cases we only want to deal with a binary random variable, which makes the beta distribution appropriate.

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\; x^{\alpha-1}\,(1-x)^{\beta-1}$$

Uses the Gamma function to normalize.

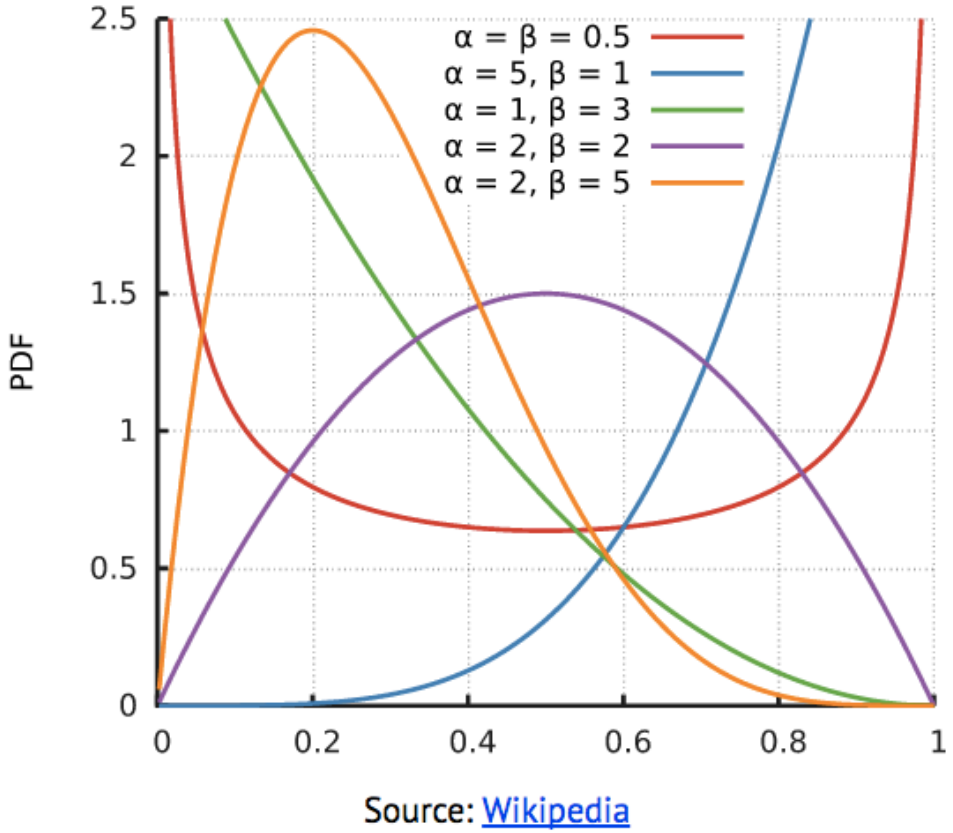
Has same algebraic form as the Binomial distribution.

The mean of the beta distribution is given by:

$$E[X] = \frac{\alpha}{\alpha+\beta}$$

THE BETA DISTRIBUTION

The likelihood of a given P depends on the parameters α and β



Source: Wikipedia

How might we interpret this?

Just like a feature X, which has a distribution, an estimated model parameter can also have a distribution (which might reflect our uncertainty about our estimation of it). Often times we want to estimate a probability, which is bounded between 0 and 1. The Beta Distribution gives us a way to express uncertainty about a probability.

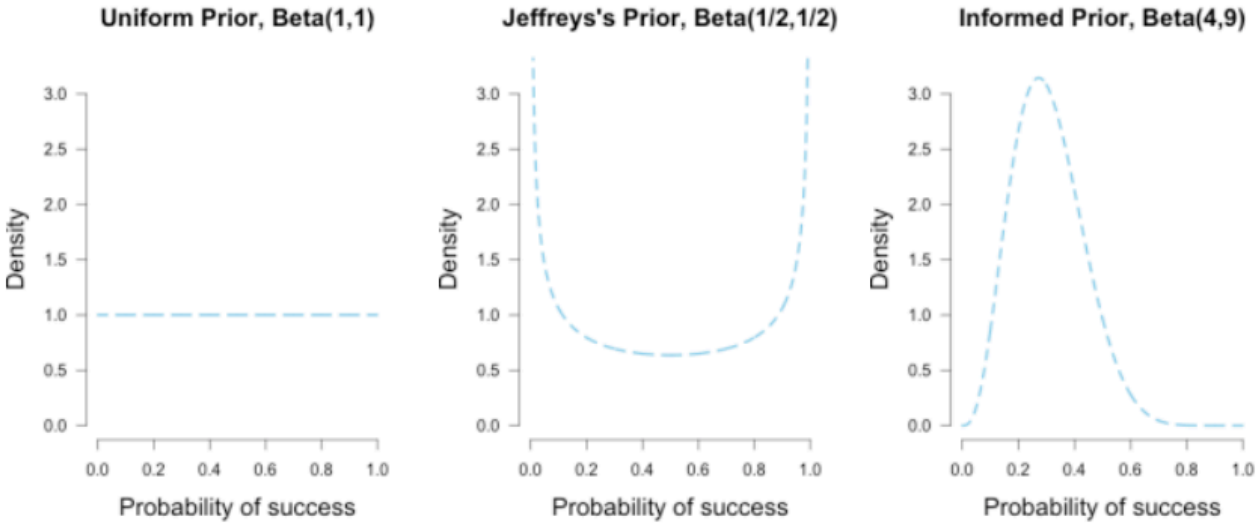
THE BETA PRIOR

The beta prior gives us greater flexibility to incorporate any prior belief in the probability of binary X occurring. With the beta prior, we assume the prior belief that P(X) is given just by the beta distribution with parameters α and β. The smoothed estimate of our probability is then:

$$P = \frac{s + \alpha}{n + \alpha + \beta}$$

The beta prior gives us the ability to estimate our prior directly from the data. I.e., if we want μ = P(X), we can estimate P(X) from the data. We control the degree of smoothing by choosing α and solving for β. The higher α, the more we assert our prior belief into the estimate of θ.

COMMON PRIORS



Best when you have no prior information (aka LaPlace Smoothing

Strong belief in either 0/1

Have some data already, use this when updating.

SMOOTHING: BIAS AND VARIANCE

Smoothing is similar to model parameter regularization, where we induce a bias on our estimate in order to reduce its variance.

$$P = \frac{s + \alpha}{n + \alpha + \beta}$$

Bias – our estimate of P is biased towards the prior α / (α + β) is the. If N=0 (i.e. no data), our estimate will be centered around the prior. The influence of the prior is determined by the magnitude of (α + β)

Variance - the variance of P is largely a function of N. When N is large, we can trust S/N more. If N is large, such that N>> (α + β), our estimate of P becomes less biased because the prior holds less weight.

In other words, when we receive more evidence, we don't have to rely so much on our prior beliefs.

BAYESIAN UPDATING

Goal: Compute the mean of a Bernoulli random variable at different intervals in time

Take a Bayesian approach and use the beta-binomial distribution!

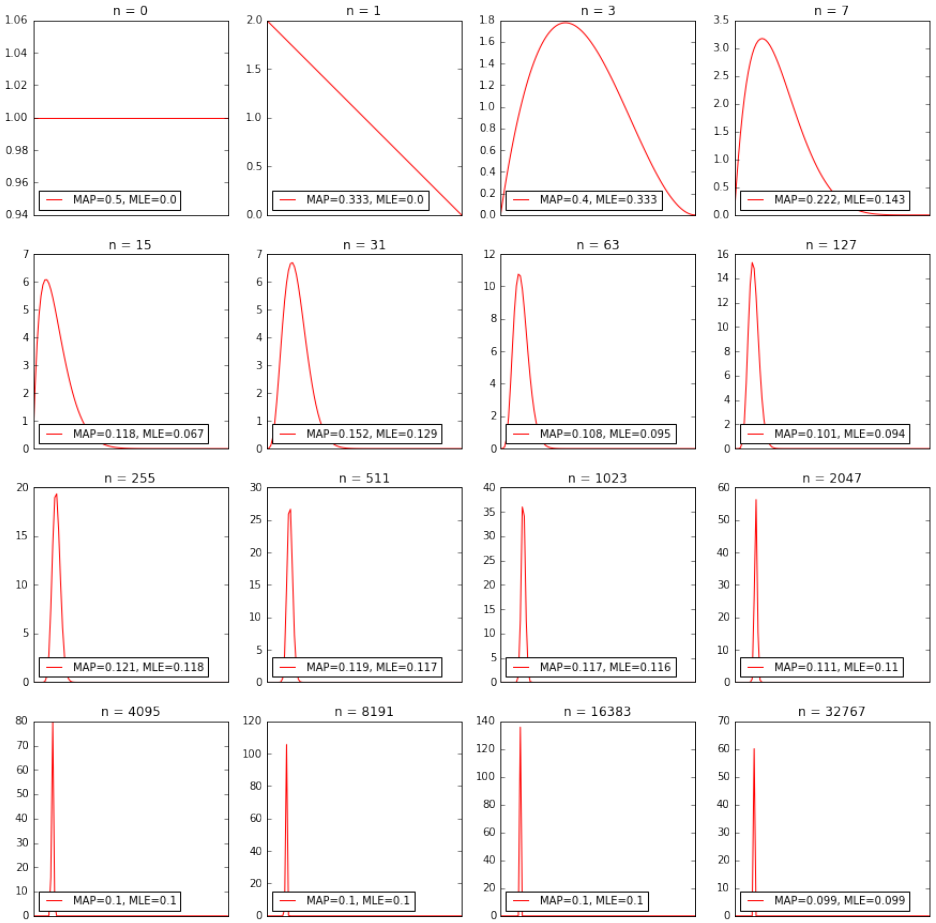
1. Chose a suitable prior before taking any samples: $P(\Theta) = \text{Beta}(1, 1)$
2. Draw N_1 samples and count S_1 successes
3. Compute $\Theta = \text{Beta}(\alpha + S_1, \beta + N_1 - S_1).\text{mean}() = (S_1 + \alpha) \ / (N_1 + \alpha + \beta)$

Need / Receiving more information?

4. Reset: $P(\Theta) = \text{Beta}(\alpha + S_1, \beta + N_1 - S_1)$
5. Draw N_2 more samples, count S_2 more successes
6. Compute $\Theta = \text{Beta}(\alpha + S_1 + S_2, \beta + N_1 + N_2 - (S_1 + S_2)).\text{mean}() = (S_2 + S_1 + \alpha) \ / (N_2 + N_1 + \alpha + \beta)$

7. Repeat steps 4 – 7 until satisfied

BAYESIAN UPDATING



In this example we have true P ~ 10%. Absent any data we choose Beta(1,1) as our prior, which gives all values in [0,1] equal likelihood.

With just a few examples we see peak emerge, and after a few more the peak is close to the true mean.

As exponentially more data comes in, our peak doesn't shift much, but the width gets smaller, indicating our certainty on where the true parameter value lies.