Bayesian Inference

November 10, 2020

Bayesian approaches

- Typically contrasted with **frequentist** approaches
- Treat parameters as uncertain, data as fixed

Bayes' Rule

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

Terms

$$\underset{\text{posterior}}{p(\theta|x)} = \frac{\overset{\text{likelihood}}{p(x|\theta)} \quad \overset{\text{prior}}{p(\theta)}}{\underset{\text{evidence}}{p(x)}}$$

Posterior

The posterior distribution is proportional to the prior times the likelihood: $p(\theta|x) \propto p(x|\theta)p(\theta)$

The posterior distribution *is a distribution* over θ.

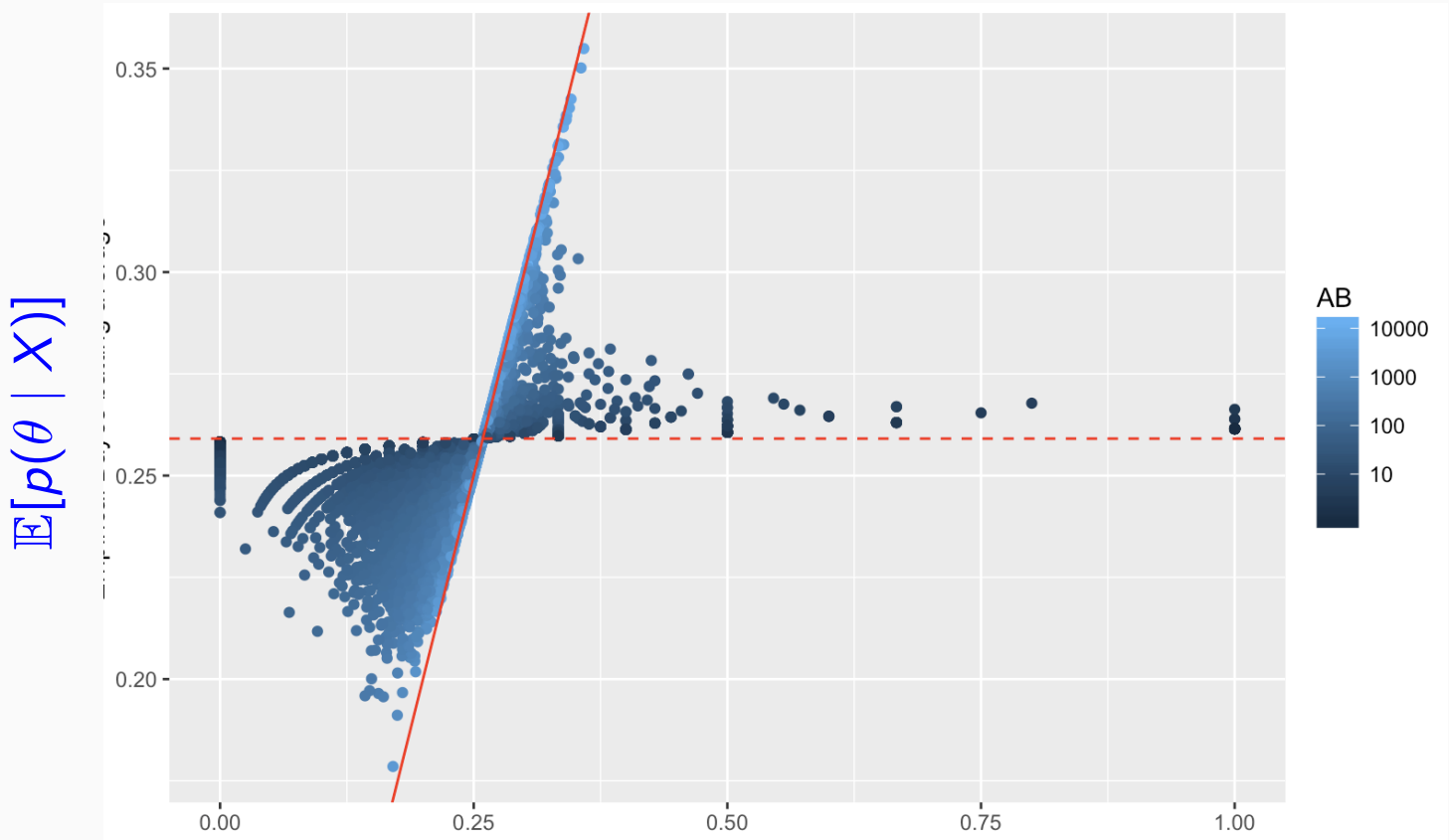
Evidence

The evidence, or *marginal likelihood*, can be used for model comparison.

Simple Motivation: Batting Averages

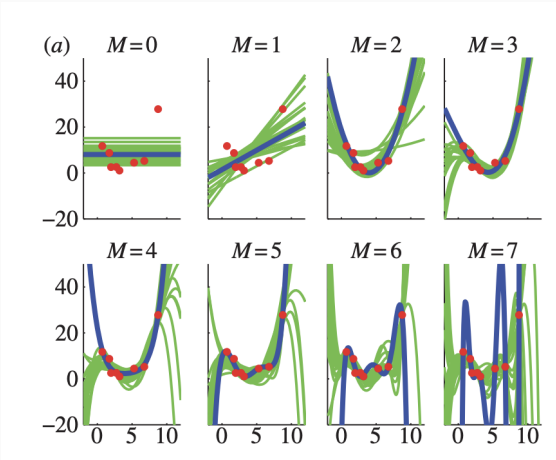
Let

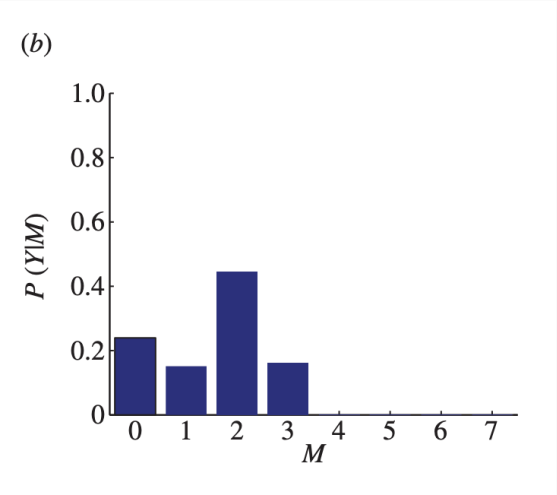
- x be observed data (batting average after n at bats)
- θ be parameters (a player's 'true' batting average)



Bayesian Occam's Razor

Maximum Likelihood (ML) solutions tend to overfit. Bayesian marginalization reduces overfitting.





Models $y = f(x) + \epsilon$ of various complexity (polynomials of various order, M) were fit to 8 data points.

- Plotted are ML polynomials (least squares fits to the data under Gaussian noise) and posterior samples from a Bayesian model (which used a Gaussian prior for the coefficients, and an inverse gamma prior on the noise).
- The ML estimate can look very different from a typical sample from the posterior!

The evidence is plotted as a function of model order. Model orders M=0 to M=3 have considerably higher evidence than other model orders. We see that Bayesian marginalization has reduced overfitting. (The maximum likelihood model, the $M = 7$ model, fits the data perfectly, but overfits wildly, predicting the function will shoot up or down between neighboring data points.)

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110553.

Posterior predictive distribution

Given

$p(\theta|x)$ - posterior
$p(\theta)$ - prior
$p(x|\theta)$ - likelihood

Posterior predictive distribution

Consider the probability of new data x'. Posterior predictive distribution is:

$$p(x'|x) = \int p(x', \theta|x)\, d\theta = \int p(x'|\theta, x)p(\theta|x)\, d\theta = \int p(x'|\theta)p(\theta|x)\, d\theta$$

Incorporates the knowledge and uncertainty about θ that we still had after seeing data x.

Bayesian inference: conjugate example

Sometimes, we can compute the posterior distribution by hand, given prior and likelihood.

Setup: flipping a coin

Probability that it lands heads is (unknown) θ.

Prior probability over θ assumed to follow a $Beta(3, 3)$ distribution:

$$p(\theta) = \frac{\theta^{3-1}(1-\theta)^{3-1}}{B(3,3)}$$

Note: $\theta \sim Beta(a, b)$ means $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$

Will collect data by flipping coin once. Likelihood of observing heads ($x = 1$) or tails ($x = 0$) is given by a Bernoulli distribution:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

.

Bayesian inference: conjugate example

Setup: flipping a coin

Probability that it lands heads is (unknown) θ.

Prior probability over θ assumed to follow a $Beta(3,3)$ distribution:

$$p(\theta) = \frac{\theta^{3-1}(1-\theta)^{3-1}}{B(3,3)}$$

Note: $\theta \sim Beta(a,b)$ means $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$

Will collect data by flipping coin once. Likelihood of observing heads ($x=1$) or tails ($x=0$) is given by a Bernoulli distribution:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

.

Computing the posterior after observing x=1

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \theta^1(1-\theta)^0\theta^2(1-\theta)^2 = \theta^3(1-\theta)^2 \implies \theta|x \sim Beta(4,3)$$

Bayesian inference: tractability notes

Conjugacy

We have conjugacy when the prior and the posterior distributions are in the same family (e.g. in the previous example, the prior and posterior are beta distributions).

Generally

Generally, computing the posterior distribution is much harder than in this example!

Consider the denominator in $p(\theta|x) = \dfrac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)}$ - integrals are hard

In nonconjugate examples, we need approaches to work with the posterior distribution when we cannot calculate it directly. Stay tuned!