CS340 Machine learning
Bayesian statistics 1

Fundamental principle of Bayesian statistics

- In Bayesian stats, everything that is uncertain (e.g., θ) is modeled with a probability distribution.
- We incorporate everything that is known (e.g., D) is by conditioning on it, using Bayes rule to update our prior beliefs into posterior beliefs.

$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

In praise of Bayes

- Bayesian methods are conceptually simple and elegant, and can handle small sample sizes (e.g., one-shot learning) and complex hierarchical models without overfitting.
- They provide a single mechanism for answering all questions of interest; there is no need to choose between different estimators, hypothesis testing procedures, etc.
- They avoid various pathologies associated with orthodox statistics.
- They often enjoy good frequentist properties.

Why isn’t everyone a Bayesian?

- The need for a prior.
- Computational issues.

The need for a prior

- Bayes rule requires a prior, which is considered “subjective”.
- However, we know learning without assumptions is impossible (no free lunch theorem).
- Often we actually have informative prior knowledge.
- If not, it is possible to create relatively “uninformative” priors to represent prior ignorance.
- We can also estimate our priors from data (*empirical Bayes).*
- We can use posterior predictive checks to test goodness of fit of both prior and likelihood.

Computational issues

- Computing the normalization constant requires integrating over all the parameters

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int p(\theta')p(D|\theta')d\theta'}$$

- Computing posterior expectations requires integrating over all the parameters

$$Ef(\Theta) = \int f(\theta)p(\theta|D)d\theta$$

Approximate inference

- We can evaluate posterior expectations using Monte Carlo integration

$$E f(\Theta) = \int f(\theta) p(\theta|D) d\theta \approx \frac{1}{N} \sum_{s=1}^{N} f(\theta^s) \quad \text{where } \theta^s \sim p(\theta|D)$$

- Generating posterior samples can be tricky
  - Importance sampling
  - Particle filtering
  - Markov chain Monte Carlo (MCMC)
- There are also deterministic approximation methods
  - Laplace
  - Variational Bayes
  - Expectation Propagation

Not on exam

Conjugate priors

- For simplicity, we will mostly focus on a special kind of prior which has nice mathematical properties.
- A prior $p(\theta)$ is said to be *conjugate* to a likelihood $p(D|\theta)$ if the corresponding posterior $p(\theta|D)$ has the same functional form as $p(\theta)$.
- This means the prior family is *closed under Bayesian updating.*
- So we can recursively apply the rule to update our beliefs as data streams in (online learning).
- A natural conjugate prior means $p(\theta)$ has the same functional form as $p(D|\theta)$.

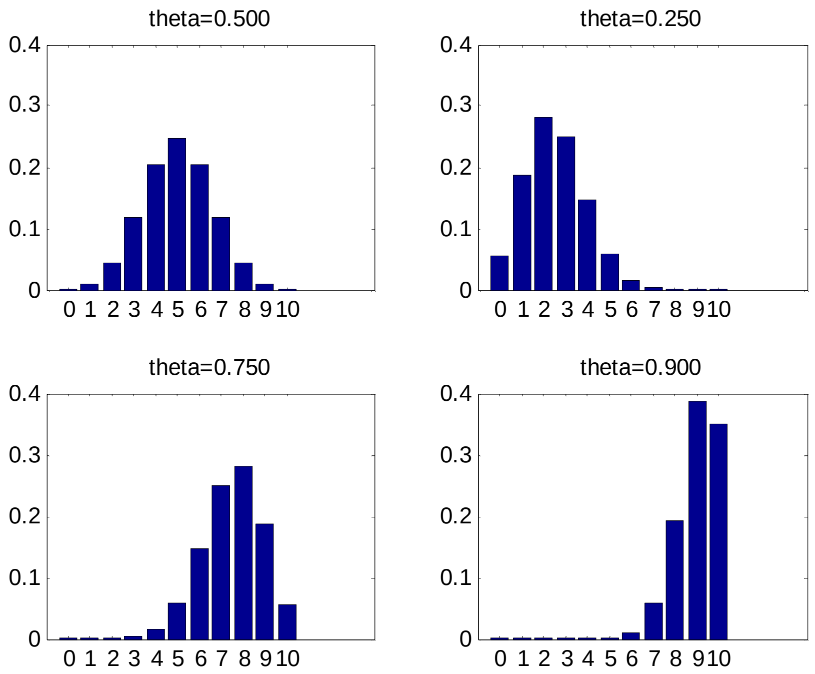
Example: coin tossing

- Consider the problem of estimating the probability of heads θ from a sequence of N coin tosses, D = $(X_1, \ldots, X_N)$
- First we define the likelihood function, then the prior, then compute the posterior. We will also consider different ways to predict the future.

Binomial distribution

- Let X = number of heads in N trials.
- We write X ~ Binom(θ, N).

$$P(X = x|\theta, N) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$



Bernoulli distribution

- Binomial distribution when N=1 is called the Bernoulli distribution.
- We write X ~ Ber(θ)

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

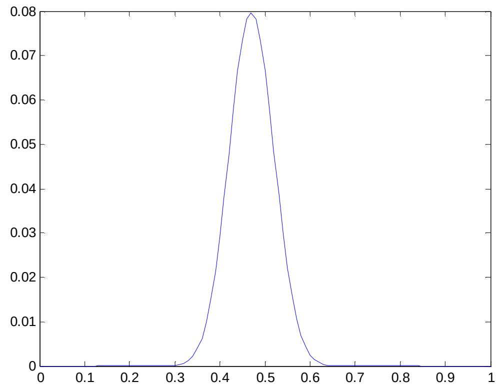
- So p(X=1) = θ, p(X=0) = 1-θ

Fitting a Bernoulli distribution

- Suppose we conduct N=100 trials and get data D = (1, 0, 1, 1, 0, ….) with N_1 heads and N_0 tails. What is θ?
- A reasonable best guess is the value that maximizes the likelihood of the data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$$

$$L(\theta) = p(D|\theta)$$



Bernoulli likelihood function

- The likelihood is

$$
\begin{aligned}
L(\theta) &= p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \\
&= \prod_{n} \theta^{I(x_n=1)} (1-\theta)^{I(x_n=0)} \\
&= \theta^{\sum_n I(x_n=1)} (1-\theta)^{\sum_n I(x_n=0)} \\
&= \theta^{N_1} (1-\theta)^{N_0}
\end{aligned}
$$

We say that N_0 and N_1 are sufficient statistics of D for θ

This is the same as the Binomial likelihood function, up to constant factors.

Bernoulli log-likelihood

- We usually use the log-likelihood instead

$$\begin{aligned}\ell(\theta) &= \log p(D|\theta) = \sum_n \log p(x_n|\theta) \\ &= N_1 \log \theta + N_0 \log(1-\theta)\end{aligned}$$

- Note that the maxima are the same, since log is a monotonic function

$$\arg\max L(\theta) = \arg\max \ell(\theta)$$

Computing the Bernoulli MLE

- We maximize the log-likelihood

$$
\begin{aligned}
\ell(\theta) &= N_1 \log \theta + N_0 \log(1-\theta) \\
\frac{d\ell}{d\theta} &= \frac{N_1}{\theta} - \frac{N-N_1}{1-\theta} \\
&= 0 \\
&\Rightarrow \\
\hat{\theta} &= \frac{N_1}{N}
\end{aligned}
$$

Empirical fraction of heads eg. 47/100

Black swan paradox

- Suppose we have seen N=3 white swans. What is the probability that swan X_{N+1} is black?
- If we plug in the MLE, we predict black swans are impossible, since $N_b=N_1=0$, $N_w=N_0=3$

$$\hat{\theta}_{MLE} = \frac{N_b}{N_b + N_w} = \frac{0}{N}, \quad p(X = b|\hat{\theta}_{MLE}) = \hat{\theta}_{MLE} = 0$$

- However, this may just be due to sparse data.
- Below, we will see how Bayesian approaches work better in the small sample setting.

The beta-Bernoulli model

- Consider the probability of heads, given a sequence of N coin tosses, $X_1, \ldots, X_N$.
- Likelihood

$$p(D|\theta) = \prod_{n=1}^{N} \theta^{X_n}(1-\theta)^{1-X_n} = \theta^{N_1}(1-\theta)^{N_0}$$

- Natural conjugate prior is the Beta distribution

$$p(\theta) = Be(\theta|\alpha_1, \alpha_0) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

- Posterior is also Beta, with updated counts

$$p(\theta|D) = Be(\theta|\alpha_1 + N_1, \alpha_0 + N_0) \propto \theta^{\alpha_1 - 1 + N_1}(1-\theta)^{\alpha_0 - 1 + N_0}$$

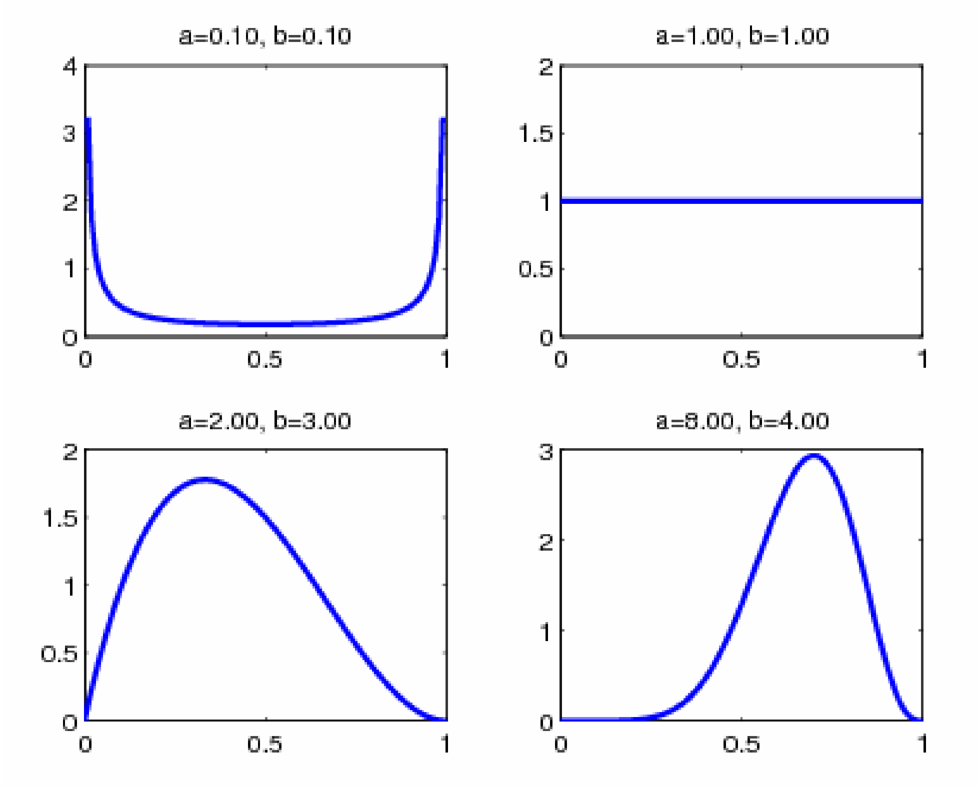
Just combine the exponents in θ and $(1-\theta)$ from the prior and likelihood

The beta distribution

- Beta distribution $p(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$
- The normalization constant is the beta function

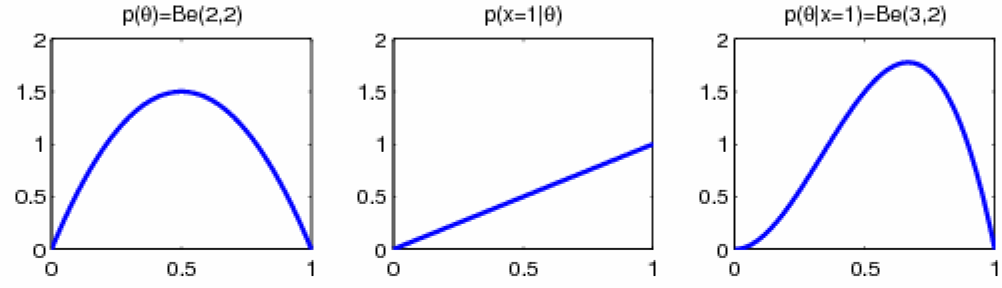
$$B(\alpha_1, \alpha_0) = \int_0^1 \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} d\theta = \frac{\Gamma(\alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha_1 + \alpha_0)}$$

$$E[\theta] = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

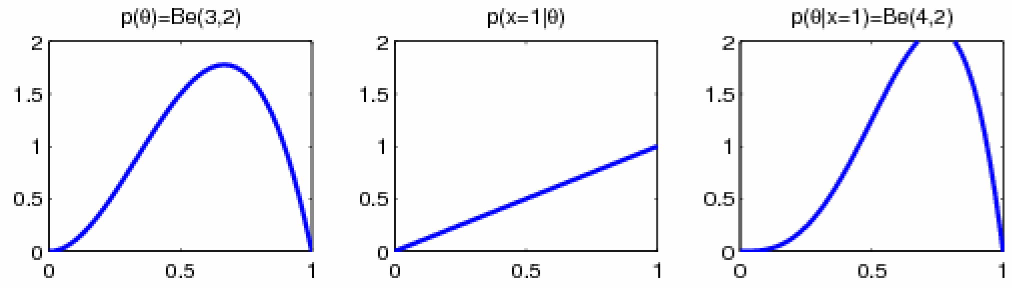


Updating a beta distribution

- Prior is Beta(2,2). Observe 1 head. Posterior is Beta(3,2), so mean shifts from 2/4 to 3/5.



- Prior is Beta(3,2). Observe 1 head. Posterior is Beta(4,2), so mean shifts from 3/5 to 4/6.



Setting the hyper-parameters

- The prior *hyper-parameters* α_1, α_0 can be interpreted as *pseudo counts.*
- The *effective sample size* (strength) of the prior is $\alpha_1+\alpha_0$.
- The prior mean is $\alpha_1/(\alpha_1+\alpha_0)$.
- If our prior belief is p(heads) = 0.3, and we think this belief is equivalent to about 10 data points, we just solve

$$\alpha_1 + \alpha_0 = 10, \quad \frac{\alpha_1}{\alpha_1 + \alpha_0} = 0.3$$

Point estimation

- The posterior $p(\theta|D)$ is our *belief state.*
- To convert it to a single best guess (point estimate), we pick the value that minimizes some loss function, e.g., MSE -> posterior mean, 0/1 loss -> posterior mode

$$\hat{\theta} = \arg\min_{\theta'} \int L(\theta', \theta) p(\theta|D) d\theta$$

- There is no need to choose between different estimators. The bias/ variance tradeoff is irrelevant.

Posterior mean

- Let $N=N_1 + N_0$ be the amount of data, and $M=\alpha_0+\alpha_1$ be the amount of virtual data.

The posterior mean is a convex combination of prior mean α_1/M and MLE N_1/N

$$
\begin{aligned}
E[\theta|\alpha_1, \alpha_0, N_1, N_0] &= \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_0 + N_0} = \frac{\alpha_1 + N_1}{N + M} \\
&= \frac{M}{N+M}\frac{\alpha_1}{M} + \frac{N}{N+M}\frac{N_1}{N} \\
&= w\frac{\alpha_1}{M} + (1-w)\frac{N_1}{N}
\end{aligned}
$$

$w = M/(N+M)$ is the strength of the prior relative to the total amount of data

We *shrink* our estimate away from the MLE towards the prior (a form of regularization).

MAP estimation

- It is often easier to compute the posterior mode (optimization) than the posterior mean (integration).
- This is called maximum a posteriori estimation.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$

- This is equivalent to penalized likelihood estimation.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(D|\theta) + \log p(\theta)$$

- For the beta distribution,

$$MAP = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_0 - 2}$$

Posterior predictive distribution

- We integrate out our uncertainty about θ when predicting the future (hedge our bets)

$$p(X|D) \quad = \quad \int p(X|\theta)p(\theta|D)d\theta$$

- If the posterior becomes peaked

$$p(\theta|D) \rightarrow \delta(\theta - \hat{\theta})$$

we get the *plug-in principle.*

$$p(x|D) = \int p(x|\theta)\delta(\theta - \hat{\theta})d\theta = p(x|\hat{\theta})$$

Posterior predictive distribution

- Let α_i' = updated hyper-parameters.
- In this case, the posterior predictive is equivalent to plugging in the posterior mean parameters

$$
\begin{aligned}
p(X=1|D) &= \int_0^1 p(X=1|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta \, \text{Beta}(\theta|\alpha_1', \alpha_0')d\theta = E[\theta] = \frac{\alpha_1'}{\alpha_0' + \alpha_1'}
\end{aligned}
$$

- If $\alpha_0 = \alpha_1 = 1$, we get *Laplace's rule of succession* (add one smoothing)

$$
p(X=1|N_1, N_0) = \frac{N_1 + 1}{N_1 + N_0 + 2}
$$

Solution to black swan paradox

- If we use a Beta(1,1) prior, the posterior predictive is

$$p(X = 1|N_1, N_0) = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

  so we will never predict black swans are impossible.
- However, as we see more and more white swans, we will come to believe that black swans are pretty rare.