1 MLE for the Bernoulli/ binomial model

$$X_i \sim Ber(\theta) \tag{1}$$
$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{2}$$

$$\begin{aligned}
\ln\big(p(D|\theta)\big) &= \ln\Big(\theta^{N_1}(1-\theta)^{N_0}\Big) \\
&= \ln\Big(\theta^{N-1}\Big) + \ln(1-\theta)^{N_0} \\
&= N_1 \ln\theta + N_0 \ln(1-\theta) \\
\tfrac{\mathrm{d}}{\mathrm{d}\theta} \ln p(D|\theta) &= \frac{N_1}{\theta} - \frac{N_0}{1-\theta}
\end{aligned}$$

The log-likelihood will take a maximum when the derivative equals 0.

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

2 Marginal likelihood for Beta-Bernoulli model

$$p(X_{1:N}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2})...p(x_N|x_{N-1}) \tag{3}$$

$$p(X=k|D_{1:N}) = \frac{N_k + \alpha_k}{\sum_i N_i + \alpha_i} \tag{4}$$

$$(\alpha-1)! = \Gamma(\alpha) \tag{5}$$

Given $D = H, T, T, H, H \overset{\triangle}{=} 1, 0, 0, 1, 1$

$$\begin{aligned}
p(X=1|\alpha) &= \frac{\alpha_1}{\alpha} \\
p(X=0|\alpha, D_1) &= \frac{\alpha_0}{\alpha+1} \\
p(X=0|\alpha, D_{1:2}) &= \frac{\alpha_0+1}{\alpha+2} \\
p(X=1|\alpha, D_{1:3}) &= \frac{\alpha_0+1}{\alpha+3} \\
p(X=1|\alpha, D_{1:4}) &= \frac{\alpha_0+2}{\alpha+4}
\end{aligned}$$

$$\begin{aligned}
p(D) &= p(D_{1:5}) \\
&= p(D_1)\cdot p(D_2|D_1)\cdot p(D_3|D_{1:2})\cdot p(D_4|D_{1:3})\cdot p(D_5|D_{1:4}) && \text{by (3)} \\
&= \frac{\alpha_1}{\alpha}\cdot\frac{\alpha_0}{\alpha+1}\cdot\frac{\alpha_0+1}{\alpha+2}\cdot\frac{\alpha_1+1}{\alpha+3}\cdot\frac{\alpha_1+2}{\alpha+4} \qquad \text{by (4)} \\
&= \frac{\big[\alpha_1(\alpha_1+1)(\alpha_1+2)\big]\,\big[\alpha_0(\alpha_0+1)\big]}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)} \\
&= \frac{\big[(\alpha_1)...(\alpha_1+N_1-1)\big]\,\big[(\alpha_0)...(\alpha_0+N_0-1)\big]}{(\alpha)...(\alpha+N-1)} \\
&= \frac{(\alpha_1+N_1-1)!}{(\alpha_1-1)!}\cdot\frac{(\alpha_0+N_0-1)!}{(\alpha_0-1)!}\cdot\frac{(\alpha-1)!}{(\alpha+N-1)!} \\
&= \frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)}\cdot\frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0)}\cdot\frac{\Gamma(\alpha)}{\Gamma(\alpha+N)} \\
&= \frac{\Gamma(\alpha_1+N_1)\,\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0+\alpha_1+N)}
\end{aligned}$$

3 Posterior predictive for a Beta-Binomial model

$$\begin{aligned}
p(x|n,D) &= Bb(x|\alpha_0', \alpha_1', n) \\
&= \frac{B(x+\alpha_1', n-x+\alpha_0')}{B(\alpha_1', \alpha_0')}\binom{n}{x}
\end{aligned}$$

Given $n=1$

$$\begin{aligned}
Bb(1|\alpha_0, \alpha_1, 1) &= \frac{B(1+\alpha_1, \alpha_0)}{B(\alpha_1, \alpha_0)} \binom{1}{1} \\
&= \frac{\Gamma(1+\alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha_0+\alpha_1+1)} \cdot \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \\
&= \frac{\alpha_1\Gamma(\alpha_1)\Gamma(\alpha_0)}{(\alpha_0+\alpha_1)\Gamma(\alpha_0+\alpha_1)} \cdot \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \\
&= \frac{\alpha_1}{\alpha_0+\alpha_1} \\
&= \frac{\alpha_1}{\alpha}
\end{aligned}$$

4 Beta updating from censored likelihood

Let n represent the number of coin tosses. Let X represent the number of heads. Given $n = 5$ and $X < 3$, we need to compute the posterior $p(\theta|X < 3)$ under a $B(1, 1)$ prior up to normalization constants.

$$\begin{aligned}
P(\theta) &= \frac{P(\theta)P(D|\theta)}{P(D)} \\
&= \frac{P(\theta) \cdot P(X < 3|\theta)}{P(X < 3)} \\
P(\theta) &\propto P(\theta) \cdot P(X < 3) \\
&\propto B(1, 1) \cdot \sum_{k=0}^{2} P(k|\theta, 5) \\
&\propto \sum_{k=0}^{2} \binom{5}{k} \theta^k (1-\theta)^{5-k}
\end{aligned}$$

5 Uninformative prior for log-odds ratio

Let $\phi = log\frac{\theta}{1-\theta}$. $p(\phi) = 1$ is equivalent to $p(\phi) = k$, where k is a constant and $0 < k < 1$.

$$\int_\phi p(\phi) d\phi = \int_\phi k d\phi = 1$$

$$d\phi = \frac{d\phi}{d\theta} d\theta$$

$$\frac{d\phi}{d\theta} = \frac{d}{d\theta}\left(ln \frac{\theta}{1-\theta}\right)$$

$$= \frac{d}{d\theta}\big(\ln\theta - \ln(1-\theta)\big)$$

$$= \frac{1}{\theta} - \frac{1}{1-\theta} \cdot -1$$

$$= \frac{1}{\theta} + \frac{1}{1-\theta}$$

$$\int_\phi k d\phi = \int_\theta k(\frac{1}{\theta} + \frac{1}{1-\theta}) d\theta$$

$$1 = k \int_\theta \theta^{-1}(1-\theta)^{-1} d\theta$$

We recognize the final integral as the normalization constant for a $Beta(\theta|0,0)$ distribution.

6 MLE for the Poisson distribution

$$P(X = k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k \in \{0, 1, 2...\}$$

$$p(x|\lambda) = e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} \cdot ...$$

$$= \prod_{i=1}^{N} e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$$

Now we take the log-likelihood and find its maximum.

$$\begin{aligned}
\ell(\lambda) &= \ln p(x|\lambda) \\
&= \ln\left(\prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \\
&= \ln\left(\frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\right) ... \\
&= \sum_{i=1}^{N} \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \\
&= \sum_{i=1}^{N} \ln(e^{-\lambda}\lambda^{x_i}) - \ln(x_i!) \\
&= \sum_{i=1}^{N} \ln(e^{-\lambda}) + \ln(\lambda^{x_i}) - \ln(x_i!) \\
&= \sum_{i=1}^{N} -\lambda + x_i \ln\lambda - \ln(x_i!) \\
&= -N\lambda + (\ln\lambda)\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \ln(x_i!)
\end{aligned}$$

$$\begin{aligned}
\ell'(\lambda) &= \frac{1}{\lambda}\sum_{i=1}^{N} x_i - N = 0 \\
\frac{\sum_{i=1}^{N} x_i}{\lambda} &= N \\
\lambda_{MLE} &= \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

7 Bayesian analysis of the Poisson distribution

7.1 a.

$$\begin{aligned}
p(\lambda|D) &= \frac{p(D|\lambda)p(\lambda)}{p(D)} \\
&\propto p(D|\lambda)p(\lambda) \\
&= \frac{e^{-N\lambda} \cdot \lambda^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!} \cdot \frac{\lambda^{a-1}e^{-\lambda b}}{k} \\
&\propto e^{-N\lambda-\lambda b} \cdot \lambda^{a-1+\sum_{i=1}^N x_i} \\
&= e^{-\lambda(N+b)} \cdot \lambda^{\left[a+\sum_{i=1}^N x_i\right]-1} \\
&= Ga(\lambda|a+\sum_{i=1}^N x_i, N+b)
\end{aligned}$$

7.2 b.

$$\begin{aligned}
\frac{a+\sum_{i=1}^N x_i}{N+b} asa \to 0, b \to 0 \\
= \frac{\sum_{i=1}^N x_i}{N} \\
= \lambda_{MLE}
\end{aligned}$$

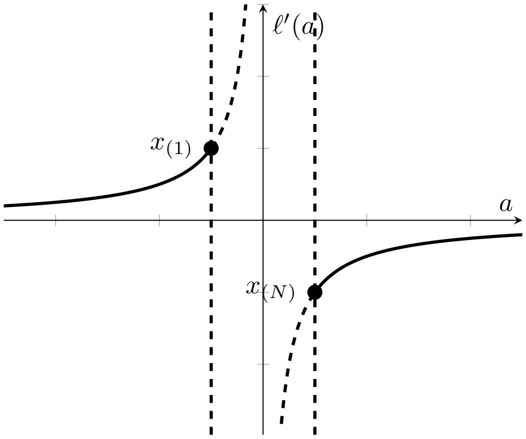
8 MLE for the uniform distribution

8.1 a.

$$\begin{aligned}
p(D|a) &= \prod_{i=1}^N p(x) \\
&= \left(\frac{1}{2a}\right)^N I(x_1, x_2, ..., x_N \in [-a, a]) \\
&= \begin{cases} \left(\frac{1}{2a}\right)^N & \text{if } -a \leq x_i \leq a, \forall i \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Now we take the derivative of the log-likelihood.

$$\begin{aligned}\ell(a) &= \ln\left(\frac{1}{2a}\right)^N \\ &= \ln 1 - \ln(2a)^N \\ &= -N\ln(2a) \\ \ell'(a) &= -N\frac{1}{2a}\cdot 2 \\ &= -\frac{N}{a}\end{aligned}$$



For $a < 0$, the likelihood is increasing, so it will be maximized where $a = x_{(1)}$, assuming that $|x_{(1)}| \geq |x_{(N)}|$. Similarly, for $a > 0$, the likelihood is decreasing, so it is maximized where $a = x_{(n)}$, assuming that $|x_{(N)}| \geq |x_{(1)}|$. This function is only defined when $\max_i |x_i| \leq a$. This means that ℓ is maximized at $a = \max(|x_{(1)}|, |x_{(n)}|)$.

8.2 b.

$p(x_{n+1}) = \frac{1}{b-a} = \frac{1}{2a}$

8.3 c.

Our approach is not Bayesian, so we will assign zero probability to $x_{n+1} > a$ and $x_{n+1} < -a$. A better solution would be to derive $\hat{a}_{MAP}$ and give a plug-in approximation.

9 Bayesian analysis of the uniform distribution

We must derive the posterior, $p(D|\theta)$ given the following:

$$p(D,\theta) = \frac{Kb^K}{\theta^{N+K+1}} \mathbb{I}(\theta \geq \max(D,b)) \tag{6}$$

Let $m = \max(D)$.

$$\begin{aligned} p(D) &= \int_m^\infty \frac{Kb^K}{\theta^{N+K+1}}\, d\theta \\ &= \begin{cases} \frac{K}{(N+K)b^N} & \text{if } m \leq b \\ \frac{Kb^K}{(N+K)m^{N+K}} & \text{if } m > b \end{cases} \end{aligned} \tag{7}$$

$$\begin{aligned} p(\theta|D) &= \frac{p(\theta,D)}{p(D)} \\ &= \begin{cases} \frac{Kb^K}{\theta^{N+K+1}} \cdot \frac{(N+K)b^N}{K} & \text{if } m \leq b \leq \theta \\ \frac{Kb^K}{\theta^{N+K+1}} \cdot \frac{(N+K)m^{N+K}}{Kb^K} & \text{if } b < m \leq \theta \end{cases} \\ &= \begin{cases} (N+K) \cdot b^{N+K} \cdot \theta^{-(N+K+1)} & \text{if } m \leq b \leq \theta \\ (N+K) \cdot m^{N+K} \cdot \theta^{-(N+K+1)} & \text{if } b < m \leq \theta \end{cases} \\ &\propto \begin{cases} \text{Pareto}(\theta|N+K, b) & \text{if } m \leq b \leq \theta \\ \text{Pareto}(\theta|N+K, m) & \text{if } b < m \leq \theta \end{cases} \\ &= \text{Pareto}(\theta|N+K, \max(m,b)) \end{aligned}$$

10 Taxicab (tramcar) problem

$$\text{Pareto}(\theta|N+K, \max(m,b)) \tag{8}$$

10.1 a.

Given a non-informative prior, Pareto$(\theta|0,0)$:

$$\begin{aligned} p(\theta|D) &= \text{Pareto}(\theta|1+0, \max(100,0)) \\ &= \text{Pareto}(\theta|1, 100) \end{aligned}$$

10.2 b.

$E[\theta|D] = \frac{km}{k-1}$ and $k = 1$, so the posterior mean is not defined.
The mode is $\max(D) = 100$.

$$
\begin{aligned}
\int_{100}^{x} km^k\theta^{-(k+1)}\,d\theta &= \frac{1}{2}\\
100\int_{100}^{x} \theta^{-2}\,d\theta &= \frac{1}{2}\\
\left[-\frac{1}{\theta}\right]_{100}^{x} &= \frac{1}{200}\\
-\frac{1}{x}+\frac{1}{100} &= \frac{1}{200}\\
x &= 200
\end{aligned}
$$

The median is 200.

10.3 c.

$$
\begin{aligned}
p(D'|D,\alpha) &= \int_\theta p(D'|\theta)p(\theta|D,\alpha)\,d\theta)\\
&= \int_\theta \frac{1}{\theta}\mathbb{I}(x\le\theta)\cdot N\cdot m^N\cdot\theta^{-(N+1)}\,d\theta\\
&= Nm^N\int_\theta \theta^{-(N+2)}\,d\theta\\
&= Nm^N\left[-\frac{1}{N-1}\theta^{-N-1}\right]_{\max(x,m)}^{\infty}\\
&= Nm^N\left[0-\left(-\frac{1}{N-1}\max(x,m)^{-N-1}\right)\right]\\
&= \frac{Nm^N}{(N+1)\max(x,m)^{N+1}}\\
&= \begin{cases} \frac{N}{m(N+1)} & \text{if } x<m\\ \frac{Nm^N}{x^{N+1}(N+1)} & \text{if } x\ge m \end{cases}
\end{aligned}
$$

10.4 d.

$$
\begin{aligned}
p(x=100|D,\alpha) &= \frac{1\cdot 100^1}{(1+1)100^{1+1}} = \frac{1}{200}\\
p(x=50|D,\alpha) &= \frac{1}{(1+1)100} = \frac{1}{200}\\
p(x=150|D,\alpha) &= \frac{1\cdot 100^1}{(1+1)150^{1+1}} = \frac{1}{450}
\end{aligned}
$$

10.5 e.

To improve the accuracy we could use a more informative prior. For example, we could estimate the number of cabs based on the city's population. Accuracy will also improve with more observations.

11 Bayesian analysis of the exponential distribution

11.1 a.

$$\begin{aligned}
p(x|\theta) &= \theta e^{-\theta x} \quad \text{for } x \geq 0, \theta \geq 0 \\
p(D|\theta) &= \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \theta e^{-\theta x_i}
\end{aligned}$$

$$\begin{aligned}
\ln(p(D|\theta)) &= \sum_{i=1}^{N} \ln \theta e^{-\theta x_i} \\
&= \sum_{i=1}^{N} \ln \theta + \ln e^{-\theta x_i} \\
&= \sum_{i=1}^{N} \ln \theta + -\theta x_i \\
\frac{d}{d\theta}(\sum_{i=1}^{N} \ln \theta + -\theta x_i) &= \sum_{i=1}^{N} \frac{1}{\theta} - x_i \\
0 &= \frac{N}{\theta} - \sum_{i=1}^{N} x_i \\
\hat{\theta} &= \frac{N}{\sum_{i=1}^{N} x_i} \\
\hat{\theta} &= \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i}
\end{aligned}$$

11.2 b.

$$\begin{aligned}\hat{\theta} &= \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i} \\ &= \frac{1}{\frac{1}{3}(5+6+4)} \\ &= \frac{1}{5}\end{aligned}$$

11.3 c.

$$\begin{aligned}p(\theta) &= \text{Expon}(\theta|\lambda) \\ &= \lambda e^{-\lambda\theta} \\ &= \frac{\lambda^1}{\Gamma(1)}\theta^{1-1}e^{-\lambda\theta} \\ &= \text{Ga}(\theta|1,\lambda)\end{aligned}$$

The mean of the Gamma distribution is $\frac{1}{\lambda}$, so $\hat{\lambda} = 3$

11.4 d.

$$p(\theta|D,\hat{\lambda}) \propto p(D|\theta)p(\theta|\hat{\lambda})$$

$$\begin{aligned}p(D|\theta) &= \prod_{x=1}^{N} \theta e^{-\theta x_i} \\ &= \theta e^{-\theta x_1} \cdot \theta e^{-\theta x_2} \cdot \theta e^{-\theta x_3} ... \\ &= \theta^N e^{(-\theta x_1 - \theta x_2 - \theta x_3 ...)} \\ &= \theta^N e^{-\theta \sum x_i}\end{aligned}$$

$$p(\theta|\hat{\lambda}) = \hat{\lambda} e^{-\theta\hat{\lambda}}$$

$$\begin{aligned}p(\theta|D,\hat{\lambda}) &\propto \theta^N e^{-\theta\sum x_i}\hat{\lambda}e^{-\theta\hat{\lambda}} \\ &\propto \theta^N e^{-\theta(\hat{\lambda}+\sum x_i)} \\ &= \text{Ga}(\theta|N+1, \hat{\lambda} + \sum_{i=1}^{N} x_i)\end{aligned}$$

11.5 e.

Yes, the prior is equivalent to a Gamma distribution and the posterior is also a Gamma distribution.

11.6 f.

The mean of a $\text{Ga}(\theta|a, b)$ distribution is $\frac{a}{b}$. The mean of $\text{Ga}(\theta|N + 1, \hat{\lambda} + \sum_{i=1}^{N} x_i)$ $= \frac{N+1}{\hat{\lambda}+\sum_{i=1}^{N} x_i}$

11.7 g.

The posterior accounts for the exponential prior. The prior accounts for expert knowledge, thus it is more reasonable.

12 MAP estimate for the Bernoulli with non-conjugate priors

12.1 a.

We're looking for the MAP estimate for θ, defined as $\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}} p(\theta|D)$

We can calculate the posterior up to normalization constants given the number of occurrences of heads and tails and the piecewise function that defines the prior.

$$\begin{aligned}
p(\theta|D) &\propto p(D|\theta) \cdot p(\theta) \\
&= \theta^{N_1}(1-\theta)^{N_0} \cdot \begin{cases} .5 & \text{if } \theta = .4 \text{ or } .5 \\ 0 & \text{else} \end{cases} \\
&= \begin{cases} \theta^{N_1}(1-\theta)^{N-N_1} & \text{if } \theta \in \{.4, .5\} \\ 0 & \text{else} \end{cases} \\
&= \begin{cases} .4^{N_1}(.6)^{N-N_1} & \text{if } \theta = .4 \\ .5^N & \text{if } \theta = .5 \\ 0 & \text{else} \end{cases}
\end{aligned}$$

The MAP estimate is the value of θ that maximizes the equation above.

$$\hat{\theta}_{MAP} = \begin{cases} .4 & \text{if } (.4)^{N_1}(.6)^{N-N_1} \geq .5^N \\ .5 & \text{else} \end{cases}$$

12.2 b.

If N is small, $\theta = .4$ will lead to a better estimate since the prior is close to the true value. As N grows, the data will overwhelm the prior.

13 Posterior predictive distribution for a batch of data with the Dirichlet-multinomial model

$$\begin{aligned}
p(D'|D,\alpha) &= \int_\theta p(D'|\theta)\cdot p(\theta|D)d\theta \\
&= \int_\theta \mathrm{Mu}(N_{new_1}, N_{new_2}, ...|\theta)\cdot \mathrm{Dir}(\theta|N_{old_1}+\alpha_1, N_{old_2}+\alpha_2 ...) \\
&= \frac{1}{\mathrm{B}(\alpha+N_{old})}\left(\frac{N_{new}!}{N_{new_1}!...N_{new_k}!}\right)\int_\theta \prod^k \theta_k^{N_{new_k}} \cdot \prod^k \theta_k^{\alpha_k+N_{old_k}-1}\, d\theta \\
&= \frac{1}{\mathrm{B}(\alpha+N_{old})}\left(\frac{N_{new}!}{N_{new_1}!...N_{new_k}!}\right)\int_\theta \prod^k \theta_k^{\alpha_k+N_{new_k}+N_{old_k}-1}\, d\theta
\end{aligned}$$

Note that the integral is the normalization constant for a Dirichlet distribution, $\mathrm{Dir}(\alpha + \mathbf{N_{new}} + \mathbf{N_{old}})$.

$$\begin{aligned}
&= \left(\frac{N_{new}!}{N_{new_1}!...N_{new_k}!}\right)\frac{B(\alpha+N_{old}+N_{new})}{B(\alpha+N_{old})} \\
&= \frac{N_0'!}{\prod N_k'!}\cdot\frac{\prod\Gamma(\alpha_k+N_k+N_k')}{\Gamma(\alpha_0+N_0+N_0')}\cdot\frac{\Gamma(\alpha_0+N_0)}{\prod\Gamma(\alpha_k+N_k)} \\
&\qquad \text{where } \alpha_0 = \sum_{i=1}^k \alpha_i, N_0 = \sum_{i=1}^k N_i, \text{ and } N_0' = \sum_{i=1}^k N_i' \\
&= \frac{N_0'!}{\prod N_k'!}\cdot\frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0+N_0+N_0')}\cdot\frac{\prod\Gamma(\alpha_k+N_k+N_k')}{\prod\Gamma(\alpha_k+N_k)}
\end{aligned}$$

Notice that this looks like the formula we derived in exercise 2.

14 Posterior predictive for Dirichlet-multinomial

14.1 a.

$$\begin{aligned} p(X = j|D) &= E[\theta_j|D] \qquad \text{Equation (3.51) in the textbook} \\ &= \frac{\alpha_j + N_j}{\alpha_0 + N} \\ &= \frac{10 + 260}{(10 \cdot 27) + 2000} \\ &= .119 \end{aligned}$$

14.2 b.

We use the equation from **??**.

$$p(D'|D, \alpha) = \frac{N_0'!}{\prod N_k'!} \cdot \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N_0')} \cdot \frac{\prod \Gamma(\alpha_k + N_k + N_k')}{\prod \Gamma(\alpha_k + N_k)}$$

For a single trial, $N_0' = N_j' = 1$, so the first term is equal to 1. Now we examine the last term.

$$\begin{aligned} p(x = j|D, \alpha) &= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N_0')} \cdot \frac{\prod_{k \neq j} \Gamma(\alpha_k + N_k + N_k')}{\prod_{k \neq j} \Gamma(\alpha_k + N_k)} \cdot \frac{\Gamma(\alpha_j + N_j + N_j')}{\Gamma(\alpha_j + N_j)} \\ &= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N_0')} \cdot \prod_{k \neq j} \frac{\Gamma(\alpha_k + N_k + 0)}{\Gamma(\alpha_k + N_k)} \cdot \frac{\Gamma(\alpha_j + N_j + N_j')}{\Gamma(\alpha_j + N_j)} \\ &= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N_0')} \cdot 1 \cdot \frac{\Gamma(\alpha_j + N_j + N_j')}{\Gamma(\alpha_j + N_j)} \\ &= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + 1)} \cdot \frac{\Gamma(\alpha_j + N_j + 1)}{\Gamma(\alpha_j + N_j)} \\ &= \frac{\Gamma(\alpha_0 + N_0)}{(\alpha_0 + N_0)\Gamma(\alpha_0 + N_0)} \cdot \frac{(\alpha_j + N_j)\Gamma(\alpha_j + N_j)}{\Gamma(\alpha_j + N_j)} \\ &= \frac{\alpha_j + N_j}{\alpha_0 + N_0} \end{aligned}$$

For independent samples

$$\begin{aligned} p(x_{2001} = a, x_{2002} = p|D, \alpha) &= p(x_{2001} = a|D, \alpha) \cdot p(x_{2002} = p|D', \alpha) \\ &= \frac{10 + 100}{270 + 2000} \cdot \frac{10 + 87}{270 + 2001} \\ &= .0021 \end{aligned}$$

15 Setting the beta hyper-parameters

First, we express a in terms of m and b.

$$\begin{aligned} m &= \frac{a}{a+b} \\ a &= m(a+b) \\ a(1-m) &= mb \\ a &= \frac{mb}{1-m} \end{aligned}$$

We can now solve for b in terms of m and v by plugging in this value of a into the equation for the variance of a Beta distribution with parameters a and b.

$$\begin{aligned} \text{var} &= \frac{ab}{(a+b)^2(a+b+1)} \\ &= \frac{mb^2}{1-m} \cdot \frac{1}{\left(\frac{mb}{1-m}+b\right)^2 \left(\frac{mb}{1-m}+b+1\right)} \\ &= \frac{mb^2}{1} \cdot \frac{(1-m)^2}{\big(mb+b(1-m)\big)^2 (mb+b-mb)+(1-m)} \\ v &= \frac{m(1-m)^2}{b-m+1} \end{aligned}$$

Now, solving for b...

$$\begin{aligned} v(b-m+1) &= m(1-m)^2 \\ vb - mv + v &= m(1-m)^2 \\ b &= \frac{m}{v}(1-m)^2 + m - 1 \end{aligned}$$

We can plug b back into our original equation for a.

$$\begin{aligned} a &= \frac{mb}{1-m} \\ &= \frac{m}{1-m} \cdot \left(\frac{m(1-m)^2}{b-m+1}\right) \\ &= \frac{m^2(1-m)}{v} + \frac{m^2}{1-m} + \frac{m}{1-m} \\ &= \frac{m^2}{v}(1-m) - m \end{aligned}$$

Now that we've solved for b and a in terms of m and v, all the remains is to plug in the values given for m and v.

$$a = \frac{(.7)^2}{.2^2}(1 - .7) - .7 = 2.975$$

$$b = \frac{.7}{.2^2}(1 - .7)^2 + .7 - 1 = 1.275$$

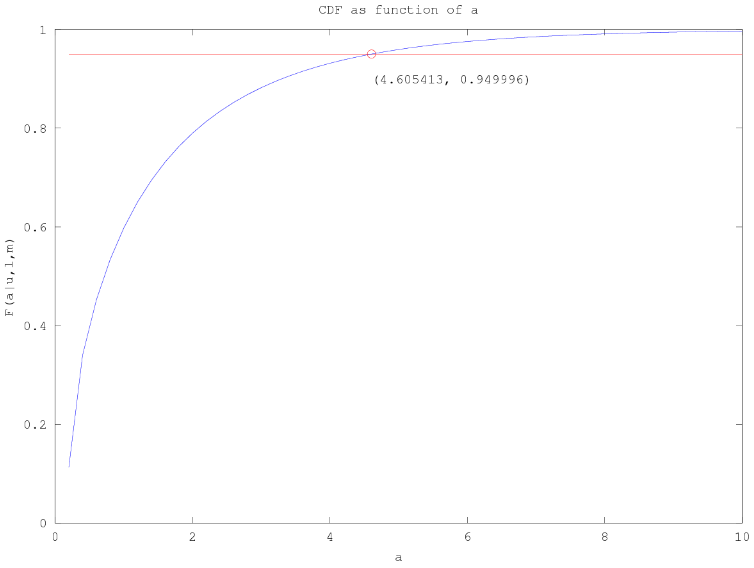
16 Setting the beta hyper-parameters II

We know that the cumulative probability density over the range $l \leq \theta \leq u = .95$. We can express the CDF as a function of a given u, l, and m.

$$\begin{aligned} F(a|u, l, m) &= \int_u^l p(\theta) d\theta \\ &= \frac{1}{B(a, b)} \int_u^l \theta^{a-1}(1-\theta)^{b-1} d\theta \\ &= \frac{1}{B(a, \frac{a(1-m)}{m})} \int_u^l \theta^{a-1}(1-\theta)^{\frac{a(1-m)}{m}-1} d\theta \end{aligned}$$

Now we want to find the value of a which minimizes the difference between this function and the value 0.95. Formally, we want to minimize $(F(a|u, l, m) - .95)^2$. We can then easily determine b given a and m using the formula for the mean of a Beta distribution.

The Octave code for this solution is included in the code/ directory in this repository. Using this code, we obtain $a = 4.605413; b = 26.097340$. This is roughly equivalent to a prior sample with 4.6 heads and 26.1 tails, so the equivalent sample size is approximately 30.7.



17 Marginal likelihood for beta-binomial under uniform prior

To obtain the marginal likelihood, we integrate the joint distribution over θ.

$$
\begin{aligned}
p(N_1|N) &= \int_\theta p(N_1|N,\theta) \cdot p(\theta|\alpha)\; d\theta \\
&= \int_\theta Bin(N_1|N,\theta) \cdot p(\theta|\alpha)\; d\theta \\
&= \int_\theta \frac{N!}{(N-N_1)!N_1!} \cdot \theta^{N_1}(1-\theta)^{N-N_1} \cdot \frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)} \cdot \theta^{1-1}(1-\theta)^{1-1} \\
&= \frac{N!}{(N-N_1)!N_1!} \int_\theta \theta^{N_1}(1-\theta)^{N-N_1} \cdot \frac{1!}{1 \cdot 1}
\end{aligned}
$$

We recognize the integral as the normalization constant for the Beta distribution Beta$(N_1+1, N-N_1+1)$.

$$\begin{aligned}
p(N_1|N) &= \frac{N!}{(N-N_1)!N_1!}B(N_1+1, N-N_1+1) \\
&= \frac{N!}{(N-N_1)!N_1!}\frac{\Gamma(N_1+1)\Gamma(N-N_1+1)}{\Gamma(N_1+N-N_1+2)} \\
&= \frac{N_1!(N-N_1)!N!}{(N-N_1)!N_1!(N+1)!} \\
&= \frac{N!}{(N+1)!} \\
&= \frac{1}{N+1}
\end{aligned}$$

18 Bayes factor for coin tossing

$$\begin{aligned}
BF_{1,0} &= \frac{\text{likelihood}_1}{\text{likelihood}_2} \\
&= \frac{p(D|\theta_1)}{p(D|\theta_2)}
\end{aligned}$$

The likelihood of a fair coin when $N = 10$ and $N_1 = 9$.

$$\begin{aligned}
p(D|\theta_1) &= \frac{N!}{(N-N_1)!N_1!} \cdot \theta^{N_1}(1-\theta)^{N-N_1}. \\
&= \frac{10!}{(10-9)!9!} \cdot (.5)^9(.5)^1 \\
&= \frac{5}{512} \\
&= 9.766 \times 10^{-3}
\end{aligned}$$

When $N = 100$ and $N_1 = 90$

$$\begin{aligned}
p(D|\theta_1) &= \frac{N!}{(N-N_1)!N_1!} \cdot \theta^{N_1}(1-\theta)^{N-N_1}. \\
&= \frac{100!}{(100-90)!90!} \cdot (.5)^{90}(.5)^{10} \\
&= 1.366 \times 10^{-17}
\end{aligned}$$

A uniform prior signifies that any bias is equally likely. Thus, we integrate the likelihood over every possible value of θ_2.

$$
\begin{aligned}
p(D|\theta_2) &= \int_0^1 p(D|\theta_2) \cdot p(\theta_2)\ d\theta_2 \\
&= \int_0^1 \frac{10!}{(10-9)!9!} \cdot \theta_2^9 (1-\theta)^1 \frac{1}{1-0} \cdot\ d\theta_2 \\
&= 10 \int_0^1 \theta_2^9 - \theta_2^{10}\ d\theta_2 \\
&= 10 \left[\frac{1}{10} \theta_2^{10} \right]_0^1 - 10 \left[\frac{1}{11} \theta_2^{11} \right]_0^1 \\
&= \frac{1}{11}
\end{aligned}
$$

Note that this is again the marginal likelihood, $\frac{1}{N+1}$, which we derived in 3.17. Thus, for $N = 100$ and $N_1 = 90$, $p(D|\theta_2) = \frac{1}{101} = .0099$.

So for 10 trials, $BF_{1,0} = \frac{1}{11} \cdot \frac{512}{5} = 9.31$. For 100 trials, $BF_{1,0} = \frac{\frac{1}{101}}{1.366 \times 10^{-17}} = 7.25 \times 10^{-14}$.

As we would expect, as the number of trials increases the likelihood that the coin is biased drastically increases, as does the Bayes Factor.

19 Irrelevant features with naive Bayes

19.1 a.

$$
\begin{aligned}
\log_2 p(C|\mathbf{x_i}) &= \log_2 \big(p(\mathbf{x_i}|C) p(C) \big) \\
&= \log_2 p(\mathbf{x_i}|C) + \log_2 p(C) \\
&= \phi(\mathbf{x_i})^T \beta_c + \log_2 p(C)
\end{aligned}
$$

$$
\begin{aligned}
\log_2 \frac{p(C=1|\mathbf{x_i})}{p(C=2|\mathbf{x_i})} &= \log_2 p(C=1|\mathbf{x_i}) - \log_2 p(C=2|\mathbf{x_i}) \\
&= \phi(\mathbf{x_i})^T \beta_1 + log(.5) - (\phi(\mathbf{x_i})^T \beta_2 + log(.5)) \\
&= \phi(\mathbf{x_i})^T \left[\beta_1 - \beta_2 \right]
\end{aligned}
$$

19.2 b.

The presence or absence of a word will have no effect on the class posterior when $\beta_{1,w} - \beta_{2,w} = 0$ (which implies that $\theta_{1,w} = \theta_{2,w}$).

19.3 c.

No, it will not be ignored.

A word is ignored by our classifier if $\theta_{c,w} = k \quad \forall c \in C$ where k is some constant. We are told that word w occurs in every document, so it obviously occurs in every document in class c for any value of c. Therefore, $x_{i,w} = 1 \quad \forall i$ and

$$\sum_{i \in c} x_{i,w} = \sum_{i \in c} 1 = n_c$$

where n_c is the count of documents in class c. Plugging into our equation for the posterior mean estimate, we get

$$\hat{\theta}_{c,w} = \frac{1 + n_c}{2 + n_c}$$

Clearly, $\hat{\theta}_{c,w}$ depends on n_c. For example, unless $n_1 = n_2$, then

$$\frac{1 + n_1}{2 + n_1} \neq \frac{1 + n_2}{2 + n_2}$$

19.4 d.

The simplest way to discard irrelevant terms would be to discard words which occur either in every document or in no document. Alternatively, we could rank the words according to their mutual information with the class label and discard terms with low information.

20 Class conditional densities for binary data

20.1 a.

Let us first suppose that $D = 2$. That is $\mathbf{x} \in \mathbb{R}^2$. We can represent $p(\mathbf{x}|y = c)$ in tabular form.

	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	$P(x_1 = 0, x_2 = 0)$	$P(x_1 = 0, x_2 = 1)$
$x_1 = 1$	$P(x_1 = 1, x_2 = 0)$	$P(x_1 = 1, x_2 = 1)$

Since we know that the class-conditional marginal probability $P(\mathbf{x}|y = c)$ must sum to 1, we can fully express the class-conditional probability given any 3 of the parameters above. Now, if we were to add a third dimension such that $\mathbf{x} \in \mathbb{R}^3$, we would double the number of parameters, since for each parameter in the 2-dimensional scenario, we would add two parameters: one corresponding to $x_3 = 0$ and one corresponding to $x_3 = 1$. So for $D = 3$, we have $8 = 2^3$ parameters, and we can express $p(\mathbf{x}|y = c)$ in terms of $2^3 - 1$ parameters.

More generally, in a full Bayesian model with D binary features we can express the class-conditional distribution $p(\mathbf{x}|y = c)$ in terms of $2^D - 1$ parameters.

20.2 b.

If the sample size is very small, a full Bayesian model is likely to overfit the training data. In this case, the naive model will probably give lower test set error.

20.3 c.

As the sample size grows, our model will be less prone to overfitting, so the full model may give better results if the number of features is very small. However, since our feature space grows exponentially with D, for large values of D we would need massive amounts of training data to ensure that we don't overfit.

20.4 d.

20.4.1 Full

Since the class prior is uniform, the MAP estimate reduces to the MLE, which is just the empirical ratio of the count of occurrences of $\mathbf{x}$ for examples in class c to the total number of examples in class c.

We can fit a model with N training examples, D features, and C classes as follows. For each class, for each training example in the class, for each feature, we calculate the index corresponding to $\mathbf{x}$ increment the count in the appropriate class-specific frequency table. This procedure requires $O(NDC)$ time.

The space requirements grow exponentially with the number of features, since each of the C frequency tables will consist of 2^D addresses.

20.4.2 Naive

For each of the D features we'll compute a class-conditional MAP estimate for feature j, $\hat{\theta}_{j,c}$. As with the full model, under a uniform prior the MAP is equivalent to the MLE which is simply the empirical ratio of the number of examples in class c where feature $x_j = 1$ to the total number of examples in class c.

The time complexity of fitting this model is $O(ND)$. Unlike with the full model, whose space requirements are exponential, the space requirements for the Naive model are $O(D)$.

20.5 e.

20.5.1 Full

Using the model for prediction requires that we find $\underset{c}{\text{argmax}} p(y = c|\mathbf{x})$. We are told to assume that we can convert $\mathbf{x}$ into an array index in $O(D)$ time, and we will repeat this for each of the C classes, so the time complexity is $O(DC)$.

20.5.2 Naive

For the Naive model, we need to take the product of D independent likelihoods for each of C classes. Thus, the time complexity is again $O(DC)$.

20.6 f.

20.6.1 Full

To account for the h missing features, we have to marginalize over them by summing over every possible value of $\mathbf{x}$. If we have h missing features, then we have 2^h possible values of $\mathbf{x}$ in our frequency table over which we must sum. This requires $O(C \cdot v \cdot 2^h)$ time.

20.6.2 Naive

Marginalizing over the missing h features is much easier in the Naive case. We can consider the features independently, so there is no combinatorial explosion. We simply sum over all possible values for each of the D features. Since the number of possible values is constant (2, since the features are binary), this still only requires $O(DC)$ time (or, in terms of h and v, $O(C \cdot (h + v))$).

21 Mutual information for naive Bayes classifiers with binary features

$$
\begin{aligned}
I(X,Y) &= \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \\
&= \sum_{x_j \in \{0,1\}} \sum_{y} p(x_j|y)p(y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \\
&= \sum_{c} p(x_j = 1|y = c)p(y = c) \log \frac{p(x_j = 1, y = c)}{p(x_j = 1)} + p(x_j = 0|y = c)p(y = c) \log \frac{p(x_j = 0, y = c)}{p(x_j = 0)} \\
&= \sum_{c} \left[\theta_{jc}\pi_c \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j} \right]
\end{aligned}
$$

22 Fitting a naive Bayes spam filter by hand

$$
\begin{aligned}
\hat{\theta}_{spam} &= \frac{3}{3+4} = \frac{3}{7} \\
\hat{\theta}_{secret|spam} &= \frac{2}{3} \\
\hat{\theta}_{secret|non-spam} &= \frac{1}{4} \\
\hat{\theta}_{sports|non-spam} &= \frac{2}{4} = \frac{1}{2} \\
\hat{\theta}_{dollar|non-spam} &= \frac{1}{3}
\end{aligned}
$$