6.790 Homework 1

Sept 10, 2024

Questions 1–3 are relatively stand-alone warm-ups. Questions 4–6 are more extended practice and illustrations of the ideas of this material. Question 7 requires coding. *Do not submit your code!*

There are some rhetorical questions in blue boxes. You don't need to answer them—they're just for thinking about.

Please hand in your work via Gradescope via the link at https://gradml.mit.edu/info/homeworks/. If you were not added to the course automatically, please use Entry Code R7RGGX to add yourself to Gradescope.

1. Latex is not required, but if you are hand-writing your solutions, please write clearly and carefully. You should include enough work to show how you derived your answers, but you don't have to give careful proofs.

2. Homework is due on Tuesday September 17 at 11PM.

3. Lateness and extension policies are described at https://gradml.mit.edu/info/class_policy/.

Contents

1 Normal fish [10 Points] 3

1.1 Fish tale . . . 3

1.2 A simple model . . . 5

1.3 A more complex model . . . 5

2 Parameter estimation [10 points] 5

2.1 Force field . . . 5

2.2 Pigeons . . . 6

3 Bayesian belief update [10 Points] 7

3.1 Beta-Binomial practice . . . 7

3.2 What's new? . . . 8

4 Which dice factory? [15 points] 9

5 Emergency Room [15 Points] 11

6 Abby Normal [15 Points] 15

7 Coding Question: Two Gaussians [25 Points] 18

8 Optional 20

8.1 Throwing rocks: asymmetric loss 20

8.2 Copy that: discrete Bayes update and decision theory 21

8.3 Dirichlet Priors 23

8.4 More fun with Dirichlet 26

Solution: Don't look at the solutions until you have tried your absolute hardest to solve the problems. This is especially true for optional problems that you didn't work on—it's a good idea to come back to them when studying for exams.

1 Normal fish [10 Points]

1.1 Fish tale

(Bishop 1.11) We find ourselves with a data set consisting of the measured weights of a bunch of fish caught during an afternoon of fishing. We decide to model the distribution of these weights using a Gaussian distribution.

Why might this not be a great modeling choice?

Solution:

Maybe they come from different species, so we could expect the distribution to be multi-modal. Also, the Gaussian has infinite tails, so it will assign positive probability to fish with negative weight.

Our goal is to select parameters μ, σ^2 of the Gaussian distribution in order to maximize the likelihood of our data, $\mathcal{D} = \{x^{(1)}, \dots, x^{(n)}\}$. The parameters that maximize the log likelihood of the data, will also maximize the likelihood (due to its monotonicity) and the form is easier to deal with. Recall that the pdf of a Gaussian distribution is given by

$$p_X(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \ .$$

If we assume that the process whereby we caught the fish made their weights independent and identically distributed, then

$$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_i p_X(x^{(i)} \mid \mu, \sigma^2) \ .$$

The log likelihood function is then

$$\log p(\mathcal{D} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi) \ .$$

By setting its derivatives with respect to μ and σ^2 equal to zero and solving , verify that the maximum likelihood estimates of μ and σ are given by

$$\mu_{\mathbf{ml}} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$$

$$\sigma^2_{\mathbf{ml}} = \frac{1}{N} \sum_{n=1}^{N} (x^{(n)} - \mu_{\mathbf{ml}})^2$$

.

Under what assumptions about the log likelihood function is this a valid approach for finding a global maximum?

This solution may be different than the estimator you have previously seen for σ^2. See the discussion at the bottom of Bishop page 27 for an explanation.

Solution: Taking the partial derivatives of the log likelihood with respect to μ and σ^2 results in

$$\frac{\partial \log \Pr(x \mid \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x^{(n)} - \mu)$$

$$\frac{\partial \log \Pr(x \mid \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^{N} (x^{(n)} - \mu)^2 - \frac{N}{2\sigma^2}$$

Setting the partial derivative with respect to μ to 0 gives

$$0 = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x^{(n)} - \mu) = \frac{1}{\sigma^2} \sum_{n=1}^{N} x^{(n)} - \frac{1}{\sigma^2} \sum_{n=1}^{N} \mu = \frac{1}{\sigma^2} \sum_{n=1}^{N} x^{(n)} - \frac{1}{\sigma^2} N\mu$$

so

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$$

For the Gaussian, we are fortunate that our estimate for the mean is independent of the variance.

Setting the partial derivative with respect to σ^2 (note – not σ) to 0 gives

$$0 = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^{N} (x^{(n)} - \mu)^2 - \frac{N}{2\sigma^2}$$

$$\frac{N}{2\sigma^2} = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^{N} (x^{(n)} - \mu)^2$$

so

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x^{(n)} - \mu)^2$$

We can relate this quantity back to by replacing μ with $\mu_{\mathbf{ml}}$ found above, since as we noticed, the $\mu_{\mathbf{ml}}$ does not depend on the variance.

We are finding a value for θ, the parameter to be estimated (here, (μ, σ^2)), such that $\nabla_\theta L(\theta) = 0$ where L is the loss function to be minimized (here, $-\log p$). From calculus class, we know that this is a necessary condition of θ being a local extremum of $L : U \to \mathbb{R}$ (where U is an open subset of $\mathbb{R}^n$). If the loss function L is convex, this is also a sufficient condition of θ being a global minimum of L.

1.2 A simple model

As it happens, we caught 6 mega-guppies (a tasty type of fish), with these weights:

$$\mathcal{D}_0 = \{0.9, 1, 1.1, 1.2, 3, 3.1\} \ .$$

We looked in the USDA handbook which told us that the variance of the weight of North American mega-guppies is $\sigma^2 = 0.5^2 = 0.25$.

Find the maximum likelihood value of $\mu_{\mathbf{ml}}$ for $\mathcal{D}_0$ under this assumption. What is the data likelihood $p(\mathcal{D}_0|\mu_{\mathbf{ml}})$?

Solution:

$$\mu_{\mathbf{ml}} = 1.71666$$

$$p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}) = 5.387577e - 06$$

$$\log(p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}})) = -12.131415$$

1.3 A more complex model

Now, what if we ignore the USDA value of σ^2 and decide to estimate it ourselves? Find the maximum likelihood estimates $\mu_{\mathbf{ml}}$ and $\sigma^2_{\mathbf{ml}}$ of μ and σ^2 for our data set $\mathcal{D}_0$. What is the data likelihood $p(\mathcal{D}_0|\mu_{\mathbf{ml}}, \sigma^2_{\mathbf{ml}})$?

What are the advantages and disadvantages of this model versus the one with the fixed variance?

Solution:

$$\mu_{\mathbf{ml}} = 1.716666$$

$$\sigma^2_{\mathbf{ml}} = 0.898055$$

$$p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}, \sigma_{\mathbf{ml}}) = 0.000277$$

$$\log(p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}, \sigma_{\mathbf{ml}})) = -8.191061$$

This new model fits the data better. But it might be that it would have been better to use the other variance because it was based on a larger sample. But it might have been better to use our estimate because the local population of fish has a different distribution. We will spend a lot of time in class thinking about how to make trade-offs like this. The problem is called *model selection*.

2 Parameter estimation [10 points]

2.1 Force field

A supervillain has our hero trapped in an invisible one-dimensional force-field (hero can only move in one dimension) and we know that the field has finite extent. Using a drone flying overhead, we make several measurements of the hero's position.

We wish to estimate the boundaries of the force-field given samples of the hero's position.

If we knew that our data are drawn uniformly from a finite interval, $[a, b]$, then we might want to find $a_{\mathbf{ml}}, b_{\mathbf{ml}}$ to maximize the likelihood of $\mathcal{D}$.

For our data set $\mathcal{D} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$, what are the maximum likelihood parameter estimates $a_{\mathbf{ml}}$ and $b_{\mathbf{ml}}$? What is the data likelihood $p(\mathcal{D}|a_{\mathbf{ml}}, b_{\mathbf{ml}})$?

Is this model of the hero data a good one? Why or why not?

Solution:

It might not be good if we have reason to think that for the hero, some parts of the force field are more interesting or comfortable than others. Also, if we are sampling positions finely in time, they will be very highly correlated with one another (not iid).

Solution:

The likelihood of the data is:

$$\prod_{i=1}^{n} \begin{cases} (b_{\mathbf{ml}} - a_{\mathbf{ml}})^{-1} & \text{if } a_{\mathbf{ml}} \leqslant x^i \leqslant b_{\mathbf{ml}} \\ 0 & \text{otherwise} \end{cases}$$

We can see that if $a_{\mathbf{ml}} > x^i$ or $b_{\mathbf{ml}} < x^i$, for any x^i, then the likelihood of the whole data set must be 0. So, we should pick $b_{\mathbf{ml}}$ to be as small as possible subject to the constraint that $b_{\mathbf{ml}} \geqslant x^i$, which means $b_{\mathbf{ml}} = \max_i x^i$. Similarly, $a_{\mathbf{ml}} = \min_i x^i$.

For $\mathcal{D}_0 = \{0.9, 1, 1.1, 1.2, 3, 3.1\}$ (the data from the previous question):

$$a_{\mathbf{ml}} = 0.9 \quad b_{\mathbf{ml}} = 3.1$$

$$p(\mathcal{D}_0 \mid a_{\mathbf{ml}}, b_{\mathbf{ml}}) = 0.008820$$

$$\log(p(\mathcal{D}_0 \mid a_{\mathbf{ml}}, b_{\mathbf{ml}})) = -4.730744$$

2.2 Pigeons

Pigeons[1], when put in a situation where $\Pr(y = 1) = p$ and $\Pr(y = 0) = 1 - p$, will select option 1 with probability p and option 0 with probability $1 - p$. What is the expected 0-1 loss for the pigeons' decision rule? What is the optimal decision rule and its expected loss?

Actually, people[2] do this too!

[1] "Probability-Matching in the Pigeon", Donald H. Bullock and M. E. Bitterman, *The American Journal of Psychology* , Vol. 75, No. 4 (Dec., 1962), pp. 634-639

[2] "Banking on a Bad Bet: Probability Matching in Risky Choice is Linked to Expectation Generation," *Psychological Science*, Vol. 22, No. 6 (2011).

Solution: The loss is 0 when the pigeon's random choice g agrees with the independent draw from the underlying distribution y. So, the loss is:

$$1 - (\Pr(g = 0)\Pr(y = 0) + \Pr(g = 1)\Pr(y = 1))$$

Recalling that $\Pr(g = 1) = \Pr(y = 1) = p$, we have that loss is

$$2p(1 - p)$$

Note that we saw that the optimal decision rule is to pick the mode and that the loss of that rule is:

$$1 - max(p, (1 - p))$$

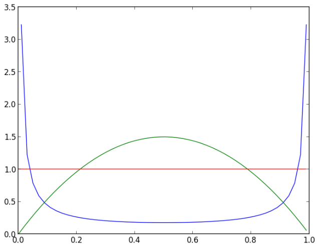
Note that for $p = 0.5$ the losses are the same. But, for other values, say $p = 0.6$, the pigeons' loss is 0.48 and the optimal loss is 0.4. Proving that pigeons are not so good at decision theory.

3 Bayesian belief update [10 Points]

3.1 Beta-Binomial practice

(a) Label which of the lines in the figure below correspond to:

1. Beta(0.1, 0.1)
2. Beta(1,1)
3. Beta(2,2)



Solution:

1. Beta(0.1, 0.1) is blue
2. Beta(1,1) is red
3. Beta(2,2) is green

We are estimating the probability that a coin comes up heads.

(b) What does it mean to have a prior of $\text{Beta}(2, 2)$?

Solution: Before seeing any data, we believe that the distribution for the parameter μ of a binomial random variable, which describes the numbers of heads and tails, is distributed as $\text{Beta}(\mu; 2, 2)$. This is as if we had previously seen 2 heads and 2 tails.

(c) If that's the prior, what is the posterior after seeing 3 heads and 2 tails?

Solution: The posterior is $\text{Beta}(\mu; 5, 4)$

(d) What are the mean and mode of that posterior?

Solution: The mean is 5/9; the mode is 4/7. Note that without a prior, we would have had $\mu_{\mathbf{ml}} = 3/5$ which is a more "extreme" value than both the mean and the mode of the posterior distribution. The impact of the extra "head" observation is moderated by the prior.

(e) What does it mean to have a prior of $\text{Beta}(2, 3)$?

Solution: It's as if we had previously seen 2 heads and 3 tails, starting with a uniform prior on μ.

(f) If that's the prior, what is the posterior after seeing 3 heads and 2 tails?

Solution: The posterior is $\text{Beta}(\mu; 5, 5)$

(g) What are the mean and mode of that posterior?

Solution: The mean is 1/2; the mode is 1/2.

3.2 What's new?

(Bishop 2.7) Consider a bernoulli random variable x with mean μ with prior distribution for μ given by the beta distribution:

$$\text{Beta}(\mu; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \quad (2.13)$$

and suppose we have observed m occurrences of $x = 1$ and l occurrences of $x = 0$. Show that the posterior mean value of μ lies between the prior mean and the maximum likelihood estimate for μ.

To do this, show that the posterior mean can be written as λ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, where $0 \leqslant \lambda \leqslant 1$. This illustrates the concept of the posterior distribution being a compromise between the prior distribution and the maximum likelihood solution.

Solution: We will show that

$$E[\mu|D] = \lambda E[\mu] + (1-\lambda)\mu_{\mathbf{ml}}$$

using a beta distribution for the prior.

So,

$$E[\mu] = \frac{a}{a+b} \quad \text{(using 2.15)}$$

$$E[\mu|D] = p(\mu|m,l;a,b) = \frac{m+a}{m+a+l+b} \quad \text{(using 2.20)}$$

$$\mu_{\mathbf{ml}} = \frac{m}{m+l} \quad \text{(using 2.8)}$$

Therefore,

$$\frac{m+a}{m+a+l+b} = \lambda\frac{a}{a+b} + (1-\lambda)\frac{m}{m+l}$$

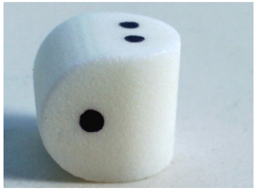
$$\lambda = \left(\frac{m+a}{m+a+l+b} - \frac{m}{m+l}\right)\frac{(a+b)(m+l)}{a(m+l)-m(a+b)}$$

$$\lambda = \frac{(a+b)}{m+a+l+b} = \frac{1}{1+\frac{m+l}{a+b}}$$

Because a, b, m and l are positive, $\lambda \in (0,1)$

4 Which dice factory? [15 points]

You have just purchased a two-sided die, which can come up either 1 or 2:



You want to use your crazy die in some betting games with friends later this evening, but first you want to know the probability that it will roll a 1.

You know it came either from factory 0 or factory 1, but not which.

Factory 0 produces dice that roll a 1 with probability ϕ_0. Factory 1 produces dice that roll a 1 with probability ϕ_1. You believe initially that with probability η_0 that it came from factory 1.

(a) Without seeing any rolls of this die, what would be your predicted probability that it would roll at 1?

Solution: Define θ as a binary random variable which is one if the die came from factory 0 and Y as the random variable associated with a dice roll. Then by conditional probability, we have

$$\begin{aligned}\Pr(Y = 1) &= \Pr(Y = 1|\theta = 0)\Pr(\theta = 0) + \Pr(Y = 1|\theta = 1)\Pr(\theta = 1)\\ &= \phi_0(1 - \eta_0) + \phi_1\eta_0\end{aligned}$$

(b) If we roll the die and observe the outcome, what can we infer about where the coin was manufactured?

Solution: Having observed an outcome y, we can apply Bayes' rule.

$$\begin{aligned}\Pr(\theta = 1 \mid Y = y) &= \frac{\Pr(y \mid \theta = 1)\Pr(\theta = 1)}{\Pr(y)}\\ &= \frac{\phi_1^y(1-\phi_1)^{1-y}\eta_0}{\phi_0^y(1-\phi_0)^{1-y}(1-\eta_0) + \phi_1^y(1-\phi_1)^{1-y}\eta_0}\\ \eta_1 &= g(\eta_0, y)\end{aligned}$$

In the second equality, we used exponentiation as a way to select amongst the two possible choices in general. It doesn't always come out so cleanly.

So, η_1 are the parameters of the posterior.

(c) More concretely, let's assume that:

- $\phi_0 = 1$: dice from factor 0 always roll a 1
- $\phi_1 = 0.5$: dice from factory 1 are fair (roll at 1 with probability 0.5)
- $\eta_0 = 0.7$: we think with probability 0.7 that this die came from factory 1

Now we roll it, and it comes up 1! What is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

Solution:

$$\eta_1 = \frac{0.5 \cdot 0.7}{0.5 \cdot 0.7 + 1 \cdot 0.3} \approx 0.54$$

The predictive distribution over the next roll is

$$\Pr(Y = 1) = \eta_1\phi_1 + (1 - \eta_1)\phi_0 \approx 0.71$$

(d) You roll it again, and it comes up 1 again.

Now, what is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

Solution: The update is the same, but starting from the posterior we had before.

$$\eta_2 = \frac{0.5 \cdot 0.54}{0.5 \cdot 0.54 + 1 \cdot 0.46} \approx 0.37$$

(e) Instead, what if it rolls a 2 on the second roll?

Solution: We know for sure where this coin came from!

$$\eta_2 = \frac{0.5 \cdot 0.54}{0.5 \cdot 0.54 + 0 \cdot 0.46} = 1 \ \ .$$

(f) In the general case (not using the numerical values we have been using) prove that if you have two observations, and you use them to update your prior in two steps (first conditioning on one observation and then conditioning on the second), that no matter which order you do the updates in you will get the same result.

Solution: Let us denote our 2 observations by y_a, y_b. Next, observe that $\Pr(y_a, y_b|\theta = \alpha) = \Pr(y_a|\theta = \alpha)\Pr(y_b|\theta = \alpha)$, this follows from the fact that the rolls are independent given we know factory, and since $\Pr(y_a, y_b) = \sum_\alpha \Pr(y_a|\theta = \alpha)\Pr(y_b|\theta = \alpha)\Pr(\theta = \alpha) = \sum_\alpha \Pr(y_b|\theta = \alpha)\Pr(y_a|\theta = \alpha)\Pr(\theta = \alpha) = \sum_\alpha \Pr(y_b, y_a|\theta = \alpha)\Pr(\theta = \alpha)$, we have that $\Pr(y_a, y_b) = \Pr(y_b, y_a)$. Thus, the probability of observing y_a, y_b does not change with the order in which they appear. By Bayes' Rule, $\Pr(\theta = \alpha|y_a, y_b)$ can be rewritten as,

$$\begin{aligned}
\Pr(\theta = \alpha|y_a, y_b) &= \frac{\Pr(y_a, y_b|\theta = \alpha)\Pr(\theta = \alpha)}{\Pr(y_a, y_b)} \\
&= \frac{\Pr(y_a|\theta = \alpha)\Pr(y_b|\theta = \alpha)\Pr(\theta = \alpha)}{\Pr(y_b, y_a)} \\
&= \frac{\Pr(y_b, y_a|\theta = \alpha)\Pr(\theta = \alpha)}{\Pr(y_b, y_a)} \\
&= \Pr(\theta = \alpha|y_b, y_a)
\end{aligned}$$

5 Emergency Room [15 Points]

You are a young doctor, working off your federal medical school tuition grant in southern North Dakota. It's your fourth day on the job. You are all alone in the emergency room (ER) when Pat comes in complaining of chest pain.

You have to predict whether Pat is having a heart attack (H) or indigestion (I). Your loss function is:

$$L(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{if } g = \text{"H" and } a = \text{"I"} \\ 10 & \text{if } g = \text{"I" and } a = \text{"H"} \end{cases}$$

You have seen three previous patients who exhibited chest pain, none of whom were actually having a heart attack.

(a) You use those three data points to make a point estimate of the probability that Pat is having a heart attack and then use it to make the prediction that minimizes the empirical risk. What do you predict? What is the empirical risk of that prediction?

> Do you think the empirical risk of this predictor is a good measure of how useful it will be?

Solution: First we begin by estimating the maximum likelihood estimate (MLE) of the Binomial distribution.

Recall that the binomial distributed random variable Y with n total draws and a probability p of success has the probability mass function (PMF)

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}.$$

In our case, we have that $y = 3$, $n = 3$, so taking the log of the likelihood we observe that

$$\log(P(Y = y|n, p)) = 3\log(p).$$

which increases monotonically with p, and therefore the MLE $\hat{p} = 1$.

Recall that the empirical risk of making a guess g is

$$\frac{1}{3}\sum_{i=1}^{3} L(\hat{g}, y_i).$$

Now note that the risk of $g = I$ is zero since we make zero mistakes if we had always guessed indigestion, and the risk of $g = H$ is 1, since we would have made an average of one mistake per patient. Therefore the empirical risk minimizing decision is $g = I$ with risk zero.

This is a terrible decision for two reasons: first, it ignores our intuition that heart attacks occur with probability greater than zero, and second, we would make this decision even if mistaking heart attack for indigestion had an arbitrarily large (finite) loss. It should be troubling that this decision completely ignores the loss function.

(b) The next morning, you think more carefully and decide it would be better to forget all your previous experience and simply view each new patient with an open mind. So, you use some ideas from this week's lectures. Let Q be a random variable representing the probability that a random patient walking into your ER will be having a heart attack. You have a uniform prior on Q.

What is the prediction that minimizes risk for a random patient walking into your ER? What is the risk of that prediction?

Solution: Now we have a model where the probability of any patient having indigestion is controlled by a random variable Q instead of a probability p. First we will write down the probability that the next patient Y has indigestion.

$$\begin{aligned} P(a = ''I''|n = 1) &= \int_0^1 P(a = ''I''|n = 1, p = p)P(Q = p)dp \\ &= \int_0^1 pP(Q = p)dp \\ &= E[Q] \end{aligned}$$

For this problem, since Q is uniform, we have that $P(a = ``I''|n = 1) = E[Q] = 0.5$.

Now that we have the probability that the next patient has indigestion $a = ``I''$, we can now calculate the risk as

$$\begin{aligned} R_{g=H} &= E[L(H, a)] = P(a = I) = 0.5 \\ R_{g=I} &= E[L(I, a)] = 10 - 10P(a = I) = 5 \end{aligned}$$

So the optimal decision is now to guess $g = H$ which gives risk 0.5.

(c) Later that afternoon, you figure it would be better to combine approaches. So, what if you started with a uniform prior, but then observed three patients all of whom had indigestion?

What would be your posterior distribution on Q? What prediction should you make? What is the risk (under the posterior distribution) of that prediction?

Solution: Following the same argument as part b, we will first derive the probability that the next patient has indigestion. In this case, this requires us to calculate the posterior. For clarity we write down all three parts of our update:

$$\begin{aligned} \text{Prior:} \quad & Q \sim \text{Unif}(0, 1) = \text{Beta}(1, 1) \\ \text{Likelihood:} \quad & Y \sim \text{Binomial}(3, Q) \\ \text{Posterior:} \quad & \hat{Q} \sim \text{Beta}(Y + 1, 3 - Y + 1) \end{aligned}$$

Since in our case, we hvae already observed that $Y = 3$, we know that the posterior is a $\text{Beta}(4, 1)$. Using identical arguments as part B, we derive the probability that $a = ''I''$ under the posterior distribution. This is called the posterior predictive distribution (since we are predicting the next data using our posterior).

$$
\begin{aligned}
P(a = \text{"I"}|n = 1) &= \int_0^1 P(a = \text{"I"}|n = 1, p = p)P(Q = p)dp \\
&= \int_0^1 pP(Q = p)dp \\
&= E[Q] \\
&= 4/5
\end{aligned}
$$

Finally we obtain the risks as:

$$
\begin{aligned}
R_{g=H} &= E[L(H, a)] = P(a = I) = 4/5 \\
R_{g=I} &= E[L(I, a)] = 10 - 10P(a = I) = 2
\end{aligned}
$$

So the optimal decision is still to guess $g = H$ which gives risk $4/5$.

(d) That evening, really worried that you haven't had enough experience in these matters, and beginning to question your judgment about accepting this job, you decide to call your friend Chris who is working at Mass General. Chris has seen 20 patients with indigestion and 1 with heart attack. You use Chris's experience to construct a prior distribution, and then update it with your own (3 patients with indigestion).

What would be your posterior distribution on Q? What prediction should you make? What is the risk (under the posterior distribution) of that prediction?

Solution: The difference between parts c and d is that the prior is no longer uniform. Using Chris' previous experience, we know that in the past there were 20 patients with indigestion and 1 with heart attack, this corresponds to a prior distribution of $\text{Beta}(20, 1)$

If we believed that the distribution of Q was uniform before calling Chris, the proper prior would be $\text{Beta}(21, 2)$ rather than $\text{Beta}(20, 1)$ since we would be updating a uniform prior with 20 indigestion and 1 heart attack observations. Here we are going to assume that we are truly ignorant of the distribution of Q before calling Chris

$$
\begin{aligned}
\text{Prior:} &\quad Q \sim \text{Beta}(20, 1) \\
\text{Likelihood:} &\quad Y \sim \text{Binomial}(3, Q) \\
\text{Posterior:} &\quad \hat{Q} \sim \text{Beta}(Y + 1, 3 - Y + 1)
\end{aligned}
$$

Therefore using the same arguments as part c, the posterior is $\text{Beta}(23, 1)$ and the risks are

$$
\begin{aligned}
R_{g=H} &= E[L(H, a)] = P(a = I) = 23/24 \\
R_{g=I} &= E[L(I, a)] = 10 - 10P(a = I) = 10/24
\end{aligned}
$$

Therefore we would select $g = I$ giving risk $10/24$

The optimal decision between parts c and d are very different despite the fact that the observed data (3 patients) are identical. In the case that we have little data, the optimal decision is often strongly influenced by choice and construction of the prior

(e) At 2AM, questioning the meaning of life, you are quite sure that you should have become a poet. You are so uncertain of your ability to make predictions that you call your former professor who is the head of the emergency medicine department at Gotham City Hospital. Herr Prof. Dr. Strangelove has seen 2000 patients with indigestion and 20 with heart attack. You use Dr. Strangelove's experience to construct a prior distribution, and then update it with your own (3 patients with indigestion).

What would be your posterior distribution on Q? What prediction should you make? What is the risk of that prediction?

Solution: Using the same argument, the posterior is $\text{Beta}(2003, 20)$ giving risks of

$$\begin{aligned} R_{g=H} &= E[L(H, a)] = P(a = I) = 2003/2024 \approx 0.990 \\ R_{g=I} &= E[L(I, a)] = 10 - 10P(a = I) = 210/2024 \approx 0.104 \end{aligned}$$

Therefore we would select $g = I$ giving risk about 0.104.

Is there a potential problem with using Dr. Strangelove's data to help construct your prior?

6 Abby Normal [15 Points]

Dr. Frahnkensteen is designing an artificial cranium, but she needs to know how big to make it; her design goal is to be a good fit to 80% of brains. So, she wants to get a good estimate of the distribution of the sizes of brains in the local population. Since brains are kind of squishy, we will just consider the total volume of the brain, a one-dimensional quantity.

The Dr. has considerable previous experience with brains and thinks their distribution is well modeled as a Gaussian distribution with with a variance of 75cc. But she's not at all sure about the mean of this current population. She thinks it might be somewhere around 1100cc.

(a) One way to express the Dr.'s uncertainty about the distribution of brain sizes in the local population is to put a Gaussian distribution *on the mean* of the local distribution.

What are the hyper-parameters of this distribution? Pick some to model Dr. F's situation (they're not completely determined by the story).

Solution: Data values are drawn from a Gaussian distribution with known variance, σ_D^2, but unknown mean. Assume a prior distribution on the mean, which is a Gaussian with parameters μ_0, σ_0^2. So:

- $\theta \in \mathbb{R}$

- $y^{(i)} \in \mathbb{R}$
- $y^{(i)} \mid \theta \sim \text{Normal}(\theta, \sigma_D^2)$
- $\theta \sim \text{Normal}(\mu_0, \sigma_0^2)$

(b) Dr. F. sends her assistant Eygor out to get a new brain from the local population. Eygor brings back one that is 1500cc! What should the posterior be?

Start by solving this problem algebraically. Write down the prior and the observation likelihood function symbolically. Then, derive a form for the posterior.

What actual numerical values do you get, given your answer to the previous question, and the observation of 1500cc?

Solution: Assume we make a single observation $y^{(1)}$. What is the posterior?

$$\begin{aligned}
\Pr(\theta \mid y^{(1)}) &\propto \Pr(y^{(1)} \mid \theta; \sigma_D^2) \Pr(\theta; \mu_0, \sigma_0^2) \\
&\propto \exp\left(-\frac{(y^{(1)} - \theta)^2}{2\sigma_D^2}\right) \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right) \\
&\propto \exp\left(-\theta^2 \left(\frac{1}{2\sigma_D^2} + \frac{1}{2\sigma_0^2}\right) + 2\theta \left(\frac{y^{(1)}}{2\sigma_D^2} + \frac{\mu_0}{2\sigma_0^2}\right)\right) \\
&\propto \exp\left(-\frac{(\theta - \mu_1)^2}{2\sigma_1^2}\right)
\end{aligned}$$

where

$$\mu_1 = \frac{\sigma_D^2 \mu_0 + \sigma_0^2 y^{(1)}}{\sigma_D^2 + \sigma_0^2} \quad ,$$

which is a weighted average of the prior mean and the data, and

$$\sigma_1^2 = \frac{\sigma_0^2 \sigma_D^2}{\sigma_0^2 + \sigma_D^2} \quad .$$

The third proportionality constant comes from the fact in this case the random variable is θ, meanwhile we know $y^{(1)}$, and therefore is a constant.
Note that the new variance is less than the prior variance and less than the variance of the observation. So, we can conclude that

$$\theta \mid y^{(1)} \sim \text{Normal}(\mu_1, \sigma_1) \quad .$$

(c) How is the new mean related to the old mean and the observation?

Solution: Rewriting the previous solution in terms of the inverse of the variance (called the precision),

$$\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y^{(1)}}{\sigma_D^2}}{\frac{1}{\sigma_D^2} + \frac{1}{\sigma_0^2}} \enspace .$$

This immediately shows that the posterior mean is a the average of the prior mean weighted by $1/\sigma_0^2$ and the observation weighted by $1/\sigma_D^2$.

(d) What can we say about how the variance behaves when an observation is made?

Solution: Once again re-writing in terms of precisions,

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_D^2}$$

which indiates that the precision is always increasing with more observations, and the variance decreasing.

(e) What is Dr. F's. posterior predictive distribution? First find it symbolically, then numerically.

Solution: Another important question in this case is, what is the *posterior predictive distribution?*? It is

$$\begin{aligned}
\Pr(y^{(n+1)} \mid \mathcal{D}) &= \int_\theta \Pr(y^{(n+1)} \mid \theta) \Pr(\theta \mid \mathcal{D}) d\theta \\
&= \int_\theta \Pr(y^{(n+1)} \mid \theta) \Pr(\theta \mid \mu_n, \sigma_n) d\theta \\
&= \mathcal{N}(y^{(n+1)}; \mu_n, \sigma_n^2 + \sigma_D^2)
\end{aligned}$$

One way to derive this is with a lot of hassling with the integral and the quadratic stuff in the exponent. Another (thanks to a paper by Murphy) is to make the following observations:

- $\theta \mid \mathcal{D} \sim \text{Normal}(\mu_n, \sigma_n^2)$
- $y^{(n+1)} \mid \theta \sim \text{Normal}(\theta, \sigma_D^2)$
- $y^{(n+1)} - \theta \sim \text{Normal}(0, \sigma_D^2)$

First note that the quantity $y^{(n+1)} - \theta$ is conditionally independent of θ (think of Y = Z + W, where Z and W are gaussians). We can see $y^{(n+1)} \mid \mathcal{D}$ as a sum of $(y^{(n+1)} - \theta) \mid \mathcal{D}$ and $\theta \mid \mathcal{D}$. The sum of two Gaussian random variables is also a Gaussian, where the new mean is the sum of the means and the new variance is the sum of the variances. So,

$$y^{(n+1)} \mid \mathcal{D} \sim \text{Normal}(\mu_n, \sigma_D^2 + \sigma_n^2)$$

(f) If Eygor brought back 10 more brains from the local morgue, would Dr. F. be able to update her prior in some way that is more efficient than doing the individual update procedure 10 times?

Solution: As parts c and d show, the posterior updates are sums and weighted means of the precision, so given 10 individuals, we would simply take the weighted mean as:

$$\mu_{11} = \frac{\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{10} \frac{y^{(i)}}{\sigma_D^2}}{\frac{10}{\sigma_D^2} + \frac{1}{\sigma_0^2}} \quad .$$

and

$$\frac{1}{\sigma_{11}^2} = \frac{1}{\sigma_0^2} + \frac{10}{\sigma_D^2}$$

7 Coding Question: Two Gaussians [25 Points]

We saw in lecture that if we know $p(X, Y)$ then we can derive an optimal decision rule, but we were sad to realize that we never really know $p(X, Y)$. One strategy for addressing this problem is to directly estimate $p(X, Y)$ and then use the estimate to derive a decision rule that would be optimal if our estimate were accurate.

In this question we consider a generative model for a dataset comprised of a mixture of two gaussians. The data is generated as follows. Let $C_0 = N(\mu_0, \Sigma_0)$ and $C_1 = N(\mu_1, \Sigma_1)$ be two gaussians where μ_0 and $\mu_1 \in \mathbb{R}^d$ are the means and Σ_0 and $\Sigma_1 \in \mathbb{R}^{d\times d}$ are two covariances. Let $y \in \{0, 1\}$ be a latent variable indicating if x is drawn from C_0 or C_1. The probability density of x is defined as follows

$$P(x) = P(x|y = 1)P(y = 1) + P(x|y = 0)P(y = 0) \tag{1}$$

Our goal is to derive and implement the bayes optimal classifier δ such that given a new point $x' \in \mathbb{R}^d$,

$$\delta(x') = \arg\max_{y\in\{0,1\}} P(x'|y) \tag{2}$$

We have provided two csv files train.csv and test.csv for the completion of this question.

(a) (Empirics) From train.csv, what is your maximum likelihood estimate for $P(y = 0)$ and $P(y = 1)$? What is your estimate for μ_0 and μ_1? What is your estimate for Σ_0 and Σ_1? Do you notice something about Σ_0 and Σ_1? (Hint: Don't overthink)

Solution: $P(y = 0) = 1/3$ and $P(y = 1) = 2/3$. $\mu_0 = (0, 0)$ and $\mu_1 = (5, 0)$. $\Sigma_0 = \Sigma_1 = I$

(b) (Theory) What are $P(y = 1|x)$ and $P(y = 0|x)$ proportional to, as a function of x?

Solution: We apply bayes rule

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x)} \propto P(x|y=1)P(y=1) \tag{3}$$

We drop the $P(x)$ in the denominator as it does not depend on y The density is then proportional to

$$P(y=1|x) \propto \exp((x-\mu_1)^T\Sigma_1(x-\mu_1))P(y=1) \tag{4}$$

and analogously for $y=0$

$$P(y=0|x) \propto \exp((x-\mu_0)^T\Sigma_0(x-\mu_0))P(y=0) \tag{5}$$

(c) (Theory) Derive an equation for the decision boundary for $x \in \mathbb{R}^d$ where

$$\ln(P(y=1|x)) = \ln(P(y=0|x)) \tag{6}$$

Here we compare the log likelihood as it simplifies the derivation. Is this decision boundary (as a function of x) linear, quadratic, etc.? How does the decision boundary simplify when $\Sigma_0 = \Sigma_1$?

Solution: Taking the log of $P(y=1|x)$ and $P(y=0|x)$ and set them equal

$$(x-\mu_1)^T\Sigma_1(x-\mu_1) + \ln(P(y=1)) = (x-\mu_0)^T\Sigma_0(x-\mu_0) + \ln(P(y=0)) \tag{7}$$

This is a quadratic in x. For $\Sigma = \Sigma_1 = \Sigma_2$ we have

$$-2x^T\Sigma\mu_1 + \mu_1\Sigma\mu_1 + \ln(P(y=1)) = -2x^T\Sigma\mu_0 + \mu_0\Sigma\mu_0 + \ln(P(y=0)) \tag{8}$$

which is linear in x.

(d) (Empirics) Using the decision boundary derived in part (c), classify the points in test.csv as $y=0$ or 1. It suffices to write down the form of the decision boundary and associated decision rule.

Solution: Let $x = (x_a, x_b) \in \mathbb{R}^2$ Decision boundary is $y=0$ when

$$\ln(1/3) > -10x_a + 25 + \ln(2/3) \tag{9}$$

rearranging we obtain

$$x_a > 0.1(25 + \ln(2/3) - \ln(1/3)) \tag{10}$$

In particular, the decision boundary is NOT $x_a = 2.5$ because of the prior.

8 Optional

8.1 Throwing rocks: asymmetric loss

You just bought a new trebuchet and you are interested in making predictions about how far it can throw a rock. Your ballistics officer tells you that optimizing the squared error of your predictions is not appropriate for the problem. If your prediction is within some constant c of the true value then they can use your predicted value to aim the trebuchet such that it hits the castle, but if your prediction off by more than c, then using the prediction for aiming will cause the trebuchet to miss. However, the fact is, it's better for your prediction to be too short than too far.

So, we will let

$$L(a, g) = \begin{cases} 0 & \text{if } |a - g| < c \\ 1 & \text{if } a - g > c \\ 2 & \text{if } g - a > c \end{cases}$$

If you know that the range of the ball for these types of trebuchets is distributed as a Gaussian with mean μ and variance σ^2, what prediction minimizes loss L?

(This is a little bit tricky. It's fine to just write out an expression in terms of c, μ, and σ^2.)

Solution: The expected loss is

$$\mathbb{E}[L(a,g)] = \int_{-\infty}^{\infty} L(a,x)p(x;\mu,\sigma^2)dx = 2\phi(a + c;\mu,\sigma^2) + \hat{\phi}(a - c;\mu,\sigma^2) \quad (11)$$

Where $\phi(x;\mu,\sigma^2) = \int_x^\infty yp(y;\mu,\sigma^2)dy$ and $\hat{\phi}(x;\mu,\sigma^2) = \int_{-\infty}^x yp(y;\mu,\sigma^2)dy$. For sake of checking solutions, it's convenient to express everything in terms of the cdf of standard normal

$$2(1 - \text{erf}(\frac{a + c - \mu}{\sigma})) + \text{erf}(\frac{a - c - \mu}{\sigma}) \quad (12)$$

Alternative solution. The best prediction is the one that minimizes the expected loss over the randomness in a:

$$\begin{aligned} \hat{g} &= \arg\min_g \mathbb{E}_a[L(a,g)] \\ &= \arg\min_g \int_a L(a,g)p(a)da \\ &= \arg\min_g \int_{-\infty}^{\infty} L(a,g)\mathcal{N}(a|\mu,\sigma^2)da \\ &= \arg\min_g \int_{-\infty}^{g-c} L(a,g)\mathcal{N}(a|\mu,\sigma^2)da + \int_{g-c}^{g+c} L(a,g)\mathcal{N}(a|\mu,\sigma^2)da + \int_{g+c}^{\infty} L(a,g)\mathcal{N}(a|\mu,\sigma^2)da \end{aligned}$$

Note that we intentionally divided up the integral so that we can easily plug in values for $L(a, g)$. In the first integral, we have that the prediction is too far, and in the last integral, the prediction is too short:

$$= \arg\min_g \int_{-\infty}^{g-c} 2\mathcal{N}(a|\mu, \sigma^2)da + \int_{g+c}^{\infty} \mathcal{N}(a|\mu, \sigma^2)da. \tag{13}$$

We can now solve the optimization problem:

$$\begin{aligned} 0 &= \frac{1}{dg}\left[\int_{-\infty}^{g-c} 2\mathcal{N}(a|\mu, \sigma^2)da + \int_{g+c}^{\infty} \mathcal{N}(a|\mu, \sigma^2)da\right] \\ &= 2\mathcal{N}(g-c|\mu, \sigma^2) - \mathcal{N}(g+c|\mu, \sigma^2) \end{aligned}$$

which we see simplifies nicely by the fundamental theorem of calculus. We now solve for g:

$$2\mathcal{N}(g-c|\mu, \sigma^2) = \mathcal{N}(g+c|\mu, \sigma^2)$$

$$2\exp\left(-\frac{(g-c-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{(g+c-\mu)^2}{2\sigma^2}\right)$$

We can take log of both sides and do some algebra to get the final answer:

$$g = \frac{-\sigma^2 \ln(2)}{2c} + \mu.$$

Intuitively, this makes sense. We slightly under-predict the mean because of our biased loss function.

8.2 Copy that: discrete Bayes update and decision theory

You have just bought a copy machine at a garage sale. You know it is one of two possible models, m_1 or m_2, but the tag has fallen off, so you're not sure which.

You do know that m_1 machines have a 0.1 "error" (bad copy) rate and m_2 machines have a 0.2 error rate.

(a) You use your machine to make 1000 copies, and 140 of them are bad. What is the maximum likelihood estimate of the machine's error rate? Explain why. (Remember that you're sure it's one of those two types of machines).

Solution: We first solve the MLE of the type of the machine, which we denote by $b \in \{1, 2\}$. Using a particular machine, the number of bad copies, denoted by k, is a random variable, as $k \sim \text{Binomial}(n, p_b)$. Thus,

$$\Pr(k \mid b) = \binom{n}{k} p_b^k (1-p_b)^{n-k} \Rightarrow \log\Pr(k \mid b) = \log C + k\log p_b + (n-k)\log(1-p_b).$$

Here, C is the value of n *choose* k. With $n = 1000$, $k = 140$, $p_1 = 0.1$ and $p_2 = 0.2$, we have

$$\log \Pr(k \mid b = 1) = \log C + 140 \log(0.1) + 860 \log(0.9) = \log C - 412.97$$
$$\log \Pr(k \mid b = 2) = \log C + 140 \log(0.2) + 860 \log(0.8) = \log C - 417.22$$

We can see that $\log \Pr(k \mid b = 1) > \log \Pr(k \mid b = 2)$, which implies that the MLE of the type of the machine is $b_{\mathbf{ml}} = 1$. It follows that the machine's error rate is $p_{b_{\mathbf{ml}}} = 0.1$.

(b) Looking more closely, you can see part of the label, and so you think that, just based on the label it has a probability 0.2 of being an m_1 type machine and a probability 0.8 of being an m_2 type machine. If you take that to be your prior, and incorporate the data from part a, what is your posterior distribution on the type of the machine?

Solution: Under the condition that the total number of copies that we made is $n = 1000$, the posterior distribution of the type of the machine, denoted by b, is

$$\Pr(b = 1 \mid k) = \frac{\Pr(k \mid b = 1) \Pr(b = 1)}{\Pr(k \mid b = 1) \Pr(b = 1) + \Pr(k \mid b = 2) \Pr(b = 2)} = \frac{0.2}{0.2 + 0.8 \frac{\Pr(k|b=2)}{\Pr(k|b=1)}}.$$

We note that $\log \Pr(k \mid b = 2) - \log \Pr(k \mid b = 1) = -4.25$. Hence

$$\frac{\Pr(k \mid b = 2)}{\Pr(k \mid b = 1)} = \exp(-4.25) = 0.0142.$$

As a result, we have

$$\Pr(b = 1 \mid k) = 0.946, \quad \text{and} \quad \Pr(b = 2 \mid k) = 0.054.$$

(c) Given that posterior, what is the probability that the next copy will be a failure?

Solution: Given the posterior, the predictive probability of the next copy being bad is

$$\Pr(b = 1 \mid k)p_1 + \Pr(b = 2 \mid k)p_2 = 0.946 \cdot 0.1 + 0.054 \cdot 0.2 = 0.1054.$$

where p_i is the failure probability of machine type m_i.

(d) You intend to sell this machine on the web. Because it's used, you have to sell it with a warranty. You can offer a gold or a silver warranty. If it has a gold warranty and the buyer runs it for 1000 copies and gets more than 150 bad copies, then you are obliged to pay \$1000 in damages; if it has a silver warranty, you have to pay damages if it generates more than 300 bad copies in 1000 copies. Your maximum reasonable asking price for a machine with a gold warranty is \$300; for a machine with a silver warranty, it is \$100. You can assume the machine will sell at these prices. What type of warranty should you offer on this machine?

Solution: Let $k = 140$ denote the number of bad copies that we have observed, and k' denote the number of bad copies the machine will generate when the buyer runs it for 1000 new copies. The probability that $k' > 150$ is

$$\Pr(k' > 150 \mid k) \;=\; \Pr(k' > 150 \mid b = 1) \Pr(b = 1 \mid k) + \Pr(k' > 150 \mid b = 2) \Pr(b = 2 \mid k).$$

When $n = 1000$, the binomial distribution is extremely peaky, with most probability mass falling around np. Hence, $\Pr(k' > 150 \mid b = 1) \simeq 0$, and $\Pr(k' > 150 \mid b = 2) \simeq 1$. Hence $\Pr(k' > 150 \mid k) \simeq \Pr(b = 2 \mid k) = 0.054$.

Similarly, we have

$$\Pr(k' > 300 \mid k) \;=\; \Pr(k' > 300 \mid b = 1) \Pr(b = 1 \mid k) + \Pr(k' > 300 \mid b = 2) \Pr(b = 2 \mid k) \simeq 0.$$

Actually, using either machine, it is very unlikely to generate over 300 bad copies for 1000 runs.

Hence, the expected profit of offering gold warranty is

$$300 - 1000 \cdot \Pr(k' > 150 \mid k) \simeq 300 - 1000 \cdot 0.054 = 246.$$

The expected profit of offering silver warranty is

$$100 - 1000 \cdot \Pr(k' > 300 \mid k) \simeq 100.$$

Therefore, offering gold warranty would generate higher expected profit, which is what we should do.

(e) Under what conditions would it be better to just throw the machine away, rather than try to sell it?

Solution: We should just throw it away when *the expected profit is zero or even negative* for both warranties that we can offer.

For this particular problem, even for the worst case scenario where we are sure with probability 1 that the machine is the worse one (with error rate 0.2), it is still very unlikely that it produces over 300 bad copies for 1000 runs (you can verify this by computing the CDF). In this (worst) case, it is still profitable to sell the machine with silver warranty.

8.3 Dirichlet Priors

Exercise borrowed from Stat180 at UCLA. See Bishop, sections 2.1 and 2.2 for background on Beta and Dirichlet distributions.

The Dirichlet distribution is a multivariate version of the Beta distribution. When we have a coin with two outcomes, we really only need a single parameter θ to model the probability of heads. But now let's consider a "thick" coin that has three possible outcomes: heads, tails, and edge. Now we need two parameters: θ_h is the probability of heads, θ_t is the probability of tails, and then the probability of an edge is $1 - \theta_h - \theta_t$.

The random variables $(V, W) \in [0, 1]$ and such that $V + W \leqslant 1$ have a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \alpha_3$ if their joint density is

$$f(v, w) = v^{\alpha_1 - 1} w^{\alpha_2 - 1} (1 - v - w)^{\alpha_3 - 1} \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \enspace .$$

This is a direct generalization of the Beta distribution. (Note that Γ refers to the Gamma function, which is a generalization of factorial.)

(a) If (θ_h, θ_t) have a Dirichlet distribution as above, what is the marginal distribution of θ_h?

Solution: To find the marginal distribution of θ_h, we integrate the joint distribution over θ_t:

$$\begin{aligned} f(\theta_h) = \int f(\theta_h, \theta_t) \, d\theta_t \quad &\propto \quad \int_0^{1-\theta_h} \theta_h^{\alpha_1 - 1} \theta_t^{\alpha_2 - 1} (1 - \theta_h - \theta_t)^{\alpha_3 - 1} \, d\theta_t \\ &= \quad \theta_h^{\alpha_1 - 1} \int_0^{1-\theta_h} \theta_t^{\alpha_2 - 1} (1 - \theta_h - \theta_t)^{\alpha_3 - 1} \, d\theta_t \end{aligned}$$

The integral looks similar to a beta function integral. Changing variables with $u = \frac{\theta_t}{1-\theta_h}$ and $du = \frac{d\theta_t}{1-\theta_h}$:

$$\begin{aligned} f(\theta_h) &\propto \theta_h^{\alpha_1 - 1} (1 - \theta_h)^{\alpha_2 + \alpha_3 - 1} \int_0^1 \left(\frac{\theta_t}{1 - \theta_h} \right)^{\alpha_2 - 1} \left(\frac{1 - \theta_h - \theta_t}{1 - \theta_h} \right)^{\alpha_3 - 1} \frac{d\theta_t}{1 - \theta_h} \\ &= \theta_h^{\alpha_1 - 1} (1 - \theta_h)^{\alpha_2 + \alpha_3 - 1} \int_0^1 u^{\alpha_2 - 1} (1 - u)^{\alpha_3 - 1} \, du \quad \propto \theta_h^{\alpha_1 - 1} (1 - \theta_h)^{\alpha_2 + \alpha_3 - 1} \end{aligned}$$

The final expression has the same functional form as a beta density. So $\theta_h \sim \text{Beta}(\alpha_1, \alpha_2 + \alpha_3)$.

(b) Suppose you are playing with a thick coin, and get results $x^{(1)} \ldots x^{(n)}$, resulting in H heads and T tails out of n throws. Given θ_h and θ_t the random variables H and T have a multinomial distribution:

$$\Pr(H, T | \theta_h, \theta_t) = \frac{n!}{H!T!(n - H - T)!} \theta_h^H \theta_t^T (1 - \theta_h - \theta_t)^{n - H - T} \enspace .$$

Assume a uniform prior on the space of possible values of θ_h and θ_t (remembering that they are constrained such that $\theta_h \geqslant 0$, $\theta_t \geqslant 0$, and $\theta_h + \theta_t \leqslant 1$). What is the posterior distribution for θ_h and θ_t?

Solution: By Bayes' rule,

$$\begin{aligned}\Pr(\theta_h, \theta_t | H, T) &\propto \Pr(H, T | \theta_h, \theta_t) P(\theta_h, \theta_t) \\ &\propto \frac{n!}{H!T!(n-H-T)!} \theta_h^H \theta_t^T (1-\theta_h-\theta_t)^{n-H-T} \\ &\propto \theta_h^H \theta_t^T (1-\theta_h-\theta_t)^{n-H-T}\end{aligned}$$

where we have absorbed all constants unrelated to θ_h and θ_t (the posterior distribution is a function of only θ_h and θ_t). Note that $P(\theta_t, \theta_h)$ was assumed uniform and so is a constant. The final expression has the same functional form as a Dirichlet density, so $\theta_h, \theta_t | H, T \sim \text{Dirichlet}(H+1, T+1, n-H-T+1)$.

(c) In this same setting, what is the predictive distribution for getting another head? That is, what's $\Pr(x^{(n+1)} = \text{heads} \mid x^{(1)} \ldots x^{(n)})$?

Solution: It is generally easier to work with the parameters θ instead of the data itself. The following decomposition holds by the law of total probability and the chain rule of probability:

$$\begin{aligned}\Pr(x^{n+1} | x^{(1)}, \ldots, x^{(n)}) &= \int \Pr(x^{n+1}, \theta_h | x^{(1)}, \ldots, x^{(n)}) d\theta_h \\ &= \int \Pr(x^{n+1} | \theta_h, x^{(1)}, \ldots, x^{(n)}) \Pr(\theta_h | x^{(1)}, \ldots, x^{(n)}) d\theta_h\end{aligned}$$

Note that the two probabilities within the integral are easier to evaluate. Since the coin flips are independent (given that we know θ_h), $\Pr(x^{n+1} = \text{heads} | \theta_h, x^{(1)}, \ldots, x^{(n)}) = \theta_h$. As for the second density, we know from question b that the posterior distribution for θ_h, θ_t is Dirichlet, hence from question a the marginal posterior distribution $\theta_h | h, t \sim \text{Beta}(h+1, n-h+2)$. The integral becomes:

$$\Pr(x^{(n+1)} = \text{heads} | x^{(1)}, \ldots, x^{(n)}) = \int \theta_h \text{Beta}(h+1, n-h+2) d\theta_h$$

$$= E_{\text{Beta}(h+1, n-h+2)}[\theta_h] = \frac{h+1}{n+3}$$

(d) Now assume a Dirichlet prior for θ_h and θ_t with parameters $\alpha_1, \alpha_2, \alpha_3$. What is the posterior in this case?

Solution: We repeat the derivation of question b for $P(\theta_h, \theta_t) \propto \theta_h^{\alpha_1 - 1} \theta_t^{\alpha_2 - 1} (1 - \theta_h - \theta_t)^{\alpha_3 - 1}$:

$$\begin{aligned}\Pr(\theta_h, \theta_t | h, t) &\propto \left[\theta_h^h \theta_t^t (1-\theta_h-\theta_t)^{n-h-t}\right] \left[\theta_h^{\alpha_1-1} \theta_t^{\alpha_2-1} (1-\theta_h-\theta_t)^{\alpha_3-1}\right] \\ &\propto \theta_h^{\alpha_1+h-1} \theta_t^{\alpha_2+t-1} (1-\theta_h-\theta_t)^{\alpha_3+(n-h-t)-1}\end{aligned}$$

Again, this has the form of a Dirichlet density, so $\theta_h, \theta_t | h, t \sim \text{Dirichlet}(\alpha_1 + h, \alpha_2 + t, \alpha_3 + (n - h - t))$.

(e) In this same case, what is the predictive distribution?

Solution: A similar derivation as in question c gives

$$\Pr(x^{n+1} = \text{heads}|x^{(1)}, \dots, x^{(n)}) = \frac{\alpha_1 + h}{\alpha_1 + \alpha_2 + \alpha_3 + n}$$

.

(f) If you assume a squared-error loss on the predicted parameter, that is,

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \quad ,$$

what is the Bayes-optimal estimate of θ_h and θ_t?

Solution: In this problem we consider θ_h and θ_t separately (they are similar). The Bayes-optimal estimate $\hat{\theta}_h$ of θ_h is the one that minimizes the expected loss over the posterior distribution of θ_h:

$$\hat{\theta}_h = \arg\min_z \int (\theta_h - z)^2 \Pr(\theta_h | h, t) d\theta_h$$

Differentiating this with respect to z and setting the result to 0, we find that $\hat{\theta}_h$ is the posterior expectation of θ_h, which is $\frac{\alpha_1+h}{\alpha_1+\alpha_2+\alpha_3+n}$ (also found in question c and question e). Similarly, $\hat{\theta}_t = \frac{\alpha_2+t}{\alpha_1+\alpha_2+\alpha_3+n}$.

(g) As $n \to \infty$, how do optimal estimates relate to the maximum likelihood estimates and to the prior?

Solution: As $n \to \infty$, the prior contributes less; $\hat{\theta}_h \to \frac{h}{n}$ and $\hat{\theta}_t \to \frac{t}{n}$, i.e., the estimates approach the MLE.

8.4 More fun with Dirichlet

Given a parameterized family of probability models $\Pr(x \mid \theta)$ and a data set $D = (x^{(1)}, \dots, x^{(n)})$ comprised of independent samples $x^{(i)} \approx \Pr(x \mid \theta)$, we fit the model to the data so as to maximize the likelihood (or log-likelihood) of all samples. This gives the maximum-likelihood (ML) estimate of the parameters:

$$\hat{\theta}_{ML} = \arg\max_\theta \log \Pr(D \mid \theta)$$

This approach does not express any prior bias as to which values of θ we should prefer when data is limited.

In the sequel, we consider a regularized approach to parameter estimation. Here, we specify a prior model $\Pr(\theta)$ over the set of allowed parameter settings Θ. Given a prior model, we may then employ Bayes' rule to compute the posterior probability of θ given the observations:

$$\Pr(\theta \mid D) = \frac{\Pr(D \mid \theta)\Pr(\theta)}{\Pr(D)}$$

where

$$\Pr(D) = \int_{\Theta} \Pr(D \mid \theta)\Pr(\theta)d\theta$$

Then, we fit the model to the data by maximizing the (log-) probability of θ conditioned on the data,

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} \log \Pr(\theta \mid D) \\ &= \arg\max_{\theta}\{\log \Pr(D \mid \theta) + \log \Pr(\theta) - \log \Pr(D)\} \\ &= \arg\max_{\theta}\{\log \Pr(D \mid \theta) + \log \Pr(\theta)\}\end{aligned}$$

Note that we have dropped the $-\log \Pr(D)$ term as this does not depend upon θ and does not affect the parameter estimate. Hence, we do not need to explicitly evaluate the integral in the denominator. This may be viewed as a penalized log-likelihood criterion, i.e. we maximize $J(\theta) = \log \Pr(D;\theta) + f(\theta)$ subject to the regularization penalty $f(\theta) = \log \Pr(\theta)$. The parameter estimate $\hat{\theta}_{MAP}$ is known as the maximum a posteriori (MAP) estimate.

In this problem you will construct MAP estimates for the probabilities of a (potentially biased) M -sided die, i.e. $x^{(i)} \in \{1, \ldots, M\}$. We consider the fully-parameterized representation $\Pr(x = k) = \theta_k$, where $0 \leqslant \theta_k \leqslant 1$ for $k = 1, \ldots, M$ and $\sum_k \theta_k = 1$. This simple model has many relevant applications.

Consider a document classification task, where we need class-conditional distributions over words in the documents. Suppose we only consider words $1, \ldots, M$ (for relatively large M). Each word in the document is assumed to have been drawn at random from the distribution $\Pr(x = k \mid y; \theta) = \theta_{k|y}$, where $\sum_k \theta_{k|y} = 1$ for each class y. Thus the selection of words according to the distribution $\theta_{k|y}$ can be interpreted as a (biased) M-sided die.

Now, the probability of generating all words $x^{(1)}, \ldots, x^{(n)}$ in a document of length n would be

$$\Pr(D \mid y; \theta) = \prod_{i=1}^{n} \Pr(x^{(i)} \mid y; \theta) = \prod_{i=1}^{n} \theta_{x^{(i)}|y}$$

assuming the document belongs to class y. Note that this model cares about how many times each word occurs in the document. It is a valid probability model over the set of words in the document.

Since we typically have very few documents per class, it is important to regularize the parameters, i.e., provide a meaningful prior answer to the class conditional distributions.

Let's start by briefly revisiting ML estimation of the (biased) M-sided die. Similarly to calculations you have already performed, the ML estimate of the parameter θ from n samples is given by the empirical distribution:

$$\hat{\theta}_x = \frac{n(x)}{n}$$

where $n(x)$ is the number of times value x occurred in n samples. The count $n(x)$ is also a *sufficient statistic* for θ_x as it is all we need to know from the available n samples in order to estimate θ_x.

Next, we consider MAP estimation. To do so, we must introduce a prior distribution over the θ's. A natural choice for this problem is the Dirichlet distribution

$$\Pr(\theta;\beta) = \frac{1}{Z(\beta)} \prod_{k=1}^{M} \theta_k^{\beta_k}$$

with non-negative hyperparameters $\beta = (\beta_k > 0, k = 1, \dots, M)$ and where $Z(\beta)$ is just the normalization constant (which you saw earlier and which you do not need to evaluate in this problem).

(a) First, consider this prior model (ignoring the data for the moment). What value of θ is most likely under this prior model? That is, compute

$$\hat{\theta}(\beta) = \arg\max_{\theta} \log \Pr(\theta;\beta)$$

This is the *a priori* estimate of θ before observing any data.

Solution: We wish to maximize

$$l(\theta) = \log P(\theta;\beta) = -\log Z(\beta) + \beta_x \log \theta_x$$

w.r.t parameters θ subject to $\sum_x \theta_x = 1$. Use Lagrange multipliers.

$$L(\theta,\mu) = l(\boldsymbol{\theta}) + \mu\left(1 - \sum_x \theta_x\right) = -\log Z(\beta) + \mu + \sum_x (\beta_x \log \theta_x - \mu\theta_x)$$

Minimizing θ for fixed μ

$$\frac{\partial L}{\partial \theta_x} = \frac{\beta_x}{\theta_x} - \mu = 0$$

This gives

$$\hat{\theta}_x = \frac{\beta_x}{\mu}$$

Using the same approach we used in the Exercises from Week 1 (problem 2), *i.e.*, $\sum_x \hat{\theta}_x = 1$, we have $\mu = \sum_x \beta_x$ so that

$$\hat{\theta}_x = \frac{\beta_x}{\sum_k \beta_k}$$

(b) Next, given the data D, compute the MAP estimate of θ as a function of the hyperparameters β and the data D (use the sufficient statistics $n(x)$):

$$\hat{\theta}_{MAP}(D;\beta) = \arg\max_{\theta} \log \Pr(\theta \mid D;\beta)$$

Note that you do not need to calculate $Z(\beta)$ in order to perform this optimization; you can optimize the penalized log-likelihood $J(\theta) = \log \Pr(D \mid \theta) + f(\theta;\beta)$ with a simple penalty function $f(\theta;\beta)$, as discussed above. Thus we do not have to evaluate the full posterior distribution $\Pr(\theta \mid D;\beta)$ in order to perform the regularization.

Solution:

We wish to maximize the penalized log-likelihood

$$l(\theta) = \log P(D|\theta) + \log P(\theta;\beta) = -\log Z(\beta) + \sum_{i=1}^{n} \log \theta_{x^{(i)}} + \sum_{x} \{\beta_x \log \theta_x\}$$

We can rewrite the quantity $\sum_{i=1}^{n} \log \theta_{x^{(i)}} = \sum_x n_x \log \theta_x$ and we have,

$$l(\theta) = \log P(D|\theta) + \log P(\theta;\beta) = -\log Z(\beta) + \sum_{i=1}^{n} \log \theta_{x^{(i)}} + \sum_{x} \{(n_x + \beta_x) \log \theta_x\}$$

w.r.t. θ subject to the constraint $\sum_x \theta_x = 1$. We minimize the Lagrangian

$$L(\theta,\mu) = l(\theta) + \mu\left(1 - \sum_x \theta_x\right) = \mu - \log Z(\beta) + \sum_x \{(n(x) + \beta_x) \log \theta_x - \mu\theta_x\}$$

Having

$$\frac{\partial L}{\partial \theta_x} = \frac{n(x) + \beta_x}{\theta_x} - \mu = 0$$

gives

$$\theta_x = \frac{n(x) + \beta_x}{\mu}$$

As before, we get $\mu = \sum_x (n(x) + \beta_x)$ so that the MAP estimate is

$$\hat{\theta}_x = \frac{n(x) + \beta_x}{\sum_x (n(x) + \beta_x)}$$

(c) Show that your MAP estimate may be expressed as a convex combination of the a priori estimate $\hat{\theta}(\beta)$ and the ML estimate $\hat{\theta}_{ML}(D)$. The means that we may write

$$\hat{\theta}_{MAP}(D;\beta) = (1-\lambda)\hat{\theta}_{ML}(D) + \lambda\hat{\theta}(\beta)$$

for some $\lambda \in [0,1]$. Note that the same convex combination holds for each component θ_x. Determine λ as a function of the number of samples n and the hyperparameters β.

Solution: Since $\sum_x n(x) = n$ and let $N = \sum_x \beta_x$, the MAP estimate becomes

$$\hat{\theta}_x = \frac{n(x) + \beta_x}{n + N} = \frac{1}{n+N}\left\{n\hat{\theta}_x^{ML} + N\hat{\theta}_x^{Prior}\right\} = \frac{n}{n+N}\hat{\theta}_x^{ML} + \frac{N}{n+N}\hat{\theta}_x^{Prior}$$

Therefore

$$\lambda = \frac{N}{n+N} = \frac{\sum_x \beta_x}{\sum_x (n(x) + \beta_x)}$$

where $x = 1, \ldots, M$.

As this shows, one way of thinking of a prior distribution is that it is a proxy for any data we have observed in the past but no longer have available. The normalized parameters $\hat{\beta}_i = \beta_i / N$, where $N = \sum_i \beta_i$, express our prior estimate of the parameters θ while the normalization parameter N expresses how strongly we believe in that prior estimate.