6.867: Exercises (Week 2)

Sept 8, 2016

Contents

Solution: Don't look at the solutions until you have tried your absolute hardest to solve the problems.

1 One parameter, two estimators

In this problem, we're going to explore the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. Consider a simple model for its distribution (in a later problem we will consider a slightly different model):

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter b, such that $0 < b$.

We are interested in estimates of the mean of the distribution.

(a) What's the mean of the Model 1 distribution?

Solution: The model density is $\frac{1}{b}$ (over $[0, b]$) giving a mean $\frac{b}{2}$.

Let's start by considering the situation in which the data were, in fact, drawn from an instance of the model under consideration: a uniform distribution on $[0, b]$ (for model 1),

In model 1, the ML estimator for b is $b_{\mathbf{ml}} = \max_i x^{(i)}$. The likelihood of the data is:

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

We can see that if $b_{\mathbf{ml}} < x^{(i)}$, 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 $b_{\mathbf{ml}} = \max_i x^{(i)}$.

To understand the properties of this estimator we have to start by deriving their PDFs. The minimum and maximum of a data set are also known as their first and nth *order statistics*, and sometimes written $x^{[1]}$ and $x^{[n]}$ (we're using square brackets to distinguish these from our notation for samples in a data set).

In model 1, we just need to consider the distribution of $b_{\mathbf{ml}}$. Generally speaking, the pdf of the maximum of a set of data drawn from pdf f, with cdf F, is:

$$f_{b_{\mathbf{ml}}}(x) = nF(x)^{n-1} f(x) \tag{1}$$

The idea is that, if x is the maximum, then $n-1$ of the other data values will have to be less than x, and the probability of that is $F(x)^{n-1}$, and then one value will have to equal x, the probability of which is $f(x)$. We multiply by n because there are n different ways to choose the data value that could be the maximum.

(b) What is the maximum likelihood estimate of the mean, $\mu_{\mathbf{ml}}$, of the distribution?

Solution: Given the MLE $b_{\mathbf{ml}}$ of b, which is $x^{[n]}$ the maximum of the data set, the MLE of the mean is $\frac{b_{\mathbf{ml}}}{2} = \frac{x^{[n]}}{2}$ (from our expression of the mean in part a).

(c) What is $f_{b_{\mathbf{ml}}}$ for this particular case where the data are drawn uniformly from 0 to b?

Solution: $f(x) = \frac{1}{b}$, $F(x) = \frac{x}{b}$, hence $f_{b_{\mathbf{ml}}}(x) = n\frac{x^{n-1}}{b^n}$ over $[0, b]$, and is zero otherwise.

(d) Write an expression for the expected value of $\mu_{\mathbf{ml}}$, as an integral,

Solution: The pdf of the max of n data points was given in Equation 1 above. Given that the max value is x, the mean is $\frac{x}{2}$ from Q1. Hence:

$$E[\mu_{\mathbf{ml}}] = \int_0^b \frac{x}{2} f_{b_{\mathbf{ml}}}(x)\, dx = \int_0^b \frac{x}{2} n \frac{x^{n-1}}{b^n}\, dx = \frac{b}{2}\frac{n}{n+1}$$

In fact, there's a nice closed form expression, which you can use in the following questions:

$$E[\mu_{\mathbf{ml}}] = \frac{b}{2}\frac{n}{(n+1)} \quad .$$

(e) What is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased? (Reminder: $\text{bias}^2(\theta_{\mathbf{ml}}) = (E_D[\theta_{\mathbf{ml}}] - \theta)^2$.)

Solution: Using the given equations,

$$\text{bias}^2(\mu_{\mathbf{ml}}) = (E[\mu_{\mathbf{ml}}] - \mu)^2 = \left(\frac{b}{2}\frac{n}{n+1} - \frac{b}{2}\right)^2 = \frac{b^2}{4(n+1)^2}$$

Because bias^2 is not zero, it is biased. However, since $\text{bias}^2 \to 0$ as $n \to \infty$, it is asymptotically unbiased.

(f) Write an expression for the variance of $\mu_{\mathbf{ml}}$, as an integral.

Solution:

$$\begin{aligned} V[\mu_{\mathbf{ml}}] &= E\left[\mu_{\mathbf{ml}}^2\right] - [E[\mu_{\mathbf{ml}}]]^2 \\ &= \int_0^b \left(\frac{x}{2}\right)^2 f_{b_{\mathbf{ml}}}(x)\, dx - [E[\mu_{\mathbf{ml}}]]^2 \\ &= \int_0^b \frac{x^2}{4} n \frac{x^{n-1}}{b^n}\, dx - \left[\frac{b}{2}\frac{n}{n+1}\right]^2 \\ &= \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} \end{aligned}$$

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} \quad .$$

(g) What is the mean squared error of $\mu_{\mathbf{ml}}$? (Reminder: $\text{MSE}(\theta_{\mathbf{ml}}) = \text{bias}^2(\theta_{\mathbf{ml}}) + \text{var}(\theta_{\mathbf{ml}})$.)

Solution:

$$\text{MSE}(\mu_{\mathbf{ml}}) = \text{bias}^2(\mu_{\mathbf{ml}}) + \text{var}(\mu_{\mathbf{ml}}) = \frac{b^2}{4(n+1)^2} + \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} = \frac{b^2}{2(n+1)(n+2)}$$

(h) So far, we have been considering the error of the *estimator*, comparing the estimated value of the mean with its actual value. We will often want to use the estimator to make predictions, and so we might be interested in the expected error of a prediction.

Assume the loss function for your predictions is $L(g, a) = (g - a)^2$. Given an estimate $\hat{\mu}$ of the mean of the distribution, what value should you predict?

What is the expected loss (risk) of this prediction? Take into account both the error due to inaccuracies in estimating the mean as well as the error due to noise in the generation of the actual value.

Solution: Let μ be the mean of the actual distribution, and $\mu_{\mathbf{ml}}$ be the maximum likelihood estimator of mean. Let a be the actual value of next sample, and g be the predicted value. Since the loss function is symmetric, the predicted value should be $g = \mu_{\mathbf{ml}}$. The loss of such prediction:

$$\begin{aligned}
& E_{\mu_{\mathbf{ml}}}\left[E_a[(g-a)^2]\right] && (2)\\
=& E_{\mu_{\mathbf{ml}}}\left[\int_0^{2\mu}(g-a)^2 P(a|\mu)\,da\right] && (3)\\
=& E_{\mu_{\mathbf{ml}}}\left[\int_0^{2\mu}(\mu_{\mathbf{ml}}-a)^2\frac{1}{2\mu}da\right] && (4)\\
=& E_{\mu_{\mathbf{ml}}}\left[\frac{(2\mu-\mu_{\mathbf{ml}})^3+{\mu_{\mathbf{ml}}}^3}{6\mu}\right] && (5)\\
=& \int_0^{\mu} P(\mu_{\mathbf{ml}}|\mu)\frac{(2\mu-\mu_{\mathbf{ml}})^3+{\mu_{\mathbf{ml}}}^3}{6\mu}d\mu_{\mathbf{ml}} && (6)\\
=& \frac{n}{6\mu^n}\int_0^{\mu}(8\mu^2{\mu_{\mathbf{ml}}}^{n-1}-12\mu{\mu_{\mathbf{ml}}}^n+6{\mu_{\mathbf{ml}}}^{n+1})d\mu_{\mathbf{ml}} && (7)\\
=& \frac{n}{6\mu^n}\left(\frac{8\mu^{n+2}}{n}-\frac{12\mu^{n+2}}{n+1}+\frac{6\mu^{n+2}}{n+2}\right)=\frac{n^2+3n+8}{3(n+1)(n+2)}\mu^2 && (8)
\end{aligned}$$

One thing to notice is that, when $n \to \infty$, the risk converges $\frac{1}{3}\mu^2$, which is the variance of uniform distribution.

We might consider something other than the MLE for Model 1 (labeled o for other). Consider the estimator

$$\mu_o = \frac{x^{[n]}(n+1)}{2n} \quad .$$

where $x^{[n]}$ is the maximum of the data set.

(i) Write an expression for the expected value of this version of μ_o as an integral. Then solve the integral.

Solution:

$$E[\mu_o] = \int_0^b \frac{x(n+1)}{2n} f_{b_o}\, dx = \int_0^b \frac{x(n+1)}{2n} n \frac{x^{n-1}}{b^n}\, dx = \frac{b}{2}$$

Where f_{b_o} also assumes the data is drawn uniformly from 0 to b.

(j) What is the squared bias of this estimator for μ_o? Is this estimator unbiased? Is it asymptotically unbiased?

Solution:

$$\text{bias}^2(\mu_o) = (E[\mu_o] - \mu)^2 = \left(\frac{b}{2} - \frac{b}{2}\right)^2 = 0$$

The estimator is unbiased (and asymptotically unbiased).

(k) Write an expression for the variance of μ_o as an integral.

Solution:

$$\begin{aligned} V[\mu_o] &= \int_0^b \left(\frac{x(n+1)}{2n}\right)^2 f_{b_o}\, dx - [E[\mu_o]]^2 \\ &= \int_0^b \frac{x^2(n+1)^2}{4n^2} n \frac{x^{n-1}}{b^n}\, dx - \frac{b^2}{4} \\ &= \frac{b^2}{4n(n+2)} \end{aligned}$$

The closed form for the variance is

$$V[\mu_o] = \frac{b^2}{4n(n+2)} \quad .$$

(l) What is the mean squared error of this version of μ_o?

Solution: Since the bias is zero, the MSE is the same as the variance $V[\mu_o]$.

(m) What are the relative advantages of the estimator from the previous question and this one?

> **Solution:** The unbiased estimator is strictly better. It always has smaller MSE, even if the variance is higher.

2 One problem, two models

In this problem, we're going to continue exploring the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. We will consider two different models for its distribution:

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter b, such that $0 < b$.
- **Model 2:** The data are drawn from a uniform distribution on the interval $[a, b]$. This model has two positive real parameters, a and b, such that $0 < a < b$.

We are interested in comparing estimates of the mean of the distribution, derived from each of these two models.

2.1 Using Model 2

(a) What's the mean of the Model 2 distribution?

> **Solution:** The model density is $\frac{1}{b-a}$ (over $[a, b]$) giving a mean $\frac{a+b}{2}$.

(b) Let's consider the situation in which the data were, in fact, drawn from an instance of the model under consideration: either a uniform distribution on $[0, b]$ (for model 1) or a uniform distribution on $[a, b]$ (for model 2).

In model 1, the ML estimator for b is $b_{\mathbf{ml}} = \max_i x^{(i)}$. The likelihood of the data is:

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

We can see that if $b_{\mathbf{ml}} < x^{(i)}$, 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 $b_{\mathbf{ml}} \geqslant x^{(i)}\ \forall i$, which means $b_{\mathbf{ml}} = \max_i x^{(i)}$.

By a similar argument in model 2, the ML estimator for b remains the same and the ML estimator for a is $a_{\mathbf{ml}} = \min_i x^{(i)}$. To understand the properties of these estimators we have to start by deriving their PDFs. The minimum and maximum of a data set are also known as their first and nth *order statistics*, and sometimes written $x^{(1)}$ and $x^{(n)}$.

We started our analysis of Model 1 in question 1. Now, let's do the same thing, but for the MLE for model 2. We have to start by thinking about the joint distribution of MLE's $a_{\mathbf{ml}}$ and

$b_{\mathbf{ml}}$. Generally speaking, the joint pdf of the minimum and the maximum of a set of data drawn from pdf f, with cdf F, is

$$f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y) = n(n-1)(F(y)-F(x))^{n-2}f(x)f(y) \quad .$$

Explain in words why this makes sense.

Solution: The argument for $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)$ is similar to the one for $f_{b_{\mathbf{ml}}}(x)$. However, we have to choose a minimum value x in addition to the maximum value y and ensure all other values fall between x and y. First, we factor in the probability (density) of x and y, giving the final $f(x)f(y)$ terms. The other $(n-2)$ data points must all be between x and y, which is true with probability $(F(y)-F(x))^{(n-2)}$. Finally, there are $n(n-1)$ different ways of choosing the maximum and minimum points. (Note that the ordering of these two points matters, so the multiplicative factor is *not* $\binom{n}{2} = \frac{n(n-1)}{2}$).

What is $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}$ in the particular case where the data are drawn uniformly from a to b?

Solution: $f(x) = \frac{1}{b-a}$, $F(x) = \frac{x-a}{b-a}$, hence $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y) = n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}$ for $a \leqslant x \leqslant y \leqslant b$, and is zero otherwise.

Write an expression for the expected value of $\mu_{\mathbf{ml}}$ in terms of an integral.

Here's what it should integrate to:

$$E[\mu_{\mathbf{ml}}] = \frac{a+b}{2} \quad .$$

Solution: Given that x and y are the min and max values for Model 2, the MLE is now $\frac{x+y}{2}$. Hence:

$$E[\mu_{\mathbf{ml}}] = \int\int \frac{x+y}{2} f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy = \int_a^b \int_a^y \frac{x+y}{2} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy = \frac{a+b}{2}$$

(c) What is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased?

Solution: $\text{bias}^2(\mu_{\mathbf{ml}}) = (E[\mu_{\mathbf{ml}}] - \mu)^2 = \left(\frac{a+b}{2} - \frac{a+b}{2}\right)^2 = 0$. The estimator is unbiased (and asymptotically unbiased).

(d) Write an expression for the variance of $\mu_{\mathbf{ml}}$ in terms of an integral.

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{(b-a)^2}{2(n+1)(n+2)} \quad .$$

Solution:

$$V[\mu_{ml}] = \int\int \left(\frac{x+y}{2}\right)^2 f_{a_{ml},b_{ml}}(x,y)\,dxdy - [E[\mu_{ml}]]^2$$
$$= \int_a^b \int_a^y \frac{(x+y)^2}{4} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{2(n+1)(n+2)}$$

(e) What is the mean squared error of μ_{ml}?

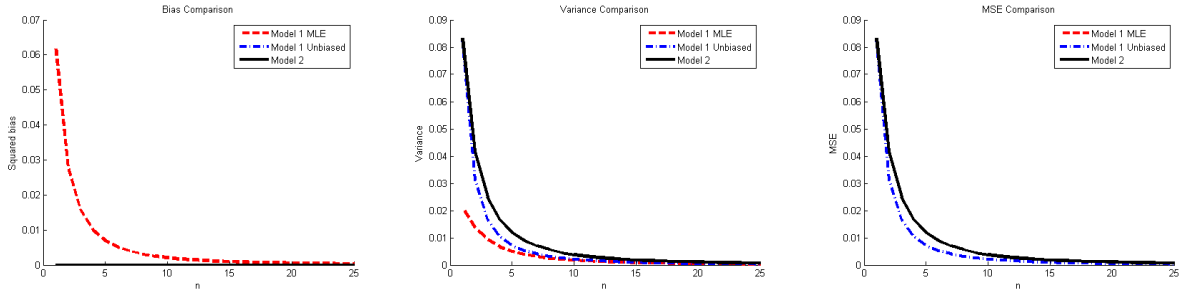
Solution: Since the bias is zero, the MSE is the same as the variance $V[\mu_{ml}]$.

2.2 Comparing Models

What if we have data that is actually drawn from the interval $[0,1]$? Both models seem like reasonable choices.

(a) Show plots that compare the bias, variance, and MSE of each of the estimators we've considered on that data, as a function of n. (Use the formulas above; don't do it by actually generating data). Write a paragraph in English explaining your results. What estimator would you use?

Solution: The plots in Figure 1 compare the bias, variance, and MSE of the models. Note that both Model 1 unbiased (blue) and Model 2 (black) have zero bias in Figure 1(a). Also, for $a = 0$, the MSE for Model 1 MLE (red) and Model 2 are the same, so they overlap in Figure 1(c) (red under black). We already know that the Model 1 unbiased estimator (blue) has lower error than the Model 1 MLE (red). Since the MSE for Model 2 and Model 1 MLE are the same, we conclude that the Model 1 unbiased estimator is superior for data from $[0,1]$ due to its lower variance.



(a) Squared bias (black overlaps blue) (b) Variance (c) MSE (black overlaps red)

Figure 1: Red = model 1 MLE, blue = model 1 unbiased, black = model 2.

(b) Now, what if we have data that is actually drawn from the interval $[.1,1]$? It seems like model 2 is the only reasonable choice. But is it?

We already know the bias, variance, and MSE for model 2 in this case. But what about the MLE and unbiased estimators for model 1? Let's characterize the general behavior when we use the estimator $\mu_{\mathbf{ml}} = x^{(n)}(n+1)/(2n)$ on data drawn from an interval $[a, b]$.

Write an expression for the expected value of $\mu_{\mathbf{ml}}$ in terms of an integral.

Solution:

$$E[\mu_{\mathbf{ml}}] = \int\int \frac{y(n+1)}{2n} f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy = \int_a^b \int_a^y \frac{y(n+1)}{2n} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy = \frac{a+bn}{2n}$$

(c) The closed form expression is

$$E[\mu_{\mathbf{ml}}] = \frac{a+bn}{2n} \quad .$$

Explain in English why this answer makes sense.

Solution: For small n (and in particular for $n = 1$), since the maximum value in fact cannot be less than a, a high value of a means that initial maximum values will be higher, and hence the estimated mean is higher. Ultimately, the estimator only depends on the maximum of the data b, and as we saw earlier the expected value is $\frac{b}{2}$. The expression above tends to this as $n \to \infty$, since with many data points, it is likely that their maximum is close to b.

(d) What is the squared bias of this $\mu_{\mathbf{ml}}$? Explain in English why your answer makes sense. Consider how it behaves as a increases, and how it behaves as n increases.

Solution: $\text{bias}^2(\mu_{\mathbf{ml}}) = (E[\mu_{\mathbf{ml}}] - \mu)^2 = \left(\frac{a+bn}{2n} - \frac{a+b}{2}\right)^2 = \frac{a^2(n-1)^2}{4n^2}$. We already know it's unbiased if $a = 0$; as a increases, this is an increasingly bad (inaccurate) model. Furthermore, for fixed a, the bias increases as a function of n, because the expected answer gets closer to $\frac{b}{2}$ (and farther from the true $\frac{a+b}{2}$).

(e) Write an expression for the variance of this $\mu_{\mathbf{ml}}$ in terms of an integral.

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{(b-a)^2}{4n(n+2)} \quad .$$

To save you some tedious algebra, we'll tell you that the mean squared error of this $\mu_{\mathbf{ml}}$ is (apologies for the ugliness; let us know if you find a beautiful rewrite)

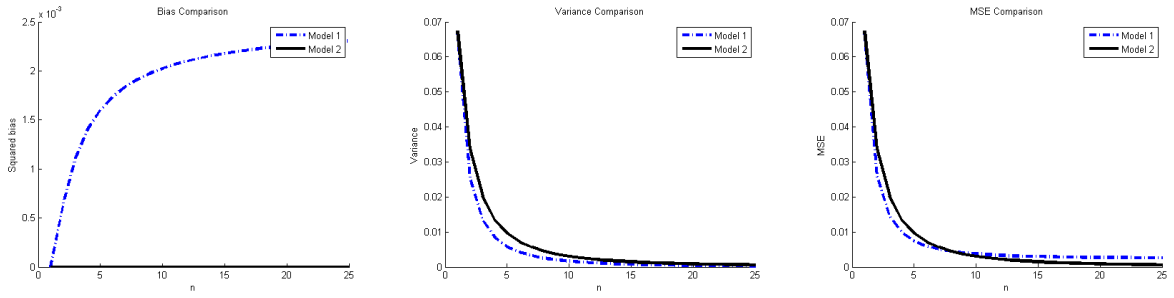
$$\frac{b^2 n - 2abn + a^2(2 - 2n + n^3)}{4n^2(n+2)} \quad .$$

Solution:

$$V[\mu_{\mathbf{ml}}] = \int\int \left(\frac{y(n+1)}{2n}\right)^2 f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy - [E[\mu_{\mathbf{ml}}]]^2$$
$$= \int_a^b \int_a^y \frac{y^2(n+1)^2}{4n^2} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy - \frac{(a+bn)^2}{4n^2} = \frac{(b-a)^2}{4n(n+2)}$$

(f) Show plots that compare the bias, variance, and MSE of this estimator with the regular model 2 estimator on data drawn from $[0.1, 1]$, as a function of n. Are there circumstances in which it would be better to use this estimator? If so, what are they and why? If not, why not?

Solution: The MSE plots in Figure 2 cross over at around 8. That is, for $n < 8$, using the model 1 estimator is better, and beyond that the model 2 estimator should be used. Although model 1 has lower variance than model 2, the bias in using model 1 takes over for larger n.

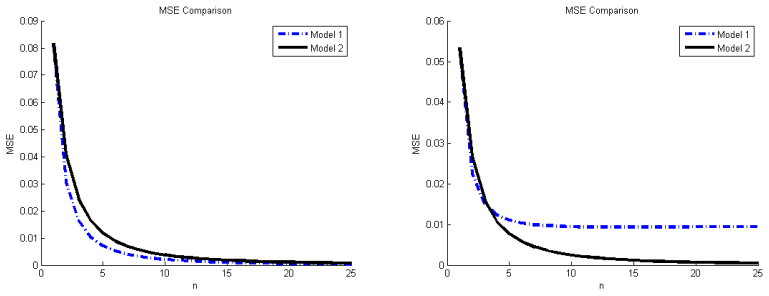


(a) Squared bias (b) Variance (c) MSE

Figure 2: MSE Plots: Blue = model 1, black = model 2. Data from [.1,1].

(g) Show plots of MSE of both estimators, as a function of n on data drawn from $[.01, 1]$ and on data drawn from $[.2, 1]$. How do things change? Explain why this makes sense.

Solution: See Figure 3. For data from $[.01, 1]$, model 1 is a very good approximation. Although the model 1 estimator is still biased, because a is very small, the effect of the bias is much smaller, and model 1 is superior for a larger range of n due to its lower variance. In contrast, for data from $[.2, 1]$, model 1 is less accurate compared to its application in Figure 2. The model 1 estimator is more biased and is inferior for $n > 3$.

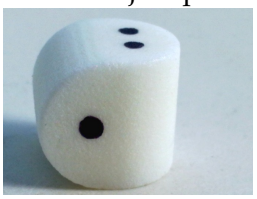


(a) [.01,1] (b) [.2,1]

Figure 3: Blue = model 1, black = model 2.

3 Which dice factory?

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 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:

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

(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!

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

(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.

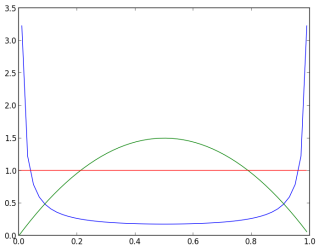
Solution: Let us denote our 2 observations by y_a, y_b. Next, observe that $\Pr(y_a, y_b|\theta = \alpha) = \Pr(y_a|\theta = \alpha)\Pr(y_b|\theta = \alpha)$, this follows from the fact that the rolls are independent given we know factory, and since $\Pr(y_a, y_b) = \sum_\alpha \Pr(y_a|\theta = \alpha)\Pr(y_b|\theta = \alpha)\Pr(\theta = \alpha) = \sum_\alpha \Pr(y_b|\theta = \alpha)\Pr(y_a|\theta = \alpha)\Pr(\theta = \alpha) = \sum_\alpha \Pr(y_b, y_a|\theta = \alpha)\Pr(\theta = \alpha)$, 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,

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

4 Beta-Binomial

(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 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 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 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 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 Beta$(\mu; 5, 5)$

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

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

5 Emergency Room, Reconsidered

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.

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

For this problem, since Q is uniform, we have that $P(a = \text{"I"}|n = 1) = E[Q] = 0.5$.

Now that we have the probability that the next patient has indigestion $a = \text{"I"}$, 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 $a = \text{"I"}$ 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

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

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

Dr. Frahnkensteen is designing an artificial cranium, but he needs to know how big to make it; his design goal is to be a good fit to 80% of brains. So, he 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 he's not at all sure about the mean of this current population. He thinks it might be somewhere around 1100cc.

(a) One way to express the Dr.'s uncertainty about the distribution of brain sizes in his 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 his 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

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

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

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

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

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

(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),

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

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 his 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:

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

and

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

7 Copy that

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 Bayes (Bishop 2.7)

Consider a binomial random variable x given by:

$$\binom{N}{m} \mu^m (1-\mu)^{N-m} \quad (2.9)$$

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, $\lambda \in (0,1)$

9 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

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

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:

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

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:

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

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

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

.

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

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

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.

10 Parameter estimation

Given a parameterized family of probability models $\Pr(x \mid \theta)$ and a data set $D = (x^{(1)}, \ldots, x^{(n)})$ 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 $\beta = (\beta_k > 0, k = 1,\ldots,M)$ 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.

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

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

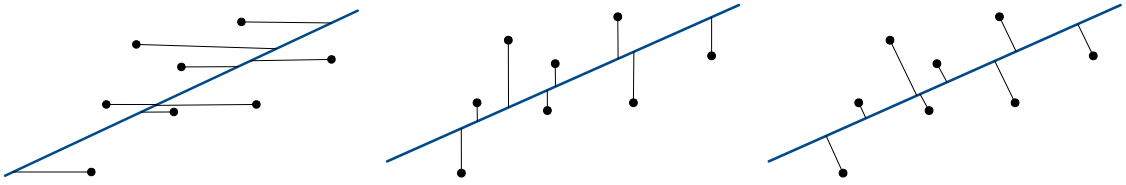
$$\lambda = \frac{N}{n+N} = \frac{\sum_x \beta_x}{\sum_x (n(x) + \beta_x)}$$

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

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.

11 Residue removal

Which of the following images shows the error that is minimized in ordinary least-squares regression? Write down the formula and explain how it's related to the small lines in the picture.



Solution: The second one shows least-squares regression; we are minimizing error in the y direction, conditioned on given values in the x direction.

12 Weighted least squares regression

You are trying to build a predictor with data that you gathered on two different days with two different instruments. We know that data set 1, consisting of n pairs, $(x^{(i)}, y^{(i)})$ has a conditional Gaussian distribution

$$y \sim \text{Normal}(\mathbf{x} \cdot \theta, \sigma_1^2) \ ,$$

and data set 2, consisting of m pairs $(u^{(i)}, v^{(i)})$ has a conditional Gaussian distribution that differs only in the variance:

$$v \sim \text{Normal}(\mathbf{u} \cdot \theta, \sigma_2^2) \ ,$$

The parameter vector θ and all of the $x^{(i)}$ and $u^{(i)}$ are vectors in $\mathbb{R}^d$, and the $y^{(i)}$ and $v^{(i)}$ are in $\mathbb{R}$.

(a) Derive the maximum-likelihood estimator for $\theta \in \mathbb{R}^d$. You can assume that there is no special θ_0. **We strongly recommend that you do this in matrix-vector form.**

Solution: For dataset 1

$$\log L_1(X; \theta) = \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (x^{(i)} \cdot \theta - y^{(i)})^2 + C_1 = \frac{1}{2\sigma_1^2} \|X\theta - Y\|^2 + C_1$$

For dataset 2

$$\log L_1(U; \theta) = \frac{1}{2\sigma_2^2} \sum_{i=1}^{m} (u^{(i)} \cdot \theta - v^{(i)})^2 + C_2 = \frac{1}{2\sigma_2^2} \|U\theta - V\|^2 + C_2$$

where $X = [x^{(1)}, \cdots, x^{(n)}]^T, Y = [y^{(1)}, \cdots, y^{(n)}]^T, V = [v^{(1)}, \cdots, v^{(m)}]^T$, $U = [u^{(1)}, \cdots, u^{(m)}]^T$ and C_1, C_2 are constants from the Gaussian PDF.

The joint objective of MLE is

$$\begin{aligned} J &= \log L_1 + \log L_2 \\ &= \frac{1}{2\sigma_1^2}(X\theta - Y)^T(X\theta - Y) + \frac{1}{2\sigma_2^2}(U\theta - V)^T(U\theta - V) + C_1 + C_2 \\ &= \frac{1}{2}\theta^T \left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]\theta - \theta^T \left[\frac{1}{\sigma_1^2}X^TY + \frac{1}{\sigma_2^2}U^TV\right] + C \end{aligned}$$

To maximize J with respect to θ, take its derivative with respect to θ, set to 0 and solve. Recall that A^TA is symmetric for any matrix A and that $\frac{\partial}{\partial x}x^TAx$ for symmetric A is $2Ax$.

$$\frac{\partial J}{\partial \theta} = \left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]\theta - \left[\frac{1}{\sigma_1^2}X^TY + \frac{1}{\sigma_2^2}U^TV\right] = 0$$

The solution is

$$\hat{\theta} = \left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]^{-1}\left[\frac{1}{\sigma_1^2}X^TY + \frac{1}{\sigma_2^2}U^TV\right]$$

(b) Argue that it makes sense for extreme relative values of σ_1 and σ_2.

Solution: Since σ_1 and σ_2 can be viewed as weighting coefficients, if

$$\begin{cases} \sigma_1 = \sigma_2 & \longrightarrow \text{ unweighted (or equally weighted) least square regression} \\ \sigma_1 \gg \sigma_2 & \longrightarrow \text{ least square regression on } (u^{(i)}, v^{(i)}) \\ \sigma_1 \ll \sigma_2 & \longrightarrow \text{ least square regression on } (x^{(i)}, y^{(i)}) \end{cases}$$

13 Ridge Regression

(a) (Bishop 3.4)
Consider a linear model of the form

$$y(x, \mathbf{w}) = w_o + \sum_{i=1}^{D} w_i x_i \qquad (3.105)$$

together with a sum-of-squares error function of the form

$$Err_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x^{(n)}, \mathbf{w}) - t_n\}^2 \qquad (3.106)$$

where t_n is the true value for $x^{(n)}$. Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $E[\epsilon_i] = 0$ and $E[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$ (where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise) show that minimizing Err_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Solution:

Let

$$\begin{aligned} \tilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i(x_i^{(n)} + \epsilon_i) \\ &= y_n + \sum_{i=1}^{D} w_i\epsilon_i \end{aligned}$$

where $y_n = y(x^{(n)}, \mathbf{w})$ (the prediction from the current model for the n^{th} data point) and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Note that we have used (3.105). Using (3.106), we define

$$\begin{aligned}
\tilde{Err} &= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n - t_n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2) \\
&= \frac{1}{2}\sum_{n=1}^{N}\left(y_n^2 - 2y_n\sum_{i=1}^{D} w_i\epsilon_i + \left(\sum_{i=1}^{D} w_i\epsilon_i\right)^2 - 2t_n y_n - 2t_n\sum_{i=1}^{D} w_i\epsilon_i + t_n^2\right)
\end{aligned}$$

If we now take the expectation of $\tilde{Err}$ under the distribution of ϵ_i, we see that the second and fifth term disappear, since $E[\epsilon_i] = 0$ (it's a zero mean normal), while for the third term we get:

$$E\left[\left(\sum_{i=1}^{D} w_i\epsilon_i\right)^2\right] = \sum_{i=1}^{D} w_i^2\sigma^2$$

since the ϵ_i are all independent with variance σ^2. From this and (3.106) we get:

$$E\left[\tilde{Err}\right] = Err_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2\sigma^2$$

where the regularization coefficient $\lambda = N\sigma^2$. Compare to Bishop 3.27.

(b) (HTF[1] Ex. 3.12 in on-line version, with some notation changed)

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. First we *center* the data, computing a new matrix C, where $c_j^{(i)} = (x_j^{(i)} - \bar{x}_j)$; that is, that we subtract the average value of each feature j, $\bar{x}_j$, from each of the values for feature j in the data.

We augment the centered matrix C with d additional rows $\sqrt{\lambda}I$, and augment $\mathbf{y}$ with d zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data examples that satisfy them.

Solution: The centering step is necessary so that the bias parameter w_0 is not regularized (there is generally no reason to believe the data has mean zero, and it can be removed by such a centering step anyway). If we center both X and $\mathbf{y}$ by subtracting their averages

[1] Hastie, Tibshirani and Friedman, The Elements of Statistical Learning (http://statweb.stanford.edu/ tibs/ElemStatLearn/)

resulting in C and $\mathbf{z}$, and augment them as suggested, ordinary least squares (without bias) minimizes the following objective with respect to $\mathbf{w}$:

$$\begin{aligned}\sum_{i=1}^{n+d}(z^{(i)}-\mathbf{w}\cdot\mathbf{c}^{(i)})^2 &= \sum_{i=1}^{n}((y^{(i)}-\bar{y})-\mathbf{w}\cdot(\mathbf{x}^{(i)}-\bar{\mathbf{x}}))^2+\sum_{j=1}^{d}(0-\sqrt{\lambda}w_j)^2\\ &= \sum_{i=1}^{n}((y^{(i)}-\mathbf{w}\cdot\mathbf{x}^{(i)})-(\bar{y}-\mathbf{w}\cdot\bar{\mathbf{x}}))^2+\lambda\sum_{j=1}^{d}w_j^2\end{aligned}$$

The final expression is similar to the ridge regression objective. In fact, if we recall that in ordinary least squares $w_0 = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-\mathbf{w}\cdot\mathbf{x}^{(i)}) = \bar{y}-\mathbf{w}\cdot\bar{\mathbf{x}}$, we see that the bias term essentially falls out from the centered matrices. If we replace $(\bar{y}-\mathbf{w}\cdot\bar{\mathbf{x}})$ within the squared summand with w_0, then the objective from this approach is equivalent to that of ridge regression, and the resulting $\mathbf{w}$ will be the same.

(c) (based on HTF Ex. 3.6 in on-line version, with some notation changed)
Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\tau I)$, and Gaussian sampling model $\mathbf{y}\sim\mathcal{N}(X\mathbf{w},\sigma^2 I)$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2. Assume that the data are "centered" as described in the previous problem, so that we don't need a bias term.

Solution: By Bayes' rule, $p(\mathbf{w}|\mathbf{y})\propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, hence in the log domain:

$$\begin{aligned}\log p(\mathbf{w}|\mathbf{y}) &= \text{const}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\mathbf{w}\cdot\mathbf{x}^{(i)})^2-\frac{1}{2\tau}\sum_{j=1}^{d}w_j^2\\ &= \text{const}-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y^{(i)}-\mathbf{w}\cdot\mathbf{x}^{(i)})^2+\frac{\sigma^2}{\tau}\sum_{j=1}^{d}w_j^2\right]\end{aligned}$$

where we have accumulated constant terms that do not depend on $\mathbf{w}$. The mode (and mean) of the posterior Gaussian distribution is given by the parameters that maximize the log probability above, which is equivalent to minimizing the expression in the square brackets on the RHS. The latter expression is equivalent to the ridge regression objective for $\lambda=\frac{\sigma^2}{\tau}$, so the estimates are the same.