Selected Exercises of Machine Learning: A Probabilistic Perspective

Mike Craig

Last Updated October 8, 2017

1 Introduction

This is a collection of certain exercise solutions I did as I was reading through this book. As I am not a mathematician, the ”proofs” listed here are rough and probably not rigorous. They only serve as a means to explain the concepts in a way that aids my own understanding.

2 Probability

Exercises

Exercise 2.1. **My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities, two boys or two girls, have probabilities 1/4 and 1/4.**

a. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?

b. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

a. Let G represent one girl, and B represent one boy. Since the neighbor has two children, we can state the entire sample space:

$$S = \{BB, BG, GB, GG\}$$

When the neighbor answers the question, this changes our beliefs about the other child. Using Bayes' theorem:

$$\begin{aligned} P(G=1|B \geq 1) &= \frac{P(B \geq 1|G=1)P(G=1)}{P(B \geq 1)} && (1) \\ &= \frac{2/2 \times 1/2}{3/4} && (2) \\ &= \frac{2}{3} && (3) \end{aligned}$$

b. If we instead happen to see one of his children, this is a different way of looking at the problem. In this situation, learning the gender of one child tells us nothing about the gender of the other child. Therefore, the gender of the second child is a coin flip, 1/2.

Exercise 2.2. **Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.**

a. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he is guilty". This is known as the prosecutor's fallacy. What is wrong with this argument?

b. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that

the defendant is guilty, and thus has no relevance". This is known as the defender's fallacy. What is wrong with this argument?

a. The defendant sharing the blood type does not mean that the defendant himself has a 99% probability of being guilty, just that he shares the same blood type as the guilty party, just like he shares the same blood type with 1% of the population. In a large enough city, there would be a large number of people fitting this description in a small geographical radius.

b. This statement assumes that the defendent is just as guilty (or just as non-guilty) as anyone else in that group of 8000 people. If there truly is no other evidence to tie this defendent to this crime, then that may be so, but if there were any other evidence (drives a similar car as the criminal, lives in the same area, or frequents the same locations), the probability that the defendant is guilty could be much higher.

Exercise 2.3. **Show that the variance of a sum is $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$, where $Cov[X,Y]$ is the covariance between X and Y.**

$$Var[X] + Var[Y] + 2Cov[X,Y] = E[(X-\mu_x)^2] + E[(Y-\mu_y)^2] + 2E[(X-\mu_x)(Y-\mu_y)] \tag{4}$$

$$= E[X^2 - 2X\mu_x + \mu_x^2] + E[Y^2 - 2Y\mu_y + \mu_y^2] + E[2XY - 2X\mu_y - 2Y\mu_x + 2\mu_x\mu_y] \tag{5}$$

$$= E[X^2 - 2X\mu_x + \mu_x^2 + Y^2 - 2Y\mu_y + \mu_y^2 + 2XY - 2X\mu_y - 2Y\mu_x + 2\mu_x\mu_y] \tag{6}$$

$$= E[X^2 + 2XY - 2X(\mu_x + \mu_y) + Y^2 - 2Y(\mu_x + \mu_y) + 2\mu_x\mu_y] \tag{7}$$

Note that $E(X+Y) = E(X) + E(Y) = \mu_x + \mu_y = \mu_{xy}$. Given this,

$$E[X^2 + 2XY - 2X(\mu_x + \mu_y) + Y^2 - 2Y(\mu_x + \mu_y) + 2\mu_x\mu_y] \tag{8}$$

$$= E[X^2 + 2XY - 2X\mu_{xy} + Y^2 - 2Y\mu_{xy} + 2\mu_x\mu_y] \tag{9}$$

$$= E[(X + Y - \mu_{xy})^2] \tag{10}$$

$$= Var[X+Y] \tag{11}$$

Exercise 2.4. **After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)**

Since the test is 99% accurate, we know that $P(Y|D) = P(N|\ D) = 0.99$, where Y means a positive test result, N a negative test result, D means you have the disease, and D means you do not have the disease. We also know the prior probability of having the disease: $P(D) = 0.0001$. Using Bayes' rule:

$$\begin{aligned}
P(D|Y) &= \frac{P(Y|D)P(D)}{P(Y)} \qquad (12)\\
&= \frac{0.99 \times 0.0001}{P(Y|D)P(D) + P(Y|\ D)P(\ D)} \qquad (13)\\
&= \frac{0.000099}{(0.99 \times 0.0001) + (0.01 \times 0.9999)} \qquad (14)\\
&= \frac{0.000099}{0.000099 + 0.009999} \qquad (15)\\
&= 0.0098 \qquad (16)
\end{aligned}$$

So, there's about a 1% chance that you have the disease even though you tested positive for it.

Exercise 2.5. **Solve the Monty Hall problem.**

The key here is that the host will never open a door that has the car in it. So from that sense, the contestant does not disturb the original distribution, but it does give you additional information.

Let's say you originally pick the door with the car. Under these circumstances, which door the host will open is uniform (1/2). In this situation, it is worse for you to switch, and you'll only find yourself in this situation if you pick the door correctly the first time (1/3 chance).

Let's say you did not originally pick the door with the car. Under these circumstances, the door that the host will open is completely deterministic. Two doors will remain, one with the car, and the host will never choose that one. In this situation, it is better for you to switch, because you'll be gauranteed a car.

So, when you guess correctly the first time and switch, you're guaranteed to lose the car. If you guess incorrectly the first time and switch, you're guaranteed to win the car. The probability of guessing correctly the first time is 1/3, so switching will give you a winning probability of 2/3.

Exercise 2.6. **a. Let $H \in \{1, ..., K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector**

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), ..., P(H = K|e_1, e_2))$$

Which of the following sets of numbers are sufficient for the calculation?

i. $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$
ii. $P(e_1, e_2), P(H), P(e_1, e_2|H)$
iii. $P(e_1|H), P(e_2|H), P(H)$

b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficient now?

a. Let's use Bayes' Theorem to decompose this a bit:

$$\begin{aligned} P(H|e_1, e_2) &= \frac{P(e_1, e_2, H)}{P(e_1, e_2)} && (17) \\ &= \frac{P(e_1, e_2|H)P(H)}{P(e_1, e_2)} && (18) \\ &= \frac{P(e_1|e_2, H)P(e_2)P(H)}{P(e_1, e_2)} && (19) \end{aligned}$$

From this, we can see that **ii** is sufficient to solve this.

b. If we assume that they are now conditionally independent, then this allows **i** to be sufficient as well. Note that **iii** is still not sufficient, but if we know that E_1 and E_2 were unconditionally independent, this would be sufficient as well.

Exercise 2.7. **Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counterexample.**

Exercise 2.8. **In the text we said $X \perp Y|Z$ iff $p(x, y|z) = p(x|z)p(y|z)$ for all x, y, z such that $p(z) > 0$. Now prove the following alternative definition: $X \perp Y|Z$ iff there exist functions g and h such that $p(x, y|z) = g(x, z)h(y, z)$ for all x, y, z such that $p(z) > 0$.**

For this to be true, $g(x, z)h(y, z) = p(x|z)p(y|z)$ So by computing the marginal probabilities, we can see if they are equivalent.

$$\begin{aligned} p(x|z) &= \sum_y p(x, y|z) && (20) \\ &= \sum_y g(x, z)h(y, z) && (21) \\ &= g(x, z) \sum_y h(y, z) && (22) \end{aligned}$$

and therefore

$$\sum_y h(y, z) = \frac{p(x|z)}{g(x, z)}$$

Similarly we can find that

$$\sum_x g(x, z) = \frac{p(y|z)}{h(y, z)}$$

and we can note that

$$\sum_x \sum_y p(x,y|z) = \sum_x g(x,z) \sum_y h(y,z) = 1$$

so therefore

$$\begin{aligned}
\sum_x \sum_y p(x,y|z) &= \sum_x g(x,z) \sum_y h(y,z) && (23)\\
&= \frac{p(y|z)}{h(y,z)} \frac{p(x|z)}{g(x,z)} && (24)\\
&= 1 && (25)
\end{aligned}$$

which leads to

$$\begin{aligned}
1 &= \frac{p(y|z)}{h(y,z)} \frac{p(x|z)}{g(x,z)} && (26)\\
\frac{g(x,z)}{p(x|z)} &= \frac{p(y|z)}{h(y,z)} && (27)\\
p(x|z)p(y|z) &= g(x,z)h(y,z) && (28)
\end{aligned}$$

Exercise 2.9.
a. True or false? $(X \perp W|Z,Y) \wedge (X \perp Y|Z) \Rightarrow (X \perp Y,W|Z)$
b. True or false? $(X \perp Y|Z) \wedge (X \perp Y|W) \Rightarrow (X \perp Y|Z,W)$
a. Blowing out the component parts:

$$\begin{aligned}
(X \perp W|Z,Y) &\Leftrightarrow p(X,W|Z,Y) = p(X|Z,Y)p(W|Z,Y) && (29)\\
(X \perp Y|Z) &\Leftrightarrow p(X,Y|Z) = p(X|Z)p(Y|Z) && (30)\\
(X \perp Y,W|Z) &\Leftrightarrow p(X,Y,W|Z) = p(X|Z)p(Y|Z)p(W|Z) && (31)
\end{aligned}$$

so we can see if we can recreate the righthand side using what we know:
P(A,B) = P(A—B)P(B) P(A—B) = P(A,B)/P(B)
P(A,B—C,D) = P(A,B,D—C)/P(D)

$$\begin{aligned}
p(X,Y,W|Z) &= p(X,W|Z,Y)p(Y) && (32)\\
&= p(X|Z,Y)p(W|Z,Y)p(Y) && (33)\\
&= \frac{p(X,Y|Z)}{p(Y)} p(W|Z,Y)p(Y) && (34)\\
&= p(X|Z)p(Y|Z)p(W|Z) && (35)
\end{aligned}$$

b. Blowing out the component parts:

$$
\begin{align}
(X \perp Y|Z) &\Leftrightarrow p(X,Y|Z) = p(X|Z)p(Y|Z) \tag{36}\\
(X \perp Y|W) &\Leftrightarrow p(X,Y|W) = p(X|W)p(Y|W) \tag{37}\\
(X \perp Y|Z,W) &\Leftrightarrow p(X,Y|Z,W) = p(X|Z,W)p(Y|Z,W) \tag{38}
\end{align}
$$

Note that W is in the conditioning set for the equation on the righthand side. There is no way to remove this from the conditioning set while also removing it from the lefthand side. Therefore this is false.

Exercise 2.10. **Given the Gamma density, show that the inverse Gamma is a Gamma with a change of variables to $Y = 1/X$.**

The Gamma is given by:

$$Ga(x|a,b) = \frac{b^a}{\tau(a)} x^{a-1} e^{-xb}$$

and the Inverse Gamma is given by:

$$IG(x|a,b) = \frac{b^a}{\tau(a)} x^{-(a+1)} e^{-b/x}$$

from these, it is easy to show that:

$$
\begin{align}
Ga(\frac{1}{x}|a,b) &= \frac{b^a}{\tau(a)} (\frac{1}{x})^{a-1} e^{-b/x} \tag{39}\\
&= \frac{b^a}{\tau(a)} x^{-1\times(a-1)} e^{-b/x} \tag{40}\\
&= \frac{b^a}{\tau(a)} x^{-(a+1)} e^{-b/x} \tag{41}\\
&= IG(x|a,b) \tag{42}
\end{align}
$$

Exercise 2.11. **Show that the normalization constant for the Gaussian distribution is equal to $Z = \sigma\sqrt{2\pi}$.**

This is essentially just computing the integral

$$
\begin{align}
Z^2 &= \int_0^{2\pi} \int_0^{\infty} r\, exp(-\frac{r^2}{2\sigma^2}) dr d\theta \tag{43}\\
&= 2\pi \int_0^{\infty} r\, exp(-\frac{r^2}{2\sigma^2}) dr \tag{44}\\
&= 2\pi\sigma^2 e^{-\frac{r^2}{2\sigma^2}} |_0^{\infty} \tag{45}\\
&= 2\pi\sigma^2 \tag{46}
\end{align}
$$

Therefore $Z = \sigma\sqrt{2\pi}$.

Exercise 2.12. **Show that**

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

To show this, we must use the definitions of KL divergence, entropy, and conditional entropy. I will show the definitions here for clarity.

$$\begin{aligned}
H(X) &= -\sum_{k=1}^{K} p(X=k) log_2 p(X=k) && (47)\\
KL(X||Y) &= -H(X) + H(X,Y) && (48)\\
H(Y|X) &= \sum_x p(x) H(Y|X=x) && (49)\\
&= H(X,Y) - H(X) && (50)\\
I(X,Y) &= KL(p(X,Y)||p(X)p(Y)) && (51)
\end{aligned}$$

With these definitions, we can show that

$$\begin{aligned}
I(X,Y) &= KL(p(X,Y)||p(X)p(Y)) && (52)\\
&= -H(p(X,Y)) + H(p(X,Y), p(X), pY()) && (53)\\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y) log\, p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y) log\, p(x)p(y) && (54)\\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)(log\, p(y|x) + log\, p(x)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)(log\, p(x) + log\, p(y)) && (55)\\
&= -H(Y|X) + \sum_{x\in X}\sum_{y\in Y} p(x,y) log\, p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)(log\, p(x) + log\, p(y)) && (56)\\
&= -H(Y|X) + \sum_{x\in X} p(x) log\, p(x) - \sum_{x\in X} p(x) log\, p(x) + \sum_{y\in Y} p(y) log\, p(y) && (57)\\
&= -H(Y|X) + \sum_{y\in Y} p(y) log\, p(y) && (58)\\
&= -H(Y|X) + H(Y) && (59)\\
&= H(Y) - H(Y|X) && (60)
\end{aligned}$$

Exercise 2.13. **Evaluate $I(X_1, X_2)$ where X has a bivariate normal distribution. Evaluate $I(X_1, X_2)$ at $\rho = -1$, $\rho = 0$, $\rho = 1$.**

The entropy for both multidimensional and single dimensional Gaussians is defined by:

$$h(\mathbf{X}) = \frac{1}{2} log_2[(2\pi e)^d \, det \, \Sigma] \tag{61}$$
$$h(X) = \frac{1}{2} log_2[2\pi e \sigma^2] \tag{62}$$

Using these, we can compute:

$$\begin{aligned}
I(X_1, X_2) &= H(X_1) - H(X_1|X_2) && (63)\\
&= H(X_1) - H(X_2, X_1) + H(X_2) && (64)\\
&= log_2[2\pi e \sigma^2] - H(X_2, X_1) && (65)\\
&= log_2[2\pi e] + log_2[\sigma^2] - H(X_2, X_1) && (66)\\
&= C + log_2[\sigma^2] - H(X_2, X_1) && (67)\\
&= C + 2log_2[\sigma] - \frac{1}{2} log_2[(2\pi e)^2 \, det \, \Sigma] && (68)\\
&= C + 2log_2[\sigma] - \frac{1}{2}[2C + log_2[det \, \Sigma]] && (69)\\
&= C + 2log_2[\sigma] - C - \frac{1}{2} log_2[\sigma^4(1-\rho^2)] && (70)\\
&= 2log_2[\sigma] - \frac{1}{2} log_2[\sigma^4(1-\rho^2)] && (71)
\end{aligned}$$

where $C = log_2[2\pi e]$. Now we can plug in various values for ρ. When $\rho = -1$,

$$\begin{aligned}
I(X_1, X_2) &= 2log_2[\sigma] - log_2[\sigma^4] && (72)\\
&= 2log_2[\sigma] - 4log_2[\sigma^4] && (73)\\
&= -2log_2[\sigma] && (74)
\end{aligned}$$

When $\rho = 0$,

$$\begin{aligned}
I(X_1, X_2) &= 2log_2[\sigma] - \frac{1}{2} log_2[\sigma^4] && (75)\\
&= 2log_2[\sigma] - 2log_2[\sigma] && (76)\\
&= 0 && (77)
\end{aligned}$$

When $\rho = 1$,

$$I(X_1, X_2) = 2log_2[\sigma]$$

Exercise 2.14. **Let X and Y be discrete random variables which are identically distributed (so $H(X) = H(Y)$) but not necessarily independent. Define**

$$r = 1 - \frac{H(Y|X)}{H(X)}$$

a. Show that $r = \frac{I(X,Y)}{H(X)}$

$$\frac{I(X,Y)}{H(X)} = \frac{H(Y) - H(Y|X)}{H(X)} \quad (78)$$

$$= \frac{H(X) - H(Y|X)}{H(X)} \quad (79)$$

$$= 1 - \frac{H(Y|X)}{H(X)} \quad (80)$$

b. Show that $0 \leq r \leq 1$

$$r = 1 - \frac{H(Y|X)}{H(X)} \quad (81)$$

$$= 1 - \frac{H(X,Y) - H(X)}{H(X)} \quad (82)$$

$$= 1 - H(X,Y) = 1 - H(X) \quad (83)$$

c. When is $r = 0$**?**

$r = 0$ when the entropy of X (or equivalently Y) $= 1$.

d. When is $r = 1$**?**

$r = 1$ when the entropy of X (or equivalently Y) $= 0$. This happens when X and Y is completely deterministic.

Exercise 2.15. **Let** $p(x)$ **be the empirical distribution and** $q(x|\theta)$ **be a model. Show that** $argmin_q\, KL(p||q)$ **is obtained by** $q(x) = q(x|\hat{\theta})$**, where** θ **is the MLE.**

Since

$$KL(p||q) = \sum_k p_k log\, p_k - \sum_k p_k log\, q_k$$

minimizing this is equivalent to maximizing

$$max_q \; - KL(p||q) = - \sum_k p_k log\, p_k + \sum_k p_k log\, q_k \qquad = \sum_k p_k log\, q_k \quad (84)$$

This is the maximum likelihood equation.

Exercise 2.16. **Derive the mean, mode, variance of** θ $Beta(a, b)$

The pdf of the beta distribution is given by

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

Note that $B(\alpha, \beta)$ is a normalization constant, and in order to derive the moments of the distribution we will be using MLE, so for our purposes, we can ignore this constant.

The mode is defined as the peak of the distribution:

$$\begin{aligned} max\, x^{\alpha-1}(1-x)^{\beta-1} &\equiv max\, log[x^{\alpha-1}(1-x)^{\beta-1}] && (85) \\ &= max\, (\alpha-1)log(x) + (\beta-1)log(1-x) && (86) \end{aligned}$$

by taking the derivative and setting it to 0, we get:

$$\begin{aligned} 0 &= \frac{\alpha-1}{x} + \frac{\beta-1}{1-x} && (87) \\ \frac{-\beta+1}{-1+x} &= \frac{\alpha-1}{x} && (88) \\ x(-\beta+1) &= (\alpha-1)(x-1) && (89) \\ -\beta x + x &= -\alpha x - \alpha - x + 1 && (90) \\ \alpha x - \beta x + 2x &= \alpha - 1 && (91) \\ x &= \frac{\alpha-1}{\alpha+\beta-2} && (92) \end{aligned}$$

The mean of the distribution is defined by:

$$\begin{aligned} E[x] &= \int_0^1 xp(x)dx && (93) \\ &= \int_0^1 x\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx && (94) \\ &= \frac{\alpha}{\alpha+\beta} && (95) \end{aligned}$$

The variance of the distribution is defined by:

$$\begin{aligned} Var(X) &= E[(X-\mu)^2] = E[X^2] - E[X]^2 && (96) \\ &= E[X^2] - (\frac{\alpha}{\alpha+\beta})^2 && (97) \\ &= \int_0^1 x^2 p(x)dx - (\frac{\alpha}{\alpha+\beta})^2 && (98) \\ &= \int_0^1 x^2\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx - (\frac{\alpha}{\alpha+\beta})^2 && (99) \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} && (100) \end{aligned}$$

Exercise 2.17. **Suppose X, Y are two points sampled independently and uniformly at random from the interval $[0, 1]$. What is the expected location of the left most point?**

$$
\begin{aligned}
E[min(X, Y)] &= \int_0^1 \int_0^1 min(X, Y)p(X, Y)dxdy && (101)\\
&= \int_0^1 \int_0^1 min(X, Y)p(X)p(Y)dxdy && (102)\\
&= \int_0^1 \int_0^1 min(X, Y)dxdy && (103)\\
&= \frac{1}{2}\int_0^1 \int_0^1 X + Y - |X - Y|dxdy && (104)\\
&= \frac{1}{3} && (105)
\end{aligned}
$$

3 Generative Models for Discrete Data

Exercises

Exercise 3.1. **Derive Equation 3.22 by optimizing the log likelihood in Equation 3.11**

Equation 3.22 states $\hat{\theta}_{MLE} = \frac{N_1}{N}$.

To derive this, we must maximize the log likelihood. Formally, we have to find

$$argmax_\theta \, p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{106}$$
$$argmax_\theta \, log \, p(D|\theta) = N_1 log(\theta) + N_0 log(1-\theta) \tag{107}$$
$$argmin_\theta \, - log \, p(D|\theta) = -N_0 log(1-\theta) - N_1 log(\theta) \tag{108}$$

We will minimize this by taking the derivative equal to 0 and solving for θ:

$$0 = \frac{d}{d\theta} - N_0 log(1-\theta) - N_1 log(\theta) \tag{109}$$
$$= \frac{N_0}{1-\theta} - \frac{N_1}{\theta} \tag{110}$$
$$\frac{N_1}{\theta} = \frac{N_0}{1-\theta} \tag{111}$$
$$N_1(1-\theta) = N_0\theta \tag{112}$$
$$N_1 - N_1\theta = N_0\theta \tag{113}$$
$$N_1 = (N_0 + N_1)\theta \tag{114}$$
$$\frac{N_1}{N} = \theta \tag{115}$$

Exercise 3.2. **Derive the following:**

$$p(D) = \frac{[(\alpha_1)\cdots(\alpha_1 + N_1 - 1)][(\alpha_0)\cdots(\alpha_0 + N_0 - 1)]}{(\alpha)\cdots(\alpha + N - 1)} \tag{116}$$
$$= \frac{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1 + \alpha_0 + N)} \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{117}$$

To derive this, we must use the identity $\Gamma(\alpha) = (\alpha - 1)!$. Also note that

$$\frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} = (\alpha)\cdots(\alpha + k)$$

Using these, and the fact that $\alpha = \alpha_0 + \alpha_1$,

$$p(D) = \frac{[(\alpha_1)\cdots(\alpha_1+N_1-1)][(\alpha_0)\cdots(\alpha_0+N_0-1)]}{(\alpha)\cdots(\alpha+N-1)} \tag{118}$$

$$= \frac{(\Gamma(\alpha_1+N_1)/\Gamma(\alpha_1))(\Gamma(\alpha_0+N_0)/\Gamma(\alpha_0))}{\Gamma(\alpha+N)/\Gamma(\alpha)} \tag{119}$$

$$= \frac{\Gamma(\alpha_1+N_1)\Gamma(\alpha_0+N_0)/\Gamma(\alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha+N)/\Gamma(\alpha)} \tag{120}$$

$$= \frac{\Gamma(\alpha_1+N_1)\Gamma(\alpha_0+N_0)\Gamma(\alpha)}{\Gamma(\alpha+N)\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{121}$$

$$= \frac{\Gamma(\alpha_1+N_1)\Gamma(\alpha_0+N_0)}{\Gamma(\alpha+N)}\frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{122}$$

$$= \frac{\Gamma(\alpha_1+N_1)\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0+\alpha_1+N)}\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{123}$$

Exercise 3.3. **Show that**

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

reduces to $p(x=1|D) = \frac{\alpha_1'}{\alpha_0'+\alpha_1'}$ **when** $n=1$**.**

Let $n=1$, and $x \in \{0,1\}$. The Beta-Binomial model is given by:

$$Bb(x|a,b,n) = \binom{n}{x}\frac{B(x+a, n-x+b)}{B(a,b)}$$

Plugging what we know in,

$$Bb(1|\alpha_1',\alpha_0',1) = \binom{1}{1}\frac{B(1+\alpha_1', 1-1+\alpha_0')}{B(\alpha_1',\alpha_0')} \tag{124}$$

$$= \frac{\Gamma(1+\alpha_1')\Gamma(\alpha_0')/\Gamma(1+\alpha_1'+\alpha_0')}{\Gamma(\alpha_1')\Gamma(\alpha_0')/\Gamma(\alpha_1'+\alpha_0')} \tag{125}$$

$$= \frac{\Gamma(1+\alpha_1')\Gamma(\alpha_0')\Gamma(\alpha_1'+\alpha_0')}{\Gamma(1+\alpha_1'+\alpha_0')\Gamma(\alpha_0')\Gamma(\alpha_1')} \tag{126}$$

$$= \frac{\Gamma(1+\alpha_1')}{(\alpha_0'+\alpha_1'+1)\Gamma(\alpha_1')} \tag{127}$$

$$= \frac{(\alpha_1'+1)\Gamma(\alpha_1')}{(\alpha_0'+\alpha_1'+1)\Gamma(\alpha_1')} \tag{128}$$

$$= \frac{\alpha_1'+1}{\alpha_0'+\alpha_1'+1} \tag{129}$$

Exercise 3.4. **Suppose we toss a coin** $n=5$ **times. Let** X **be the number of heads. Let the prior probability of heads be** $p(\theta) = Beta(\theta|1,1)$**. Compute the posterior** $p(\theta|X<3)$ **up to normalization constant.**

The posterior can be given by $p(\theta|D) = p(D|\theta)p(\theta)$. By plugging in the likelihood (Binomial) and the prior (Beta), we get

$$p(\theta|D) = p(D|\theta)p(\theta) \tag{130}$$
$$\propto Beta(\theta|N_1 + \alpha, N_2 + \beta) \tag{131}$$

Since the number of heads is discrete and mutually exclusive,

$$p(\theta|X < 3) = p(\theta|X = 0) + p(\theta|X = 1) + p(\theta|X = 2) \tag{132}$$
$$\propto Beta(\theta|1, 5) + Beta(\theta|2, 4) + Beta(\theta|3, 3) \tag{133}$$
$$= \sum_{a=0}^{2} Beta(\theta|a + 1, 5 + 1 - a) \tag{134}$$

Exercise 3.5. **Let $\phi = logit(\theta) = log\frac{\theta}{1-\theta}$. Show that if $p(\phi) \propto 1$, then $p(\theta) \propto Beta(\theta|0, 0)$.**

Using the change of variables formula,

$$p(\phi) = p(logit(\theta)) \tag{135}$$
$$= \left|\frac{d\phi}{d\theta}\right| p(log\frac{\theta}{1-\theta}) \tag{136}$$
$$\propto \left|\frac{d}{d\theta}\left(\frac{\theta}{1-\theta}\right)\right| \tag{137}$$
$$= \theta^{-1}(1-\theta)^{-1} \tag{138}$$

Note that a $Beta(\theta|0, 0)$ distribution can be defined as

$$Beta(\theta|0, 0) = \frac{\theta^{0-1}(1-\theta)^{0-1}}{B(0, 0)} \tag{139}$$
$$\propto \theta^{-1}(1-\theta)^{-1} \tag{140}$$

This result is important, because it shows us that the uniform distribution transformed using the logit function is a $Beta(0, 0)$ distribution, which is an uninformative Beta distribution.

Exercise 3.6. **The Poisson pmf is defined as $Poi(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for which $x \in \{0, 1, 2, ...\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.**

The MLE is defined as

$$argmax_{\lambda} p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{141}$$
$$argmax_{\lambda} log\, p(x|\lambda) = -\lambda + xlog\lambda - logx! \tag{142}$$
$$argmin_{\lambda} - log\, p(x|\lambda) = \lambda - xlog\lambda + logx! \tag{143}$$

By taking the derivate and setting it to 0,

$$\begin{aligned}
0 &= \frac{d}{d\lambda}\left|\lambda - xlog\lambda + logx!\right| && (144)\\
&= 1 - \frac{x}{\lambda} && (145)\\
\frac{x}{\lambda} &= 1 && (146)\\
x &= \lambda && (147)
\end{aligned}$$

Therefore, the MLE solution to the Poisson is $x = \lambda$.

Exercise 3.7. **a. Derive the posterior $p(\lambda|D)$ assuming a conjugate prior $p(\lambda) = Ga(\lambda|a,b) \propto \lambda^{a-1}e^{-\lambda b}$.**

The posterior is given by

$$\begin{aligned}
p(\lambda|D) &= p(D|\lambda)p(\lambda) && (148)\\
&\propto e^{-\lambda}\frac{\lambda^x}{x!}\lambda^{a-1}e^{-\lambda b} && (149)\\
&\propto \lambda^x\lambda^{a-1}e^{-\lambda}e^{-\lambda b} && (150)\\
&= \lambda^{x+a-1}e^{-\lambda(b+1)} && (151)\\
&= Ga(a+x, b+1) && (152)
\end{aligned}$$

b. What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $Ga(a,b)$ distribution is a/b.

The posterior mean tends to $Ga(0+x, 0+1) = Ga(x,1) \to x$ when $a \to 0$ and $b \to 0$.

Exercise 3.8. **Consider a uniform distribution centered on 0 with width $2a$. The density function is given by**

$$p(x) = \frac{1}{2a}I(x \in [-a,a])$$

a. Given a data set $x_1, ..., x_n$, what is the maximum likelihood estimate of a (call it $\hat{a}$)?

$$\begin{aligned}
p(D|a) &= \prod_{i=1}^{n} p(x_i|a) && (153)\\
&= \prod_{i=1}^{n}\frac{1}{2a}I(x \in [-a,a]) && (154)\\
&= \frac{1}{(2a)^n}\prod_{i=1}^{n}I(x \in [-a,a]) && (155)
\end{aligned}$$

Since, this is the quantity we want to maximize. Note that it is maximized as a is minimal (first term). The second term nullifies the first term for all x_i that is outside the interval $[-a, a]$. This means that the posterior is maximized for the smallest interval $[-a, a]$ that captures the full range of the data. Formally, the posterior is maximized when

$$\hat{a} = max(|x_i|)$$

b. What probability would the model assign a new data point x_{n+1} using $\hat{a}$?

Since $p(x_{n+1}) = \frac{1}{2\hat{a}} I(x_{n+1} \in [-\hat{a}, \hat{a}])$, it is obvious that x_{n+1} has probability $\frac{1}{2\hat{a}}$ if the point x_{n+1} is in the range $[-\hat{a}, \hat{a}]$, and 0 otherwise.

c. Do you see any problem with the above approach? Briefly suggest (in words) a better approach.

This approach suffers from the zero-count problem. In general, any probability specification that assigns zero probability to inputs that are possible is not ideal. A better solution would to use some Bayesian approach, or add Laplace smoothing.

Exercise 3.9. **Derive the posterior $p(\theta|D)$ of the uniform with a Pareto prior, and show that it can be expressed as a Pareto distribution.**

Note that the Pareto distribution is given by

$$Pareto(\theta|b, K) = bK^b \theta^{-(b+1)} I(\theta \geq K)$$

Using this,

$$\begin{aligned}
p(\theta|D) &= \frac{p(D,\theta)}{p(D)} && (156)\\
&= \frac{\frac{Kb^K}{\theta^{N+K+1}}}{\int_m^\infty \frac{Kb^K}{\theta^{N+K+1}} d\theta} I(\theta \geq max(D)) && (157)\\
&= \begin{cases} \frac{K\theta^{N+K-1}}{K(N+K)b^{N+K}} I(\theta \geq m) & if\, m \leq b \\ \frac{K\theta^{N+K-1}}{Kb^K(N+K)m^{N+K}} I(\theta \geq m) & if\, m > b \end{cases} && (158)\\
&= \frac{\theta^{N+K-1} I(\theta \geq m)}{N+K} \begin{cases} b^{-N-K} & if\, m \leq b \\ b^{-K} m^{-K-N} & if\, m > b \end{cases} && (159)\\
&\propto \theta^{N+K-1} b^{-K} m^{-K-N} I(\theta \geq m) && (160)\\
&= Pareto(\theta| -(K+N), m) && (161)
\end{aligned}$$

Exercise 3.10. **Let's say that taxicars are numbered uniformly like $p(x) = U(0, \theta)$.**

a. Suppose we see one taxi numbered 100, so $D = \{100\}$, $m = 100$, $N = 1$. Using a non-informative prior on θ of the form $p(\theta) = Pa(\theta|0, 0) \propto 1/\theta$, what is the posterior $p(\theta|D)$?

Recall from the previous exercise that the posterior is a Pareto of the form $Pa(\theta|N + K, max(m, b))$. The posterior is then given by

$$
\begin{aligned}
p(\theta|D) &= p(D|\theta)p(\theta) && (162)\\
&= U(0,\theta)Pa(\theta|0,0) && (163)\\
&= Pa(\theta|N+0, max(100,0)) && (164)\\
&= Pa(\theta|1,100) && (165)
\end{aligned}
$$

b. Compute the posterior mean, mode and median number of taxis in the city, if such quantities exist.

We know the form of the posterior, so the posterior mean is the mean of the Pareto distribution, which is given by

$$\mu_{a,b} = \frac{ab}{a-1}$$

therefore, the mean of $Pa(\theta|1,100)$ is $\frac{100}{0}$, which is undefined.

The mode of a $Pa(\theta|a,b)$ is b, so the mode of the posterior is $m = 100$.

The median of a $Pa(\theta|a,b)$ is $2^{1/a}b$, so the mode of the posterior is $2^{1/1} \times 100 = 200$.

c. Compute the predictive density for the next taxicab number.

We can use the above equations to find the prior before witnessing the second taxicab. This prior will be the posterior after seeing the first taxicab number. This posterior is given by $Pa(\theta|1,m)$. Thus, using $b = m$ and $K = 1$, we can plug this into the equation above as

$$
\begin{aligned}
p(x|D,K,b) &= \frac{K}{(N+K)b^N}I(x \leq m) + \frac{Kb^K}{(N+K)m^{N+K}}I(x > m) && (166)\\
&= \frac{1}{(1+1)m^1}I(x \leq m) + \frac{m^1}{(1+1)x^{1+1}}I(x > m) && (167)\\
&= \frac{1}{2m}I(x \leq m) + \frac{m}{2x^2}I(x > m) && (168)
\end{aligned}
$$

d. Use the predictive density to compute the probability that the next taxi you will see (say, the next day) has number 100, 50, or 150, i.e. compute $p(x = 100|D, \alpha)$, etc.

$$
\begin{aligned}
p(x = 100|D,\alpha) &= \frac{1}{2m}I(x \leq m) + \frac{m}{20000}I(x > m) && (169)\\
p(x = 50|d,\alpha) &= \frac{1}{2m}I(x \leq m) + \frac{m}{5000}I(x > m) && (170)\\
p(x = 150|d,\alpha) &= \frac{1}{2m}I(x \leq m) + \frac{m}{45000}I(x > m) && (171)
\end{aligned}
$$

e. Briefly describe some ways we might make the model more accurate at prediction.

We are currently using an uninformative prior, which doesn't seem ideal. There are certain restrictions we could make on the distribution of taxi numbers.

Exercise 3.11. **The exponential distribution with parameter θ is given by $p(x|\theta) = \theta e^{-\theta x}$.**

a. Show that the MLE is given by $\hat{\theta} = 1/\bar{x}$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

The log likelihood is given by

$$
\begin{aligned}
log\, p(x|\theta) &= \sum_{i=1}^{N} log(\theta e^{-\theta x_i}) && (172)\\
&= \sum_{i=1}^{N} log(\theta) - \theta x_i && (173)\\
&= N log(\theta) - \sum_{i=1}^{N} \theta x_i && (174)\\
&= N log(\theta) - \theta \sum_{i=1}^{N} x_i && (175)
\end{aligned}
$$

Setting the derivative to 0,

$$
\begin{aligned}
0 &= \frac{d}{d\theta}\left| N log(\theta) - \theta \sum_{i=1}^{N} x_i \right| && (176)\\
&= \frac{N}{\theta} - \sum_{i=1}^{N} x_i && (177)\\
\sum_{i=1}^{N} x_i &= \frac{N}{\theta} && (178)\\
\theta &= \frac{N}{\sum_{i=1}^{N} x_i} && (179)\\
&= \frac{1}{\bar{x}} && (180)
\end{aligned}
$$

b. Suppose we observe $X_1 = 5$, $X_2 = 6$, $X_3 = 4$. What is the MLE given this data?

The MLE is one over the arithmetic mean, which is $1/mean(5, 6, 4) = 1/5$.

c. Assume that an expert believe θ should have a prior distribution that is also exponential $p(\theta) = Expon(\theta|\lambda)$. Choose the prior parameter, call it $\hat{\lambda}$, such that $E[\theta] = 1/3$.

Note that the exponential distribution is just a special case of the Gamma distribution. In particular, $Expon(x|\theta) = Gamma(x|1, 1/\theta)$. Since we know that the mean of the Gamma distribution is a/b, then we can find the exponential with mean of 1/3 through the Gamma:

$$Gamma(\theta|1, 3) = Expon(\theta|1/3)$$

d. What is the posterior $p(\theta|D, \hat{\lambda})$?

$$
\begin{aligned}
p(\theta|D, \hat{\lambda}) &= p(D|\theta, \hat{\lambda})p(\hat{\lambda}) && (181)\\
&= \prod_{i=1}^{N} \theta e^{-\theta x_i} \theta e^{-\theta \lambda} && (182)\\
&= \prod_{i=1}^{N} \theta^2 e^{-\theta(x_i+\lambda)} && (183)\\
&= \theta^{2N} \prod_{i=1}^{N} e^{-\theta\lambda-\theta x_i} && (184)\\
&= \theta^{2N} e^{-\theta(\lambda+\sum_{i=1}^{N} x_i)} && (185)\\
&= Gamma(\theta|2N, \lambda + \sum_{i=1}^{N} x_i) && (186)
\end{aligned}
$$

e. Is the exponential prior conjugate to the exponential likelihood?

Yes, both the prior and the likelihood are of the Gamma distribution (remember the exponential distribution is a special case of the Gamma distribution).

f. What is the posterior mean, $E[\theta|D, \hat{\lambda}]$?

The posterior is a Gamma as shown above, which has a mean of a/b.

g. Explain why the MLE and posterior mean differ. Which is more reasonable in this example?

Since the posterior comes from an informative prior ($\hat{\lambda}$), the posterior and the MLE will be different, but equal as $N \to \infty$.

In this example, the posterior is more reasonable, since the prior is more informative.

Exercise 3.12. **The book discussed using a Beta prior for a Bayesian inference of a Bernoulli rate parameter.**
a. Now consider the following prior, that believes the coin is fair, or is slightly biased towards tails:

$$
\begin{aligned}
p(\theta) &= \begin{cases} 0.5 & if\, \theta = 0.5 \\ 0.5 & if\, \theta = 0.4 \\ 0 & otherwise \end{cases} && (187)\\
&= 0.5I(\theta - 0.5 = 0) + 0.5I(\theta - 0.4 = 0) && (188)
\end{aligned}
$$

Derive the MAP estimate under this prior as a function of N_1 and N.

The posterior is given by

$$
\begin{aligned}
p(\theta|D) &= p(D|\theta)p(\theta) && (189)\\
&= \theta^{N_1}(1-\theta)^{N_0}p(\theta) && (190)\\
&= \theta^{N_1}(1-\theta)^{N_0}(0.5I(\theta-0.5=0)+0.5I(\theta-0.4=0)) && (191)\\
&= 0.5^{N_1+N_0+1}I(\theta-0.5=0)+0.5(0.4^{N_1})(0.6^{N_0})I(\theta-0.4=0) && (192)
\end{aligned}
$$

Note that the prior is so restrictive that the likelihood is 0 for all θ except for 0.4 and 0.5. Thus, we can actually compute the likelihood for both of these values of θ and find which one maximizes the likelihood.

So, for each value of θ, the posterior is

$$
\begin{aligned}
p(0.4|D) &\propto (0.4^{N_1})(0.6^{N_0}) && (193)\\
p(0.5|D) &\propto 0.5^{N_1+N_0} && (194)
\end{aligned}
$$

These are functions of N_0 and N_1, and the value of θ that maximizes the posterior will depend of these. We can find these constraints by calling one the MAP and seeing the requirements needed for N_1 and N_0. Let's say that $\theta = 0.4$:

$$
\begin{aligned}
(0.4^{N_1})(0.6^{N_0}) &\geq 0.5^{N_1+N_0} && (195)\\
N_1 log(0.4) + N_0 log(0.6) &\geq (N_1+N_0)log(0.5) && (196)\\
N_1(log(0.4)-log(0.5)) &\geq N_0(log(0.5)-log(0.6)) && (197)\\
N_1 log(\frac{4}{5}) &\geq N_0 log(\frac{5}{6}) && (198)\\
N_1 &\geq \frac{log(5/6)}{log(4/5)}N_0 && (199)\\
&\approx 0.8171N_0 && (200)
\end{aligned}
$$

Thus, when $N_1 \geq 0.8171N_0$, then $\theta_{MAP} = 0.4$, otherwise $\theta_{MAP} = 0.5$.

b. Suppose the true parameter is $\theta = 0.41$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large?

Note the "other" prior in this is when you use a $Beta(\theta|\alpha,\beta)$ prior, which leads to the MAP

$$\hat{\theta} = \frac{N_1+\alpha-1}{N_1+N_0+\alpha+\beta-2}$$

With small datasets, the prior can overwhelm the posterior. Thus, your choice of Beta could greatly influence the posterior in small datasets. For the handmade prior above, the worst that could happen is $\theta = 0.5$, which results in small error, whereas you could have worse error using a Beta prior.

For large datasets, note that the best you can do with the handmade prior is $\theta = 0.4$. When the true value is 0.41, this is not bad error, but note that using a conjugate prior with large datasets tends to the MLE solution, which, with a large enough dataset can get arbitrarily precise.

Exercise 3.13. **Derive the posterior predictive distribution for a batch of data with the dirichlet-multinomial model.**

Note that the predictive distribution for a single data point is given by

$$p(X = j|D) = \frac{\alpha_j + N_j}{\alpha_0 + N}$$

Since the assumption is that all data points are i.i.d, we can use this as a jumping off point:

$$\begin{aligned}
p(\tilde{D}|D, \alpha) &= p(x_1|D, \alpha)p(x_2|D, \alpha, x_1) \cdots p(x_n|D, \alpha, x_1, x_2, \cdots, x_{n-1}) &\quad (201)\\
&= \frac{\prod_{j=1}^{K} \prod_{i=1}^{N_j^{new}-1} \alpha_j + N_j^{old} + i}{\prod_{i=1}^{N-1} \alpha + N^{old} + i} &\quad (202)\\
&= \frac{\prod_{j=1}^{K} (\alpha_j + N_j^{old} + N_j^{new} - 1)!/(\alpha_j + N_j^{old})!}{(\alpha + N^{old} + N - 1)!/(\alpha + N^{old})!} &\quad (203)\\
&= \frac{\prod_{j=1}^{K} \Gamma(\alpha_j + N_j)/\Gamma(\alpha_j + N_j^{old})}{\Gamma(\alpha + N)/\Gamma(\alpha + N^{old})} &\quad (204)\\
&= \frac{\Gamma(\alpha + N^{old})}{\Gamma(\alpha + N)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + N_j)}{\Gamma(\alpha_j + N_j^{old})} &\quad (205)
\end{aligned}$$

Exercise 3.14. **a. Suppose we compute the empirical distribution over letters of the Roman alphabet plus the space character (a distribution over 27 values) from 2000 samples. Suppose we see the letter ”e” 260 times. What is $p(x_{2001} = e|D)$, if we assume $\theta \sim Dir(\alpha_1, ..., \alpha_{27})$, where $\alpha_k = 10$ for all k?**

Recall that the posterior predictive of the Dirichlet-multinomial model is

$$p(X = j|D) = \frac{\alpha_j + N_j}{\alpha_0 + N}$$

Given that $\alpha_k = 10$ for all k, this is simply

$$\begin{aligned}
p(x_{2001} = e|D) &= \frac{10 + 260}{\sum_{k=1}^{K} \alpha_k + 2000} &\quad (206)\\
&= \frac{270}{2270} \approx 0.119 &\quad (207)
\end{aligned}$$

b. Suppose, in the 2000 samples, we saw ”e” 260 times, ”a” 100 times, and ”p” 87 times. What is $p(x_{2001} = p, x_{2002} = a|D)$, if we assume $\theta \sim Dir(\alpha_1, ..., \alpha_{27})$, where $\alpha_k = 10$ for all k?

Note that

$$p(x_{2001} = p, x_{2002} = a|D) = p(x_{2001} = p|D)p(x_{2002} = a|D)$$

since they are conditionally independent events. Using the same framework as above, and letting $\alpha = \sum_{k=1}^{K} \alpha_k = 270$,

$$\begin{aligned} p(x_{2001} = p, x_{2002} = a|D) &= \frac{\alpha_p + N_p}{\alpha + N} \frac{\alpha_a + N_a}{\alpha + N} && (208) \\ &= \frac{97 \times 110}{(270 + 2000)^2} && (209) \\ &\approx 0.0021 && (210) \end{aligned}$$

Exercise 3.15. **Suppose $\theta \sim \beta(\alpha_1, \alpha_2)$, and we believe that $E[\theta] = m$ and $var[\theta] = v$. Using Equation 2.62, solve for α_1 and α_2 in terms of m and v. What values do you get if $m = 0.7$ and $v = 0.22$?**

Equation 2.62 states that

$$mean = \frac{a}{a+b}, mode = \frac{a-1}{a+b-2}, var = \frac{ab}{(a+b)^2(a+b+1)}$$

for a Beta distribution. Using these, we get a system of equations

$$\begin{aligned} m &= \frac{a}{a+b} && (211) \\ m(a+b) &= a && (212) \\ mb &= a(1-m) && (213) \\ b &= \frac{a(1-m)}{m} && (214) \end{aligned}$$

By plugging this into the variance function above, it can be shown that

$$a = m\left(\frac{m(1-m)}{v} - 1\right)$$

which you can then plug back into the equation for b above to get

$$b = (1-m)\left(\frac{m(1-m}{v} - 1\right)$$

If $m = 0.7$ and $v = 0.2^2 = 0.04$, then

$$a = 0.7\left(\frac{0.7(1-0.7)}{0.04} - 1\right) = 2.975$$

and

$$b = (1-0.7)\left(\frac{0.7(1-0.7)}{0.04} - 1\right) = 1.275$$

Exercise 3.16. **Suppose $\theta \sim \beta(\alpha_1, \alpha_2)$ and we believe that $E[\theta] = m$ and $p(l < \theta < u) = 0.95$. Write a program that can solve for α_1 and α_2 in terms of m, and u.**

We know the mean, so we can write

$$\begin{aligned} m &= \frac{\alpha_1}{\alpha_1 + \alpha_2} && (215) \\ m(\alpha_1 + \alpha_2) &= \alpha_1 && (216) \\ m\alpha_2 &= \alpha_1 - m\alpha_1 && (217) \\ \alpha_2 &= \frac{\alpha_1}{m} - \alpha_1 && (218) \end{aligned}$$

We are given the quantiles, which we can express as

$$\int_l^u \frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = I_u(\alpha_1, \alpha_2) - I_l(\alpha_1, \alpha_2) \qquad (219)$$

where $I_x(\alpha_1, \alpha_2) = \int_0^x Beta(\alpha_1, \alpha_2)$ is the regularized incomplete beta function. We can minimize the squared discrepancy between this and 0.95:

$$\begin{aligned} 0 &= \frac{d}{d\theta}[(I_u(\alpha_1, \alpha_2) d\theta - I_l(\alpha_1, \alpha_2) d\theta - 0.95)^2] && (220) \\ 0 &= I_u(\alpha_1, \alpha_2) - I_l(\alpha_1, \alpha_2) - 0.95 && (221) \end{aligned}$$

Since this exercise involves writing a program, the code for this program is found in an IPython notebook in this same directory.

Exercise 3.17. **Suppose we toss a coin N times and observe N_1 heads. Let $N_1 \sim Bin(N, \theta)$ and $\theta \sim Beta(1, 1)$. Show that the marginal likelihood is $p(N_1|N) = 1/(N + 1)$.**

The key here is that N_1 and N are sufficient statistics. Remember that the posterior of the Beta-Binomial model is given by

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} && (222) \\ &= Beta(\theta|N_1 + a, N_0 + b) && (223) \\ p(D) &= \frac{p(D|\theta)p(\theta)}{Beta(\theta|N_1 + a, N_0 + b)} && (224) \end{aligned}$$

The marginal likelihood is given by

$$\begin{aligned} p(N_1|N) &= \int_\theta \frac{p(N_1|\theta, N)p(\theta|N)}{p(N_1, N)} d\theta && (225) \\ &= \int_\theta \frac{Bin(N_1|\theta, N) Beta(\theta|1, 1)}{Beta(\theta|N_1 + 1, N_0 + 1)} d\theta && (226) \end{aligned}$$

Since we are marginalizing over θ, we can rewrite these using Beta functions, like

$$p(N_1|N) = \binom{N}{N_1} \frac{B(N_1+1, N-N_1+1)}{B(N_1+1, N-N_1+1)/B(N_1, N-N_1)} \tag{227}$$

$$= \binom{N}{N_1} \frac{B(N_1+1, N-N_1+1)}{B(1,1)} \tag{228}$$

$$= \binom{N}{N_1} \frac{\Gamma(N_1+1)\Gamma(N-N_1+1)}{\Gamma(N+2)} \tag{229}$$

$$= \frac{N!}{N_1!(N-N_1)!} \frac{\Gamma(N_1+1)\Gamma(N-N_1+1)}{\Gamma(N+2)} \tag{230}$$

$$= \frac{N!N_1!(N-N_1)!}{N_1!(N-N_1)!(N+1)!} \tag{231}$$

$$= \frac{N!}{(N+1)N!} \tag{232}$$

$$= \frac{1}{N+1} \tag{233}$$

Exercise 3.18. **Suppose we toss a coin $N = 10$ times and observe $N_1 = 9$ heads. Let the null hypothesis be that the coin is fair, and the alternative be that the coin can have any bias, so $p(\theta) = Unif(0, 1)$. Derive the Bayes factor $BF_{1,0}$ in favor of the biased coin hypothesis. What if $N = 100$ and $N_1 = 90$?**

The Bayes factor is defined as

$$BF_{1,0} = \frac{p(D|alt)}{p(D|null)}$$

Let's look into the null hypothesis first. The null hypothesis says that the coin is not biased, meaning $\theta = 0.5$. Thus, the likelihood is

$$p(N_1|\theta = 0.5) = \binom{N}{N_1} 0.5^{N_1} 0.5^{N-N_1} = \binom{N}{N_1} 0.5^N$$

Note that the alternative hypothesis marginalizes across all θ, and as we saw in the last exercise, this is $\frac{1}{N+1}$.

So, the Bayes Factor is

$$BF_{1,0} = \frac{1}{\binom{N}{N_1}(N+1)0.5^N} \tag{234}$$

$$= \frac{2^N}{\binom{N}{N_1}(N+1)} \tag{235}$$

Thus, is $N = 10$ and $N_1 = 9$, then the Bayes Factor is 9.31. This is moderately strong evidence to accept the alternative.

If $N = 100$ and $N_1 = 90$, then the Bayes Factor is 7.251×10^{14}. This is very strong evidence to accept the alternative.

Exercise 3.19. **This question sets up Naive Bayes as a linear classifier. a. Write down an expression for the log posterior odds ratio, in terms of the features and the parameters.**

The log posterior odds ratio is

$$\begin{aligned} log\frac{p(c=1|x_i)}{p(c=2|x_i)} &= log\frac{p(x_i|c=1,\theta)p(c=1)}{p(x_i|c=2,\theta)p(c=2)} && (236)\\ &= log\frac{p(x_i|c=1,\theta)}{p(x_i|c=2,\theta)} && (237)\\ &= logp(x_i|c=1,\theta) - logp(x_i|c=2,\theta) && (238)\\ &= \phi(x_i)^T\beta_1 - \phi(x_i)^T\beta_2 && (239)\\ &= \phi(x_i)^T(\beta_1 - \beta_2) && (240) \end{aligned}$$

b. Intuitively, words that occur in both classes are not very "discriminative", and therefore should not affect our beliefs about the class label. Consider a particular word w. State the conditions on $\theta_{1,w}$ and $\theta_{2,w}$ (or equivalently the conditions on $\beta_{1,w}$, $\beta_{2,w}$) under which the presence or absence of w in a test document will have no effect on the class posterior (such a word will be ignored by the classifier). Hint: using your previous result, figure out when the posterior odds ratio is $0.5/0.5$.

For a word w to have no effect on the posterior, the log posterior odds should equal 1. Since the model is linear, we can narrow this down to one word. We also note that we are considering words that exist in both classes, so $\phi(x_{i,w}) = 1$. So,

$$\begin{aligned} 1 &= \phi(x_i)^T(\beta_1 - \beta_2) && (241)\\ \beta_1 &= \beta_2 && (242)\\ log\frac{\theta_{1,w}}{1-\theta_{1,w}} &= log\frac{\theta_{2,w}}{1-\theta_{2,w}} && (243)\\ \frac{\theta_{1,w}}{1-\theta_{1,w}} &= \frac{\theta_{2,w}}{1-\theta_{2,w}} && (244)\\ \theta_{1,w}(1-\theta_{2,w}) &= \theta_{2,w}(1-\theta_{1,w}) && (245)\\ \theta_{1,w} - \theta_{1,2}\theta_{2,w} &= \theta_{2,w} - \theta_{1,w}\theta_{2,w} && (246)\\ \theta_{1,w} &= \theta_{2,w} && (247) \end{aligned}$$

c. Let there be n_1 documents of class 1 and n_2 be the number of documents in class 2, where $n_1 = n_2$ (since e.g., we get much more non-spam than spam; this is an example of class imbalance). If we use

the above estimate for $\theta_{c,w}$, will word w be ignored by our classifier? Explain why or why not.

Since the word is in all documents, then the given estimates $\hat{\theta}_{cw}$ are

$$\hat{\theta}_{1w} = \frac{1+n_1}{2+n_1} \tag{248}$$

$$\hat{\theta}_{2w} = \frac{1+n_2}{2+n_2} \tag{249}$$

Since $n_1 \neq n_2$, these quantities are not equal. We saw in part (b) that the necessary requirement for the model to ignore a word is for $\theta_{1w} = \theta_{2w}$, so we can be sure that the model will not ignore this word.

d. What other ways can you think of which encourage "irrelevant" words to be ignored?

Weighting each word by frequency using TF-IDF for example.

Exercise 3.20. **a. How would you specify a "full" model that doesn't use Naive Bayes assumption? How many parameters would it have?**

The Naive Bayes assumption allows for significant simplification in the model specification. Without it, the best you could do is to use the chain-rule of probability:

$$p(x_{1:D}|y=c) = p(x_1|y=c)p(x_2|x_1, y=c)\cdots p(x_D|x_1,...,x_{D-1}, y=c)$$

The Naive Bayes assumption allows us to trim down the contingency table to a workable amount. Since the features are binary, the number of parameters in the full model is 2^D.

b. Assume the number of features D is fixed. Let there be N training cases. If the sample size N is very small, which model (naive Bayes or full) is likely to give lower test set error, and why?

The number of parameters in the naive Bayes model is DC vs. 2^D for the full model. So, if N is very small while D remains fixed, it is very likely that the full model will be over-parameterized and overfit on the training set. Therefore, in this case, the naive Bayes model will perform better on the test set, since it avoids the curse of dimensionality.

c. What if the sample size N was very large?

In this case, the conditional independence assumption that the naive Bayes model uses may be too restrictive to capture the patterns in the data, and therefore the full model will likely perform better.

d. What is the computational complexity of fitting the full and naive Bayes model as a function of N and D?

Both of the models are $O(ND)$ worst case, since we can assume that it takes $O(D)$ time to convert a bit array to an array index.

e. What is the computational complexity at test time for the full and naive Bayes model?

The complexity for naive Bayes at test time is $O(CD)$. For the full model, we loop through the classes and lookup the joint probability to use as the prediction. So, the complexity is $O(CD)$. Note that in the full model, D is a much larger number than in naive Bayes.

f. Suppose the test case has missing data. Let x_v be the visible features of size v, and x_h be the hidden (missing) features of size h, where $v + h = D$. What is the computational complexity of computing $p(y|x_v, \hat{\theta})$ for the full and naive Bayes models, as a function of v and h?

The naive Bayes model could just skip over the missing features, so therefore the complexity would still be $O(CD)$. The full model, however, would have to create entries for all possible combinations of missing and non-missing features, which would be $O(2^h D)$.

Exercise 3.21. **Derive equation 3.76**

$$\begin{aligned} I(X,Y) &= \sum_{x_j}\sum_{y} p(x_j,y) log \frac{p(x_j,y)}{p(x_j)p(y)} && (250)\\ &= \sum_{x_j}\sum_{y} p(x_j|y)p(y) log \frac{p(x_j|y)p(y)}{p(x_j)p(y)} && (251)\\ &= \sum_{x_j}\sum_{y} p(x_j|y)p(y) log \frac{p(x_j|y)}{p(x_j)} && (252)\end{aligned}$$

Since it is given that the features are binary, we can expand the summation:

$$\begin{aligned} I(X,Y) &= \sum_{x_j}\sum_{y} p(x_j|y)p(y) log \frac{p(x_j|y)}{p(x_j)} && (253)\\ &= \sum_{y} p(x_j|y)p(y) log \frac{p(x_j|y)}{p(x_j)} + (1 - p(x_j|y)p(y) log \frac{1 - p(x_j|y)}{1 - p(x_j)} && (254)\\ &= \sum_{y} \theta_{jy}\pi_y log \frac{\theta_{jy}}{\theta_j} + (1 - \theta_{jy})\pi_y log \frac{1-\theta_{jy}}{1-\theta_j} && (255)\end{aligned}$$

4 Gaussian Models

Exercises

Exercise 4.1. **Let $X \sim U(1, 1)$ and $Y = X^2$. Clearly Y is dependent on X (in fact, Y is uniquely determined by X). However, show that $\rho(X, Y) = 0$. Hint: if $X \sim U(a, b)$ then $E[X] = (a + b)/2$ and $var[X] = (ba)^2/12$.**

Let's plug things into the definition of correlation:

$$\begin{aligned}
\rho(X, Y) &= \frac{cov(X, Y)}{\sigma_X \sigma_Y} \qquad (256)\\
&= \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} \qquad (257)\\
&= \frac{E[X^3] - E[X]E[X^2]}{\sigma_X \sigma_{X^2}} \qquad (258)
\end{aligned}$$

Note that to show this equals 0, we just have to show that the numerator is equal to 0. To do this, we will compute each term:

$$E[X^3] = \frac{1}{2}\int_{-1}^{1} u^3 p(u) du = 0$$

$$E[X^2] = \frac{1}{2}\int_{-1}^{1} u^2 p(u) du = \frac{1}{3}$$

$$E[X] = \frac{-1 + 1}{2} = 0$$

So we have

$$\begin{aligned}
\rho(X, Y) &= \frac{E[X^3] - E[X]E[X^2]}{\sigma_X \sigma_{X^2}} \qquad (259)\\
&= \frac{0 - 0 \times \frac{1}{3}}{\sigma_X \sigma_{X^2}} \qquad (260)\\
&= 0 \qquad (261)
\end{aligned}$$

Exercise 4.2. **Let $X \sim N(0, 1)$ and $Y = WX$, where $p(W = 1) = p(W = 1) = 0.5$. It is clear that X and Y are not independent, since Y is a function of X.**

a. Show that $Y \sim N(0, 1)$.

So, W randomly changes the sign half the time on X. Thus,

$$Y \sim \frac{1}{2}N(0, 1) - \frac{1}{2}N(0, 1)$$

Let's write the distribution out for this:

$$
\begin{aligned}
p(Y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(WX-\mu)^2}{2\sigma^2}} && (262)\\
&= \frac{1}{2\sqrt{2\pi\sigma^2}} e^{-\frac{(-X-\mu)^2}{2\sigma^2}} + \frac{1}{2\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} && (263)\\
&= \frac{1}{2\sqrt{2\pi}} e^{-\frac{(-X)^2}{2}} + \frac{1}{2\sqrt{2\pi}} e^{-\frac{X^2}{2}} && (264)\\
&= \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{X^2}{2}} + e^{-\frac{X^2}{2}}\right) && (265)\\
&= \frac{1}{2\sqrt{2\pi}} \left(2e^{-\frac{X^2}{2}}\right) && (266)\\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} && (267)\\
&= N(0,1) && (268)
\end{aligned}
$$

b. Show that $cov[X,Y] = 0$.

$$
\begin{aligned}
cov[X,Y] &= E[XY] - E[X]E[Y] && (269)\\
&= E[E[XY|W]] - E[X]E[WX] && (270)\\
&= \frac{1}{2}E[X^2] + \frac{1}{2}E[-X^2] - E[X]E[WX] && (271)\\
&= \frac{1}{2}E[X^2] + \frac{1}{2}E[X^2] - E[X]E[WX] && (272)\\
&= E[X^2] - E[X](\frac{1}{2}E[^2] + \frac{1}{2}E[-X^2]) && (273)\\
&= E[X^2] - E[X^2] && (274)\\
&= 0 && (275)
\end{aligned}
$$

Exercise 4.3. **Prove that $-1 \leq \rho(X,Y) \leq 1$**

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

This is trivial to prove with the Cauchy-Swartz inequality which states that

$$|cov(X,Y)| \leq \sqrt{\sigma_X^2 \sigma_Y^2}$$

because for $\rho(X,Y)$ to be > 1 or < -1, then $|cov(X,Y)| > \sqrt{\sigma_X^2 \sigma_Y^2}$, which is false.

Exercise 4.4. **Show that, if $Y = aX + b$ for some parameters $a > 0$ and b, then $\rho(X,Y) = 1$. Similarly show that if $a < 0$, then $\rho(X,Y) = -1$.**

$$\begin{aligned}\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} && (276)\\ &= \frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{E[(X-E[X])^2]E[(Y-E[Y])^2]}} && (277)\\ & && (278)\end{aligned}$$

Note that the quantity $(Y - E[Y])$ can be written as

$$\begin{aligned}Y - E[Y] &= aX + b - E[aX + b] && (279)\\ &= aX + b - b - aE[X] && (280)\\ &= a(X - E[X]) && (281)\end{aligned}$$

Plugging this, we get

$$\begin{aligned}\rho(X,Y) &= \frac{aE[(X-E[X])(X-E[X])]}{|a|\sqrt{E[(X-E[X])^2]E[(X-E[X])^2]}} && (282)\\ &= \frac{E[(X-E[X])(X-E[X])]}{E[(X-E[X])^2]} && (283)\\ &= \frac{E[(X-E[X])^2]}{E[(X-E[X])^2]} && (284)\\ &= 1 && (285)\end{aligned}$$

If $a < 0$, then this changes to

$$\begin{aligned}\rho(X,Y) &= \frac{aE[(X-E[X])^2]}{|a|\sqrt{E[(X-E[X])^2]E[(X-E[X])^2]}} && (286)\\ &= -\frac{E[(X-E[X])^2]}{E[(X-E[X])^2]} && (287)\\ &= -1 && (288)\end{aligned}$$

Exercise 4.5. **Derive the normalization constant for multivariate Gaussian.**

We are trying to show that

$$(2\pi)^{D/2}|\Sigma|^{1/2} = \int exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))dx$$

Using eigenvalue decomposition on $\Sigma = U\Lambda U^T$, we can write this as

$$(2\pi)^{D/2}|\Sigma|^{1/2} = \int exp(-\frac{1}{2}(x-\mu)^T U\Lambda^{-1}U^T(x-\mu))dx \tag{289}$$

$$= \int exp(-\frac{1}{2}u^T\Lambda^{-1}u)du \tag{290}$$

$$= \int exp(-\frac{1}{2}\sum_d \frac{u_d^2}{\lambda_d})du \tag{291}$$

$$= \prod_{i=1}^{D}\int exp(-\frac{u_i^2}{2\lambda_i})du \tag{292}$$

Note that this is the product of single dimensional Gaussians. We know that $\int exp(-\frac{u^2}{2\sigma^2}) = \sqrt{2\pi\sigma^2}$, and so we can rewrite this expression as

$$\prod_{i=1}^{D}\int exp(-\frac{u_i^2}{2\lambda_i})du = \prod_{i=1}^{D}\sqrt{2\pi\lambda_i} \tag{293}$$

$$= (2\pi)^{D/2}\prod_{i=1}^{D}\lambda_i^{1/2} \tag{294}$$

$$= (2\pi)^{D/2}|\Sigma|^{1/2} \tag{295}$$

Exercise 4.6. **Derive the pdf of the bivariate Guassian with Σ given.**

Note that

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix} \tag{296}$$

and

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \frac{1}{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix} \tag{297}$$

$$= \frac{1}{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}\begin{bmatrix}\sigma_1^2(x_1-\mu_1)+\rho\sigma_1\sigma_2(x_2-\mu_n) & \rho\sigma_1\sigma_2(x_1-\mu_1)+\sigma_2^2(x_2-\mu_n)\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix} \tag{298}$$

$$= \frac{1}{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}(x_1-\mu_1)(\sigma_1^2(x_1-\mu_1)+\rho\sigma_1\sigma_2(x_2-\mu_2))+(x_2-\mu_2)(\rho\sigma_1\sigma_2(x_1-\mu_1)+\sigma_2^2(x_2-\mu_2)) \tag{299}$$

$$= \frac{1}{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}\sigma_1^2(x_1-\mu_1)^2+2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2)+\sigma_2^2(x_2-\mu_2)^2 \tag{300}$$

$$= \frac{1}{1-\rho^2}\frac{\sigma_1^2(x_1-\mu_1)^2}{\sigma_1^2\sigma_2^2}+\frac{2\rho\sigma_1\sigma_2(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1^2\sigma_2^2}+\frac{\sigma_2^2(x_2-\mu_2)^2}{\sigma_1^2\sigma_2^2} \tag{301}$$

$$= \frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}+2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2}\right) \tag{302}$$

We see that this is the quantity requested of us in the exercise.

Exercise 4.7. **Compute the conditional probability distribution of the given bivariate Gaussian.**

Note that the conditional probability of two Gaussians is a Gaussian. Also the conditional probability distribution is given by

$$p(x_1|x_2) = N(x_1|\mu_{1|2},\Sigma_{1|2}) \tag{303}$$

$$\mu_{1|2} = \mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \tag{304}$$

$$\Sigma_{1|2} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{305}$$

It simplifies things greatly that we are considering a bivariate Gaussian, since Σ_{jk} becomes a scalar. Thus, after plugging in to the equations given in the problem,

$$p(x_2|x_1) = N(x_2|\mu_{2|1},\Sigma_{2|1}) \tag{306}$$

$$\mu_{2|1} = \mu_2+\sigma_1\sigma_2(\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)) \tag{307}$$

$$= \mu_2+\rho\sigma_2^2(x_1-\mu_1) \tag{308}$$

$$\Sigma_{2|1} = \sigma_1\sigma_2\frac{\sigma_2}{\sigma_1}-\sigma_1\sigma_2\rho^2\frac{\sigma_2}{\sigma_1} \tag{309}$$

$$= \sigma_2^2+\rho^2\sigma_2^2 \tag{310}$$

$$p(x_2|x_1) = N(x_2|\mu_2+\rho\sigma_2^2(x_1-\mu_1),\rho^2\sigma_2^2) \tag{311}$$

If $\sigma_2 = \sigma_1 = 1$, then

$$p(x_2|x_1) = N(x_2|\mu_2 + \rho(x_1 - \mu_1), \rho^2)$$

Exercise 4.8. **This exercise is shown in the R notebook "ch4-8.ipynb"**

Exercise 4.9. **Suppose you have two sensors with known (and different) variances ν_1 and ν_2, but unknown and same mean μ. What is the posterior $p(\mu|D)$, assuming a non-informative prior for μ?**

In section 4.4.2.1 we saw that the posterior of some observed data from some noisy measurements of this is given by

$$p(\mu|y_1, y_2, ..., y_n) = 0p(\mu|D, \Sigma) \qquad = N(\mu|m_N, V_N) \tag{312}$$

$$V_N^{-1} = V_0^{-1} + N\Sigma^{-1} \tag{313}$$

$$m_N = V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1}m_0) \tag{314}$$

$$\tag{315}$$

By assuming an uninformative prior, we are saying that $V_0 = \infty I$, which simplifies these to

$$p(\mu|D, \Sigma) = N(\mu|\bar{x}, \frac{1}{N}\Sigma) \tag{316}$$

$$\tag{317}$$

TODO

Exercise 4.10. **Derive the information form results of Section 4.3.1.**

The information form the Gaussian distribution is given by

$$N(x|\xi, \Lambda) = (2\pi)^{D/2}|\Lambda|^{1/2} exp\left[-\frac{1}{2}(x^T\Lambda x + \xi^T\Lambda^{-1}\xi - 2x^T\xi)\right] \tag{318}$$

$$\propto exp\left[-\frac{1}{2}(x^T\Lambda x + \xi^T\Lambda^{-1}\xi - 2x^T\xi)\right] \tag{319}$$

$$= exp\left[-\frac{1}{2}\begin{pmatrix}x_1\\x_2\end{pmatrix}^T\begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\\Lambda_{21} & \Lambda_{22}\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} + \begin{pmatrix}\xi_1\\\xi_2\end{pmatrix}^T\begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\\Lambda_{21} & \Lambda_{22}\end{pmatrix}^{-1}\begin{pmatrix}\xi_1\\\xi_2\end{pmatrix} - 2\begin{pmatrix}x_1\\x_2\end{pmatrix}^T\begin{pmatrix}\xi_1\\\xi_2\end{pmatrix}\right] \tag{320}$$

$$= exp[-\frac{1}{2}x_2(x_1\Lambda_{12} + x_2\Lambda_{22}) + x_1(x_1\Lambda_{11} + x_2\Lambda_{21}) + \tag{321}$$

$$\begin{pmatrix}\xi_1\\\xi_2\end{pmatrix}^T\begin{pmatrix}I & 0\\-\Lambda_{22}^{-1}\Lambda_{21} & I\end{pmatrix}\begin{pmatrix}(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})^{-1} & 0\\0 & \Lambda_{22}^{-1}\end{pmatrix}\begin{pmatrix}I & -\Lambda_{12}\Lambda_{22}^{-1}\\0 & I\end{pmatrix}\begin{pmatrix}\xi_1\\\xi_2\end{pmatrix} - 2x_1\xi_1 - 2x_2\xi_2] \tag{322}$$

$$N(x|\xi, \Lambda) = N(x|\Sigma^{-1}\mu, \Sigma^{-1})$$

The statements we are trying to prove is

$$p(x_1) = N(x_1|\mu_1, \Sigma_{11}) \tag{323}$$
$$p(x_2) = N(x_2|\mu_2, \Sigma_{22}) \tag{324}$$
$$p(x_1|x_2) = N(x_1|\mu_{1|2}, \Sigma_{1|2}) \tag{325}$$
$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{326}$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1} \tag{327}$$

Exercise 4.11. **Derive equation 4.209**

The posterior is given by

$$p(\mu, \Sigma|D) = \frac{p(D|\mu, \Sigma)p(\mu, \Sigma)}{p(D)} \tag{328}$$
$$\propto p(D|\mu, \Sigma)NIW(\mu, \Sigma|m_0, \kappa_0, \upsilon_0, S_0) \tag{329}$$
$$= (2\pi)^{ND/2}|\Sigma|^{-\frac{N}{2}}exp(-\frac{N}{2}(\mu - \bar{x})^T\Sigma^{-1}(\mu - \bar{x}) - \frac{1}{2}tr(\Sigma^{-1}S_{\bar{x}})) \tag{330}$$
$$\times NIW(\mu, \Sigma|m_0, \kappa_0, \upsilon_0, S_0) \tag{331}$$
$$\propto |\Sigma|^{-\frac{N}{2}}exp(-\frac{N}{2}(\mu - \bar{x})^T\Sigma^{-1}(\mu - \bar{x}) - \frac{1}{2}tr(\Sigma^{-1}S_{\bar{x}})) \tag{332}$$
$$\times |\Sigma|^{-\frac{\upsilon_0+D+2}{2}}exp(-\frac{\kappa_0}{2}(\mu - m_0)^T\Sigma^{-1}(\mu - m_0) - \frac{1}{2}tr(\Sigma^{-1}S_0)) \tag{333}$$
$$= |\Sigma|^{-\frac{\upsilon_0+D+2+N}{2}}exp(-\frac{N}{2}(\mu - \bar{x})^T\Sigma^{-1}(\mu - \bar{x}) - \frac{\kappa_0}{2}(\mu - m_0)^T\Sigma^{-1}(\mu - m_0) \tag{334}$$
$$- \frac{1}{2}tr(\Sigma^{-1}S_0) - \frac{1}{2}tr(\Sigma^{-1}S_{\bar{x}})) \tag{335}$$
$$= |\Sigma|^{-\frac{\upsilon_N+D+2}{2}}exp(-\frac{\kappa_N}{2}(\mu - m_N)^T\Sigma^{-1}(\mu - m_N) - \frac{1}{2}tr(\Sigma^{-1}S_N)) \tag{336}$$
$$= NIW(\mu, \Sigma|m_N, \kappa_N, \upsilon_N, S_N) \tag{337}$$

Exercise 4.12. **a. Derive the BIC score for a Gaussian with dimension D will full covariance matrix.**

The BIC is given by

$$\begin{aligned}
BIC &= log\,p(D|\hat{\mu},\hat{\Sigma}) - \frac{d}{2}log(N) &(338)\\
&= -\frac{N}{2}tr(\hat{\Sigma}^{-1}\hat{S}) - \frac{N}{2}log(|\hat{\Sigma}|) - \frac{d}{2}log(N) &(339)\\
&= -\frac{N}{2}tr(\hat{\Sigma}^{-1}\hat{\Sigma}) - \frac{N}{2}log(|\hat{\Sigma}|) - \frac{d}{2}log(N) &(340)\\
&= -\frac{Nd}{2} - \frac{N}{2}log(|\hat{\Sigma}|) - \frac{d}{2}log(N) &(341)\\
&= -\frac{1}{2}(Nd + dlog(N) + Nlog(|\hat{\Sigma}|)) &(342)
\end{aligned}$$

b. Derive the BIC for a Gaussian with diagonal covariance matrix.

Note that for diagonal matrices, the determinant is just the product of the diagonals. Thus, we can reduce the above equation to

$$\begin{aligned}
BIC &= log\,p(D|\hat{\mu},\hat{\Sigma}) - \frac{d}{2}log(N) &(343)\\
&= -\frac{1}{2}(Nd + dlog(N) + Nlog(|\hat{\Sigma}|)) &(344)\\
&= -\frac{1}{2}(Nd + dlog(N) + Nlog(\prod_{i=1}^{d}\sigma_i)) &(345)\\
&= -\frac{1}{2}(Nd + dlog(N) + N\sum_{i=1}^{d}log(\sigma_i)) &(346)
\end{aligned}$$

Exercise 4.13. **Compute the sample size needed to compute the given Bayesian credible interval.**

From the text, we see that the posterior of the mean is given by

$$\begin{aligned}
p(\mu|D,\Sigma) &= N(\mu|m_N, V_N) &(347)\\
V_N^{-1} &= V_0^{-1} + N\Sigma^{-1} &(348)\\
m_N &= V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1}m_0) &(349)
\end{aligned}$$

Plugging these into what's given in the problem we get

$$\begin{aligned}
\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} &(350)\\
\frac{N}{\sigma^2} &= \frac{1}{\sigma_N^2} - \frac{1}{\sigma_0^2} &(351)\\
N &= \frac{\sigma^2}{\sigma_N^2} - \frac{\sigma^2}{\sigma_0^2} &(352)\\
&= \sigma^2\left(\frac{1}{\sigma_N^2} - \frac{1}{\sigma_0^2}\right) &(353)
\end{aligned}$$

Plugging in the known values, we see that

$$N = \sigma^2 \left(\frac{1}{\sigma_N^2} - \frac{1}{\sigma_0^2} \right) \tag{354}$$

$$= 4 \left(\frac{1}{\sigma_N^2} - \frac{1}{9} \right) \tag{355}$$

$$\tag{356}$$

Now all we need to know is σ_N^2 to compute this. For this, we know that the interval must be of width 1. We can use this to show that

$$u - l = 1 = (\mu_N + 1.96\sigma_N) - (\mu_N - 1.96\sigma_N) \tag{357}$$

$$1 + \mu_N - 1.96\sigma_N = \mu_N + 1.96\sigma_N \tag{358}$$

$$1 = 3.92\sigma_N \tag{359}$$

$$\sigma_N = \frac{1}{3.92} \tag{360}$$

$$\sigma_N^2 = \frac{1}{3.92^2} \tag{361}$$

Now we can plug everything in to get

$$N = 4 \left(\frac{1}{\sigma_N^2} - \frac{1}{9} \right) \tag{362}$$

$$= 4 \left(3.92^2 - \frac{1}{9} \right) \tag{363}$$

$$\approx 61.02 \tag{364}$$

For sample sizes we always round up, so the sample size is $N = 62$.

Exercise 4.14. **a. Calculate the MAP estimate of μ for a 1-d Gaussian.** The posterior is given by

$$p(\mu|D) \propto p(D|\mu)p(\mu) \tag{365}$$

$$= N(\mu, m_N, V_N) \tag{366}$$

$$V_N^{-1} = V_0^{-1} + N\Sigma^{-1} \tag{367}$$

$$m_N = V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1}m_0) \tag{368}$$

Because the mode is the mean of a Gaussian, we can compute the posterior mean of the distribution and this will be the MAP estimate.

$$p(\mu|D) = N(\mu, m_N, V_N) \tag{369}$$

$$m_N = V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1}m_0) \tag{370}$$

$$= V_N(\frac{N\bar{x}}{\sigma^2} + \frac{m}{s^2}) \tag{371}$$

$$V_N = \left(\frac{1}{s^2} + \frac{N}{\sigma^2}\right)^{-1} = \frac{s^2\sigma^2}{\sigma^2 + Ns^2} \tag{372}$$

$$m_N = \frac{s^2\sigma^2}{\sigma^2 + Ns^2}\left(\frac{N\bar{x}}{\sigma^2} + \frac{m}{s^2}\right) \tag{373}$$

$$= \frac{N\bar{x}s^2\sigma^2}{\sigma^2(\sigma^2 + Ns^2)} + \frac{s^2\sigma^2 m}{s^2(\sigma^2 + Ns^2)} \tag{374}$$

$$= \frac{N\bar{x}s^2}{\sigma^2 + Ns^2} + \frac{\sigma^2 m}{\sigma^2 + Ns^2} \tag{375}$$

$$= \frac{N\bar{x}s^2 + \sigma^2 m}{\sigma^2 + Ns^2} \tag{376}$$

This is the MAP estimate of the mean.

b. Show that as the number of samples n, the MAP estimate converges to the MLE.

$$\hat{\mu}_{MAP} = \frac{N\bar{x}s^2 + \sigma^2 m}{\sigma^2 + Ns^2} \tag{377}$$

$$\hat{\mu}_{MLE} = \bar{x} \tag{378}$$

So, we need to show that as n tends to infinity, the MAP solution converges to $\bar{x}$. Thus

$$\lim_{n\to\infty} \frac{N\bar{x}s^2 + \sigma^2 m}{\sigma^2 + Ns^2} = \frac{N\bar{x}s^2}{Ns^2} = \frac{Ns^2}{Ns^2}\bar{x} = \bar{x} \tag{379}$$

c. Suppose n is small and fixed. What does the MAP estimate converge to if we increase the prior variance s^2?

This is similar to the previous section, but now we are taking the limit of s^2 to ∞.

$$\lim_{s\to\infty} \frac{N\bar{x}s^2 + \sigma^2 m}{\sigma^2 + Ns^2} = \frac{Ns^2\bar{x}}{Ns^2} = \frac{Ns^2}{Ns^2}\bar{x} = \bar{x} \tag{380}$$

So, increasing the prior variance to ∞ yields the MLE. This makes sense, since the prior with infinite variance is an uninformative prior.

c. Suppose n is small and fixed. What does the MAP estimate converge to if we decrease the prior variance s^2?

This is similar to the previous section, but now we are taking the limit of s^2 to 0.

$$\lim_{s\to 0} \frac{N\bar{x}s^2 + \sigma^2 m}{\sigma^2 + Ns^2} = \frac{\sigma^2 m}{\sigma^2} = \frac{\sigma^2}{\sigma^2}m = m \tag{381}$$

So, as the prior variance tends to 0, the MAP converges to the prior mean m. This makes sense, because a prior with a variance of 0 encodes the belief that we are absolutely certain of the prior mean.

Exercise 4.15. **a. Show how to sequentially update the covariance estimate.**

The sample covariance is given by

$$\hat{\Sigma} = C_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_n)(x_i - m_n)^T$$

Note that we can update the cumulative mean as

$$m_{n+1} = \frac{x_{n+1} + nm_n}{n+1}$$

What we are trying to show is that

$$C_{n+1} = \frac{n-1}{n}C_n + \frac{1}{n+1}(x_{n+1} - m_n)(x_{n+1} - m_n)^T \tag{382}$$

$$nC_{n+1} = (n-1)C_n + \frac{n}{n+1}(x_{n+1} - m_n)(x_{n+1} - m_n)^T \tag{383}$$

$$nC_{n+1} - (n-1)C_n = \frac{n}{n+1}(x_{n+1} - m_n)(x_{n+1} - m_n)^T \tag{384}$$

This form is a little easier to work with. Using this definition and the following definitions:

$$C_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_n)(x_i - m_n)^T \tag{385}$$

$$C_{n+1} = \frac{1}{n}\sum_{i=1}^{n+1}(x_i - m_{n+1})(x_i - m_{n+1})^T \tag{386}$$

We can show that

$$nC_{n+1} - (n-1)C_n = \sum_{i=1}^{n+1}(x_i - m_{n+1})(x_i - m_{n+1})^T - \sum_{i=1}^{n}(x_i - m_n)(x_i - m_n)^T \tag{387}$$

$$= \sum_{i=1}^{n+1} x_i x_i^T - m_{n+1}m_{n+1}^T - \sum_{i=1}^{n} x_i x_i^T - m_n m_n^T \tag{388}$$

$$= x_{n+1}x_{n+1}^T - (n+1)m_{n+1}m_{n+1}^T - nm_n m_n^T \tag{389}$$

$$= x_{n+1}x_{n+1}^T - nm_n m_n^T - (n+1)\left(\frac{x_{n+1} + nm_n}{n+1}\right)\left(\frac{x_{n+1} + nm_n}{n+1}\right)^T \tag{390}$$

$$= x_{n+1}x_{n+1}^T - nm_n m_n^T - \frac{1}{n+1}(x_{n+1} + nm_n)(x_{n+1} + nm_n)^T \tag{391}$$

$$= x_{n+1}x_{n+1}^T - nm_n m_n^T - \frac{1}{n+1}(x_{n+1}x_{n+1}^T + nx_{n+1}m_n + nm_n x_{n+1} + n^2 m_n m_n^T) \tag{392}$$

$$= \frac{n}{n+1}x_{n+1}x_{n+1}^T - nm_n m_n^T - \frac{n^2}{n+1}m_n m_n^T - \frac{n}{n+1}x_{n+1}m_n - \frac{n}{n+1}m_n x_{n+1} \tag{393}$$

$$= \frac{n}{n+1}x_{n+1}x_{n+1}^T - \frac{n(n+1)}{n+1}m_n m_n^T - \frac{n^2}{n+1}m_n m_n^T \tag{394}$$

$$= \frac{n}{n+1}x_{n+1}x_{n+1}^T - \frac{n}{n+1}m_n m_n^T \tag{395}$$

$$= \frac{n}{n+1}(x_{n+1} - m_n)(x_{n+1} - m_n)^T \tag{396}$$

which is what we intended to show.

b. What is the big-O run time of this sequential update?

This procedure is $O(d^2)$, because we only have to compute one inner product at a time.

c. Show how to incrementally update the precision matrix.

Let $u = (x_{n+1} - m_n)$. Then

$$C_{n+1}^{-1} = \left(\frac{n-1}{n}C_n + \frac{1}{n+1}uu^T\right)^{-1} \tag{397}$$

and we are trying to show that

$$C_{n+1}^{-1} = \frac{n}{n-1}\left[C_n^{-1} - \frac{C_n^{-1}uu^T C_n^{-1}}{\frac{n^2-1}{n} + u^T C_n^{-1} u}\right] \tag{398}$$

Using the matrix inversion lemma provides us with

$$C_{n+1}^{-1} = \left(\frac{n-1}{n}C_n + \frac{1}{n+1}uu^T\right)^{-1} \tag{399}$$

$$= \frac{n}{n-1}C_n^{-1} - \frac{\frac{n}{n-1}C_n^{-1}\frac{1}{n+1}uu^T\frac{n}{n-1}C_n^{-1}}{1+\frac{1}{n+1}u^T\frac{n}{n-1}C_n^{-1}u} \tag{400}$$

$$= \frac{n}{n-1}C_n^{-1} - \frac{\frac{n^2}{(n-1)^2(n+1)}C_n^{-1}uu^TC_n^{-1}}{1+\frac{n}{(n-1)(n+1)}u^TC_n^{-1}u} \tag{401}$$

$$\tag{402}$$

Note that

$$\frac{n^2}{(n-1)^2(n+1)} = \frac{n}{n-1}\frac{n}{(n-1)(n+1)}$$

and

$$(n-1)(n+1)(1+\frac{n}{(n-1)(n+1)}B) = (n-1)(n+1) + nB = n^2 - 1 + nB$$

Using these we see that

$$C_{n+1}^{-1} = \frac{n}{n-1}C_n^{-1} - \frac{\frac{n^2}{(n-1)^2(n+1)}C_n^{-1}uu^TC_n^{-1}}{1+\frac{n}{(n-1)(n+1)}u^TC_n^{-1}u} \tag{403}$$

$$= \frac{n}{n-1}\left[C_n^{-1} - \frac{nC_n^{-1}uu^TC_n^{-1}}{n^2-1+nu^TC_n^{-1}u}\right] \tag{404}$$

$$= \frac{n}{n-1}\left[C_n^{-1} - \frac{C_n^{-1}uu^TC_n^{-1}}{\frac{n^2-1}{n}u^TC_n^{-1}u}\right] \tag{405}$$

d. What's the big-O complexity of this procedure?
This procedure is also $O(d^2)$.

Exercise 4.16. **Derive an expression for the log likelihood ratio with an arbitrary covariance matrix.**

$$log\frac{p(y=1|x)}{p(y=0|x)} = log\frac{p(x|y=1)}{p(x|y=0)} + log\frac{p(y=1)}{p(y=0)} \tag{406}$$

$$= log\frac{N(x|\mu_1,\Sigma_1)}{N(x|\mu_2,\Sigma_2)} + log\frac{p(y=1)}{p(y=0)} \tag{407}$$

$$= log\frac{(2\pi)^{D/2}|\Sigma_1|^{1/2}exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}{(2\pi)^{D/2}|\Sigma_0|^{1/2}exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)} + log\frac{p(y=1)}{p(y=0)} \tag{408}$$

$$= log\frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}}exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)) + log\frac{p(y=1)}{p(y=0)} \tag{409}$$

$$= log\frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + log\frac{p(y=1)}{p(y=0)} \tag{410}$$

We can simplify this further if we make assumptions about the problem. For example, if the covariance matrix is shared ($\Sigma_j = \Sigma$), then

$$log\frac{p(y=1|x)}{p(y=0|x)} = log\frac{|\Sigma|^{1/2}}{|\Sigma|^{1/2}} - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + log\frac{p(y=1)}{p(y=0)} \tag{411}$$

$$= 1 + log\frac{p(y=1)}{p(y=0)} - \frac{1}{2}\left[tr((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)) + tr((x-\mu_0)^T\Sigma^{-1}(x-\mu_0))\right] \tag{412}$$

$$= 1 + log\frac{p(y=1)}{p(y=0)} - \frac{1}{2}\left[tr((x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + (x-\mu_0)^T\Sigma^{-1}(x-\mu_0))\right] \tag{413}$$

$$\tag{414}$$

Further, if the covariance matrix is shared and diagonal, then

$$log\frac{p(y=1|x)}{p(y=0|x)} = 1 + log\frac{p(y=1)}{p(y=0)} - \frac{1}{2}\left[tr((x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + (x-\mu_0)^T\Sigma^{-1}(x-\mu_0))\right] \tag{415}$$

$$= 1 + log\frac{p(y=1)}{p(y=0)} - \frac{1}{2}\left[\sum_{i=1}^{N}\frac{(x_i-\mu_1)^2}{\sigma_i} + \frac{(x_i-\mu_0)^2}{\sigma_i}\right] \tag{416}$$

Finally, if the covariance is shared and spherical ($\Sigma = \sigma^2 I$), then

$$log\frac{p(y=1|x)}{p(y=0|x)} = 1 + log\frac{p(y=1)}{p(y=0)} - \frac{1}{2}\left[\sum_{i=1}^{N}\frac{(x_i-\mu_1)^2}{\sigma} + \frac{(x_i-\mu_0)^2}{\sigma}\right] \tag{417}$$

$$= 1 + log\frac{p(y=1)}{p(y=0)} - \frac{N}{2\sigma}\sum_{i=1}^{N}(x_i-\mu_1)^2 + (x_i-\mu_0)^2 \tag{418}$$

Exercise 4.17. **Compute the misclassification rate of LDA and QDA on the height/weight dataset.**

The code for this can be found in ch4-17.ipynb.

Exercise 4.18. **Consider a 3 class naive Bayes classifier with one binary feature and one Gaussian feature.**
a. Compute $p(y|x_1 = 0, x_2 = 0)$.

The naive Bayes classifier can be written as

$$p(y|x_1=0,x_2=0) = p(y)p(x_1=0|y)p(x_2=0|y) \tag{419}$$

$$= Mu(y|\pi,1)Ber(x_1=0|\theta)N(x_2=0|\mu,\sigma^2) \tag{420}$$

$$= Mu(y,\begin{bmatrix}0.5\\0.25\\0.25\end{bmatrix},1)Ber(x_1=0|\begin{bmatrix}0.5\\0.5\\0.5\end{bmatrix})N(x_2=0|\begin{bmatrix}-1\\0\\1\end{bmatrix},\begin{bmatrix}1\\1\\1\end{bmatrix}) \tag{421}$$

$$\propto \begin{bmatrix}0.5\\0.25\\0.25\end{bmatrix} \oplus \begin{bmatrix}0.5\\0.5\\0.5\end{bmatrix} \oplus \begin{bmatrix}0.2419707\\0.3989423\\0.2419707\end{bmatrix} \tag{422}$$

$$= \begin{bmatrix}0.060492675\\0.0498677875\\0.0302463375\end{bmatrix} \tag{423}$$

$$= \begin{bmatrix}0.4302258233\\0.3546612864\\0.2151129116\end{bmatrix} \tag{424}$$

b. Compute $p(y|x_1 = 0)$.

We can compute this by noting that

$$p(y|x_1=0) = \sum_{i=1}^{3} p(y=i)p(x_1=0|y=i) \tag{425}$$

$$= \sum_{i=1}^{3} \pi_i Ber(x_1=0|\theta_i) \tag{426}$$

$$= 0.5\sum_{i=1}^{3}\pi_i = 0.5 \tag{427}$$

c. Compute $p(y|x_2 = 0)$.
We can compute this in a similar fashion:

$$
\begin{aligned}
p(y|x_2 = 0) &= \sum_{i=1}^{3} \pi_i p(x_2 = 0|y = i) && (428)\\
&= \sum_{i=1}^{3} \pi_i N(\mu_i, \sigma_i^2) && (429)\\
&= 0.5 \times 0.2419707 + 0.25 \times 0.3989423 + 0.25 \times 0.2419707 && (430)\\
&= 0.341706275 && (431)
\end{aligned}
$$

Exercise 4.19. **Derive the QDA decision boundary for a binary classification problem where $\Sigma_1 = k\Sigma_0$.**
The QDA formulation is given by

$$p(y = c|x, \theta) = \frac{\pi_c|2\pi\Sigma_c|^{1/2} exp(-\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1}(x - \mu_c))}{\sum_c \pi_c|2\pi\Sigma_c|^{1/2} exp(-\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1}(x - \mu_c))} \tag{432}$$

Assume that Σ is of dimensionality D. Formulating this as a binary problem and plugging in the fact that $\Sigma_1 = k\Sigma_0$, we get

$$
\begin{aligned}
p(y = 1|x, \theta) &= \frac{\pi_1|2\pi k\Sigma_0|^{1/2} exp(-\frac{1}{2}(x - \mu_1)^T k\Sigma_0^{-1}(x - \mu_1))}{\pi_0|2\pi\Sigma_0|^{1/2} exp(-\frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1}(x - \mu_0)) + \pi_1|2\pi k\Sigma_0|^{1/2} exp(-\frac{1}{2}(x - \mu_1)^T k\Sigma_0^{-1}(x - \mu_1))} && (433)\\
&= \frac{\pi_1 k^D exp(-\frac{1}{2}(x - \mu_1)^T k\Sigma_0^{-1}(x - \mu_1))}{\pi_0 exp(-\frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1}(x - \mu_0)) + \pi_1 k^D exp(-\frac{1}{2}(x - \mu_1)^T k\Sigma_0^{-1}(x - \mu_1))} && (434)\\
&= \frac{\pi_1 k^D exp(-\frac{k}{2} tr(\Sigma_0^{-1}(x - \mu_1)(x - \mu_1)^T)}{\pi_0 exp(-\frac{1}{2} tr(\Sigma_0^{-1}(x - \mu_0)(x - \mu_0)^T) + \pi_1 k^D exp(-\frac{k}{2} tr(\Sigma_0^{-1}(x - \mu_1)(x - \mu_1)^T)} && (435)\\
&= \frac{\pi_1 k^D exp(-\frac{k}{2} tr(\Sigma_0^{-1} xx^T) - \frac{k}{2} tr(\Sigma_0^{-1}\mu_1\mu_1^T))}{\pi_0 exp(-\frac{1}{2} tr(\Sigma_0^{-1} xx^T) - \frac{1}{2} tr(\Sigma_0^{-1}\mu_0\mu_0^T)) + \pi_1 k^D exp(-\frac{k}{2} tr(\Sigma_0^{-1} xx^T) - \frac{k}{2} tr(\Sigma_0^{-1}\mu_1\mu_1^T))} && (436)
\end{aligned}
$$

Let $a = tr(\Sigma_0^{-1} xx^T)$ and $b_c = tr(\Sigma_0^{-1}\mu_c\mu_c^T)$. Then

$$
\begin{align}
p(y=1|x,\theta) &= \frac{\pi_1 k^D \frac{exp(-\frac{k}{2}a)}{exp(-\frac{k}{2}b_1)}}{\pi_0 \frac{exp(-\frac{1}{2}a)}{exp(-\frac{1}{2}b_0)} + \pi_1 k^D \frac{exp(-\frac{k}{2}a)}{exp(-\frac{k}{2}b_1)}} \tag{437}\\
&= \frac{\pi_1 k^D exp(-\frac{k}{2}a)}{\pi_0 exp(-\frac{1}{2}a - \frac{k}{2}b_1 + \frac{1}{2}b_0) + \pi_1 k^D exp(-\frac{k}{2}a)} \tag{438}\\
&= \frac{1}{\pi_0 \pi_1^{-1} k^{-D} exp(\frac{k}{2}a - \frac{1}{2}a - \frac{k}{2}b_1 + \frac{1}{2}b_0) + 1} \tag{439}\\
&= \frac{1}{\pi_0 \pi_1^{-1} k^{-D} exp(\frac{1}{2}((k-1)a - kb_1 + b_0)) + 1} \tag{440}\\
&= \pi_0^{-1}\pi_1 k^D Sigm(\nu) \tag{441}
\end{align}
$$

where

$$
\begin{align}
\nu &= -\frac{1}{2}((k-1)a + kb_1 + b_0) \tag{442}\\
&= -\frac{1}{2}((k-1)tr(\Sigma_0^{-1}xx^T) + ktr(\Sigma_0^{-1}\mu_1\mu_1^T) + tr(\Sigma_0^{-1}\mu_0\mu_0^T)) \tag{443}\\
&= -\frac{1}{2}((k-1)tr(\Sigma_0^{-1}xx^T) + (k-1)tr(\Sigma_0^{-1}\mu_1\mu_1^T) + tr(\Sigma_0^{-1}\mu_1\mu_1^T) + tr(\Sigma_0^{-1}\mu_0\mu_0^T)) \tag{444}\\
&= -\frac{1}{2}((k-1)(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1) + tr(\Sigma_0^{-1}\mu_1\mu_1^T) + tr(\Sigma_0^{-1}\mu_0\mu_0^T)) \tag{445}\\
&= -\frac{1}{2}((k-1)(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1) + (\mu_1-\mu_0)^T\Sigma_0^{-1}(\mu_1-\mu_0)) \tag{446}\\
&= -\frac{k-1}{2}(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1) - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma_0^{-1}(\mu_1-\mu_0) \tag{447}
\end{align}
$$

Thus, since class 1 is a scaled version of class 2, the decision boundary scales this class as well.

Exercise 4.20. **See the textbook for the full problem description.**
a. GaussI, LinLog.

Note that we are only considering performance on the training set, and only considering a loss function that is a function of the conditional likelihood, not the joint likelihood.

Logistic regression maximizes the conditional likelihood, whereas LDA and QDA maximize the joint likelihood. Since we are only considering training set performance, maximizing the conditional likelihood will be sufficient in minimizing the given loss function. Thus

$L(GaussI) \geq L(LinLog)$.

b. GaussX, QuadLog.

Again, on the surface they seem equivalent, however the logistic model maximizes the conditional likelihood. So similarly

$L(GaussX) \geq L(QuadLog)$.

c. LinLog, QuadLog.

The QuadLog model has more parameters and is therefore more flexible. It might not perform as well on the test set, but on the training set it will likely perform better. Thus

$L(LinLog) \geq L(QuadLog)$.

d. GaussI, QuadLog.

The GaussI model is likely too restrictive, thus

$L(GaussI) \geq L(QuadLog)$.

e. In general is it true that the negative log likelihood loss function behaves similarly to the misclassification rate ($L(M) > L(M')$ implies $R(M) < R(M')$)?

This is not true because of the discretized nature of the misclassification rate. For example, one model could lower the log likelihood loss but still not be better "enough" to improve the misclassification rate.

Exercise 4.21. **TODO**

Exercise 4.22. **Class the points using the QDA model described in the text.**

a. $x = [-0.5, 0.5]$**.**

Note that the normalization constants for the problem described is given by

$$Z_c = \pi_c |2\pi\Sigma_c|^{-1/2} \tag{448}$$

$$Z_1 = \frac{1}{3}|2\pi 0.7I|^{-1/2} = 0.2273643491 \tag{449}$$

$$Z_2 = Z_3 = \frac{1}{3}|2\pi \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}|^{-1/2} = 0.2054679336 \tag{450}$$

Another quantity that is useful to compute up front is

$$M_c = (x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c) \tag{451}$$

$$M_1 = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}^T \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}^{-1} \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} = 0.714286 \tag{452}$$

$$M_2 = \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix}^T \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}^{-1} \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} = 2.83333 \tag{453}$$

$$M_3 = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}^T \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}^{-1} \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} = 0.833333 \tag{454}$$

A final quantity that will use is

$$P_c = Z_c exp(-\frac{1}{2}M_c) \tag{455}$$
$$P_1 = Z_1 exp(-\frac{1}{2}M_1) = 0.1590805683 \tag{456}$$
$$P_2 = Z_2 exp(-\frac{1}{2}M_2) = 0.0498303871 \tag{457}$$
$$P_3 = Z_3 exp(-\frac{1}{2}M_3) = 0.1354530358 \tag{458}$$

We can then plug these into the equation for QDA to get

$$p(y=1|x,\theta) = \frac{P_1}{P_1+P_2+P_3} = 0.4619547118 \tag{459}$$
$$p(y=2|x,\theta) = \frac{P_2}{P_1+P_2+P_3} = 0.1447026645 \tag{460}$$
$$p(y=3|x,\theta) = \frac{P_3}{P_1+P_2+P_3} = 0.3933426237 \tag{461}$$

Thus, we would classify this point to class 1.
b. Classify $x = [0.5, 0.5]$.
We can utilize the same machinery for this.

$$M_c = (x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c) \tag{462}$$
$$M_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}^T \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}^{-1} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = 0.714286 \tag{463}$$
$$M_2 = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}^T \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}^{-1} \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} = 0.5 \tag{464}$$
$$M_3 = \begin{bmatrix} -0.5 \\ 1.5 \end{bmatrix}^T \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}^{-1} \begin{bmatrix} -0.5 \\ 1.5 \end{bmatrix} = 3.83333 \tag{465}$$

And

$$P_c = Z_c exp(-\frac{1}{2}M_c) \tag{466}$$
$$P_1 = Z_1 exp(-\frac{1}{2}M_1) = 0.1590805683 \tag{467}$$
$$P_2 = Z_2 exp(-\frac{1}{2}M_2) = 0.1600185876 \tag{468}$$
$$P_3 = Z_3 exp(-\frac{1}{2}M_3) = 0.03022365758 \tag{469}$$

We can then plug these into the equation for QDA to get

$$p(y = 1|x, \theta) = \frac{P_1}{P_1 + P_2 + P_3} = \tag{470}$$

$$p(y = 2|x, \theta) = \frac{P_2}{P_1 + P_2 + P_3} = \tag{471}$$

$$p(y = 3|x, \theta) = \frac{P_3}{P_1 + P_2 + P_3} = \tag{472}$$