6. Probability and Distributions (Exercises Only)

Exercises

6.1

- Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y:

Y \ X	x_1	x_2	x_3	x_4	x_5
y_1	0.01	0.02	0.03	0.1	0.1
y_2	0.05	0.1	0.05	0.07	0.2
y_3	0.1	0.05	0.03	0.05	0.04

- Compute:
  - a. the marginal distributions $p(x)$ and $p(y)$
- $p(x)$:

$$P(X = x_1) = 0.16$$

$$P(X = x_2) = 0.17$$

$$P(X = x_3) = 0.11$$

$$P(X = x_4) = 0.22$$

$$P(X = x_5) = 0.34$$

- $p(y)$:

$$P(Y = y_1) = 0.26$$

$$P(Y = y_2) = 0.47$$

$$P(Y = y_3) = 0.27$$

- b. The conditional distributions $p(x|Y = y_1)$ and $p(y|X = x_3)$:

- $p(x|Y=y_1)$:

$$P(X=x_1|Y=y_1)=\frac{0.01}{0.26}$$

$$P(X=x_2|Y=y_1)=\frac{0.02}{0.26}$$

$$P(X=x_3|Y=y_1)=\frac{0.03}{0.26}$$

$$P(X=x_4|Y=y_1)=\frac{0.1}{0.26}$$

$$P(X=x_5|Y=y_1)=\frac{0.1}{0.26}$$

- $p(y|X=x_3)$:

$$P(Y=y_1|X=x_3)=\frac{0.03}{0.11}$$

$$P(Y=y_2|X=x_3)=\frac{0.05}{0.11}$$

$$P(Y=y_3|X=x_3)=\frac{0.03}{0.11}$$

6.2

- Consider the mixture of two Gaussian distributions:

$$0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)+0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}\right)$$

- a. Compute the marginal distribution for each dimension:

$$p(x_1)=0.4\mathcal{N}(10,1)+0.6\mathcal{N}(0,8.4)$$

$$p(x_2)=0.4\mathcal{N}(2,1)+0.6\mathcal{N}(0,1.7)$$

- b. Compute the mean, mode, and median for each marginal distribution:

$$\mu_{x_1}=0.4(10)=4$$

$$\mu_{x2}=0.4(2)=0.8$$

$$\text{mode}_{x_1}=4$$

$$\text{mode}_{x_2}=0.8$$

$$\text{median}_{x_1}=4$$

$$\text{median}_{x_2}=0.8$$

- c. Compute the mean and mode for the two-dimensional distribution:

$$\mu = \begin{bmatrix} 4 \\ 0.8 \end{bmatrix}$$

$$\text{median} = \mu$$

6.3

Bernoulli distribution:

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}, \qquad x \in \{0, 1\}$$

- Choose a conjugate prior for the Bernoulli likelihood and compute the posterior distribution $p(\mu|x_1, \dots, x_n)$
- prior: Beta Distribution:

$$p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- We now compute the posterior according to Bayes' Theorem:

$$p(\mu|x_1, \dots, x_n) \propto p(x|\mu)p(\mu|\alpha, \beta)$$

$$\begin{aligned} p(\mu|x_1, \dots, x_n) &\propto \mu^x (1-\mu)^{1-x} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &\propto \mu^{\alpha-1+x}(1-\mu)^{\beta+(1-x)-1} \end{aligned}$$

- The posterior is itself a beta distribution proportional to:

$$p(\mu|\alpha + x, \beta + (1-x))$$

- Therefore, the Beta Distribution is indeed a conjugate prior to the Bernoulli likelihood.

6.4

- Bag 1:
  - $p(\text{mango}) = 2/3$
  - $p(\text{apple}) = 1/3$
- Bag 2:
  - $p(\text{mango}) = 1/2$
  - $p(\text{apple}) = 1/2$
- $p(\text{heads}) = 0.6$
- $p(\text{tails}) = 0.4$

$$p(\text{tails}|\text{fruite} = \text{mango}) = \frac{p(\text{mango}|\text{tails})}{p(\text{mango})} = \frac{1/2(0.4)}{2/3(0.6) + 1/2(0.4)} = \frac{0.2}{0.6} = \frac{1}{3}$$

6.5

- Consider the time-series model

$$x_{t+1} = Ax_1 + w, \qquad w \sim \mathcal{N}(0, Q)$$

$$y_t = Cx_t + v, \qquad v \sim \mathcal{N}(0, R)$$

- w and v are i.i.d Gaussian noise variables
- $p(x_0) = \mathcal{N}(\mu_0, \Sigma_0)$
- a. What is the form of $p(x_t|x_1, \ldots, x_T)$
  - $p(x_t|x_1, \ldots, x_T)$ is *Gaussian* because the linear combination of Gaussian densities is Gaussian
- b. Assume that $p(x_{t+1}|y_1, \ldots, y_t) = \mathcal{N}(\mu_t, \Sigma_t)$

1. Compute $p(x_{t+1}|y_1, \ldots, y_t)$

$$\mathbb{E}[x_{t+1}] = \mu_{t+1} = A\mu_t$$

$$\mathbb{V}[x_{t+1}] = \Sigma_{t+1} = \mathbb{V}[Ax_t + w] = A\mathbb{V}[x_t] + \mathbb{V}[w] = A\Sigma_t A^T + Q$$

```
- Therefore:
```

$$p(x_{t+1}|y_1, \ldots y_t) = \mathcal{N}(A\mu_t, A\Sigma_t A^T + Q)$$

```
2. Compute $p(x_{t + 1}, y_{t + 1}|y_1, ..., y_t)$:
```

- First we need to compute: $p(y_{t+1}|y_1, \ldots, y_t)$

$$p(y_{t+1}|x_{t+1}y_1, \ldots, y_t) = \mathcal{N}(Cx_{t+1}, R)$$

$$p(x_{t+1}, y_{t+1}|y_1, \ldots, y_t) = p(x_{t+1}|y_1, \ldots y_t)p(y_{t+1}|x_{t+1}, y_1, \ldots, y_t)$$

$$p(x_{t+1}, y_{t+1}|y_1, \ldots, y_t) = \mathcal{N}(A\mu_t, A\Sigma_t A^T + Q)\mathcal{N}(Cx_{t+1}, R)$$

3. At time $t + 1$ we observe the value $y_{t+1} = \hat{y}$. Compute the conditional distribution $p(x_{t+1}|y_1, \ldots, y_{t+1})$

$$p(x_{t+1}|y_1, \ldots, y_t, y_{t+1}) = \frac{p(x_{t+1}, y_{t+1}|y_1, \ldots, y_t)}{p(y_{t+1}|x_{t+1}, y_1, \ldots, y_t)}$$

$$p(x_{t+1}|y_1, \ldots, y_t, y_{t+1}) = \frac{\mathcal{N}(A\mu_t, A\Sigma_t A^T + Q)\mathcal{N}(Cx_{t+1}, R)}{\mathcal{N}(CA\mu_t, C(A\Sigma_t A^T + Q)C^T + R)}$$

6.6

- Prove:

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$$

- Start with:

$$\mathbb{V}_X[x] := \mathbb{E}[(x-\mu)^2]$$

$$\begin{aligned}
\mathbb{V}_X[x] &= \mathbb{E}[x^2 = 2\mu x + \mu^2] \\
&= \mathbb{E}[x^2] - 2\mathbb{E}[\mu x] + \mathbb{E}[\mu^2] \\
&= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \\
&= \mathbb{E}[x^2] - 2\mathbb{E}[x]\mathbb{E}[x] + (\mathbb{E}[\mathbb{x}])^2 \\
&= \mathbb{E}[x^2] - 2(\mathbb{E}[x])^2 + (\mathbb{E}[\mathbb{x}])^2 \\
&= \mathbb{E}[x^2] - (\mathbb{E}[x])^2
\end{aligned}$$ ✅

6.7

It turns out we can avoid two passes by rearranging terms:

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_{\mathbb{X}}[\mathbb{x}])^2$$

- This can be computed in one pass since we accumulate x_i (to calculate the mean) and x_i^2 simultaneously, where x_i is the ith observation in the data
- Unfortunately this implementation has numerical stability issues
- Another way of understanding variance is as the sum of pairwise differences between all pairs of observations.
- We can compute the squared difference between pairs x_i and x_j

$$\sum_{i,j=1}^{N} (x_i - x_j)^2$$

- expanding out the square:

$$\sum_{i,j=1}^{N} (x_i^2 - 2x_i x_j + x_j^2)$$

$$\sum_{i,j=1}^{N} x_i^2 - 2\sum_{i,j=1}^{N} x_i x_j + \sum_{i,j=1}^{N} x_j^2$$

- since the last term is also just a sum over the square of all xs in the data set we can rewrite it to be the same as the first term:

$$2\sum_{i,j=1}^{N} x_i^2 - 2\sum_{i,j=1}^{N} x_i x_j$$

- The double sum can be distributed to the product of sums:

$$2\sum_{i,j=1}^{N} x_i^2 - 2(\sum_{i=1}^{N} x_i)(\sum_{j=1}^{N} x_j)$$

- Once again, because we're summing over all of the xs in the data set, x_j in the last term can just be written as x_i allowing us to combine the last two terms into:

$$2\sum_{i=1}^{N} x_i^2 - 2(\sum_{i=1}^{N} x_i)^2$$

- Now we can factor out a 2:

$$2\left[\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2\right]$$

- Now if you compare this to the second equation for variance we introduced:

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_{\mathbb{X}}[\mathbb{x}])^2$$

- you will see that these look extremely similar save for two differences:
  1. The equation above is the expectation instead of the sum
  2. The equation above is not multiplied by 2
- So all we need to do to equate this sum of squared differences expression to the variance is divide to divide each sum by N and divide the entire expression by 2:

$$\mathbb{E}_X[x^2] - (\mathbb{E}_{\mathbb{X}}[\mathbb{x}])^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2$$

- This is equivalent to dividing the original expression by $2N^2$

6.8

- Express the Bernoulli distribution in the natural parameter form of the exponential family:
- Bernoulli:

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- We want this to be in the form:

$$p(x) = h(x)\exp(\theta^T\phi(x) - A(\theta))$$

- We start by taking the $\exp(\log(\cdot))$ of the Bernoulli distribution:

$$\exp(\log(\mu^x(1-\mu)^{1-x}))$$

- The log of products is equal to the sum of logs:

$$\exp(\log(\mu^x) + \log((1-\mu)^{1-x}))$$
$$\exp(x\log(\mu) + (1-x)\log(1-\mu))$$

$$\exp(x\log(\mu) - x\log(1-\mu) + \log(1-\mu))$$

$$\exp(x\log(\frac{\mu}{1-\mu}) + \log(1-\mu))$$

- This is now in the natural parameter form of the exponential family with:
  - $h(x) = 1$
  - $\theta = \log\left(\frac{\mu}{1-\mu}\right)$
  - $\phi(x) = x$
- To find get the log partition function $A(\theta)$ in terms of θ we first need to find μ in terms of θ

$$\begin{aligned}
\theta &= \log\left(\frac{\mu}{1-\mu}\right) \\
\exp(\theta) &= \frac{\mu}{1-\mu} \\
\exp(\theta)(1-\mu) &= \mu \\
\exp(\theta) - \mu\exp(\theta) &= \mu \\
\exp(\theta) &= \mu + \mu\exp(\theta) \\
\exp(\theta) &= \mu(1+\exp(\theta)) \\
\mu &= \frac{\exp(\theta)}{1+\exp(\theta)}
\end{aligned}$$

- Now we rewrite:

$$\log(1-\mu)$$

- in terms of θ to get our log partition function:

$$\begin{aligned}
A(\theta) &= -\log\left(1 - \frac{\exp(\theta)}{1-\exp(\theta)}\right) \\
&= -\log\left(\frac{1-\exp(\theta)}{1-\exp(\theta)} - \frac{\exp(\theta)}{1-\exp(\theta)}\right) \\
&= -\log\left(\frac{1}{1+\exp(\theta)}\right) \\
&= \log(1+\exp(\theta))
\end{aligned}$$

6.9

- Express the Binomial Distribution as an exponential family distribution:

$$p(m) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

- Take the $\exp(\log(\cdot))$"

$$\binom{N}{m}\exp(\log(\mu^m(1-\mu)^{N-m}))$$

$$\binom{N}{m}\exp(\log(\mu^m) + \log((1-\mu)^{N-m}))$$

$$\binom{N}{m}\exp(m\log(\mu) + (N-m)\log(1-\mu))$$

$$\binom{N}{m}\exp(m\log(\mu) - m\log(1-\mu) + N\log(1-\mu))$$

$$\binom{N}{m}\exp\left(m\log\left(\frac{\mu}{1-\mu}\right) + N\log(1-\mu)\right)$$

- This is exponential family form:
- $\theta = \log(\frac{\mu}{1-\mu})$
- $\phi(m) = m$
- $h(m) = \binom{N}{m}$
- $A(\theta) = N\log(1+\exp(\theta))$
- Express the Beta Distribution as an exponential family distribution:

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

$$\exp(\log(\mu^{\alpha-1}(1-\mu)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}))$$

$$\exp(\log(\mu^{\alpha-1}) + \log((1-\mu)^{\beta-1}) + \log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}))$$

$$\exp((\alpha-1)\log(\mu) + (\beta-1)\log(1-\mu) + \log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}))$$

- $\theta = \begin{bmatrix}\alpha-1\\ \beta-1\end{bmatrix}$
- $\phi(x) = \begin{bmatrix}\log(\mu)\\ \log(1-\mu)\end{bmatrix}$
- $h(x) = 1$
- $A(\theta) = \$ - \log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}) = -[\log(\Gamma(\alpha+\beta) - \log(\Gamma(\alpha)\Gamma(\beta))] = \log(\Gamma(\alpha+\beta) + \log(\Gamma(\alpha)) + \log(\Gamma$
- Now we find the product of these two distributions:

$$\binom{N}{m}\exp\left(m\log\left(\frac{\mu}{1-\mu}\right) + N\log(1-\mu)\right) \times \exp((\alpha-1)\log(m) + (\beta-1)\log(1-m) + \log(\Gamma(\theta_1+1)$$

$$\binom{N}{m}\exp\left(m\log\left(\frac{\mu}{1-\mu}\right) + (\alpha-1)\log(m) + (\beta-1)\log(1-m) + N\log(1-\mu) + \log(\Gamma(\theta_1+1) + (\theta_2\right.$$

6.10

6.11

- Consider the two random variables x and y with joint distribution $p(x, y)$. Show that

$$\mathbb{E}_X[x] = \mathbb{E}_Y[\mathbb{E}_X[x|y]]$$

- First let's write out what the joint probability is in terms of the conditional probability $p(x|y)$:

$$p(x, y) = p(x|y)p(y)$$

$$\mathbb{E}_X[x|y] = \int xp(x|y)dx$$

- Now we write

$$\mathbb{E}_Y[f(y)] = \int f(y)p(y)dy$$

- Where $f(y) = \mathbb{E}_X[x|y]$
- Plugging the formula for the expectation of the conditional distribution in for $f(y)$ gives us:

$$\int \in xp(x|y)dxp(y)dx = \int \int xp(x|y)p(y)dxdy = \int \int xp(x, y)dxdy$$

- Which is $\mathbb{E}_X[x]$ over the joint distribution $p(x, y)$