MVE137 Exam Solutions

Probability and Statistical Learning Using Python

25 August 2022

Part I

1. (a) The RVs X_1 and X_2 are discrete uniform variables over the set $\{1, \cdots, k\}$ with probability $p = 1/k$ to assume each value. When we roll two dice any specific outcome $(X_1, X_2) = (x_1, x_2)$ has probability p^2 since the two rolls are independent.

   The RV $Y_1 = \max(X_1, X_2)$ is also a random variable over the set $\{1, \cdots, k\}$, however it is not uniform. We now calculate the probability of a couple of different outcomes:

   Given that $Y_1 = 1$, in how many different ways could (X_1, X_2) be chosen? There is only one way, namely $(X_1, X_2) = 1$.

   Given that $Y_1 = 2$, in how many different ways could (X_1, X_2) be chosen? There are three ways, namely $(X_1, X_2) \in \{(1, 2), (2, 1), (2, 2)\}$.

   We could continue in this way but since we don't know k we instead notice the following: The number of ways that we could get an outcome $Y_1 \leq y_1$ is y_1^2 since there are y_1 different values for both X_1 and X_2 that would result in this outcome. The number of ways that we can get $Y_1 = y_1$ is thus $N(Y_1 = y_1) = y_1^2 - (y_1 - 1)^2 = 2y_1 - 1$. Since the probability of each of these outcomes is $p^2 = 1/k^2$, we get $\Pr[Y_1 = y_1] = (2y_1 - 1)/k^2$

   (b) Since all faces of the individual dice are equiprobable we could swap the order of the faces and still have the same situation. Hence $\Pr[Y_1 = y_1] = \Pr[Y_2 = (k + 1 - y_1)] = (2y_2 - 1)/k^2$.

   (c) Using the results from (a) we get:

$$\mathbb{E}[Y_1] = \sum_{y_1=1}^{k} y_1(2y_1 - 1)\frac{1}{k^2} = \frac{1}{k^2}\sum_{y_1=1}^{k}(2y_1^2 - y_1) = \frac{\frac{1}{3}k(k+1)(2k+1) - \frac{1}{2}k(k+1)}{k^2}$$

$$\mathbb{E}[Y_1] = \frac{(k+1)(4k-1)}{6k}.$$

   For calculating $\mathbb{E}[Y_2]$, $Y_2 = \min(X_1, X_2)$, we use the results from (b) and thus get:

$$\mathbb{E}[Y_2] = \sum_{y_2=1}^{k}(k + 1 - y_2)(2y_2 - 1)\frac{1}{k^2} = \frac{k+1}{k^2}\sum_{y_2=1}^{k}(2y_2 - 1) - \frac{(k+1)(4k-1)}{6k}$$

$$\mathbb{E}[Y_2] = \frac{k^2(k+1)}{k^2} - \frac{k(k+1)(4k-1)}{6k^2} = \frac{(k+1)(2k+1)}{6k}.$$

   (d) Using the last equation from (c) we note:

$$\mathbb{E}[Y_2] = \frac{k^2(k+1)}{k^2} - \mathbb{E}[Y_1].$$

   Thus

$$\mathbb{E}[Y_1] + \mathbb{E}[Y_2] = k + 1.$$

   (e) We already have an expression for $\mathbb{E}[Y_1] + \mathbb{E}[Y_2]$. To calculate the other side of the equation:

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \sum_{x=1}^{k} x\frac{1}{k} = \frac{1}{k}\frac{k(k+1)}{2} = \frac{k+1}{2}.$$

   Thus $\mathbb{E}[Y_1] + \mathbb{E}[Y_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$. This is not a surprising result since we can also get it directly from the linearity of expectations

$$\mathbb{E}[Y_1] + \mathbb{E}[Y_2] = \mathbb{E}[Y_1 + Y_2] = \mathbb{E}[\max(X_1, X_2) + \min(X_1, X_2)].$$

Since there are only two dice, either $(\max, \min) = (X_1, X_2)$ or $(\max, \min) = (X_2, X_1)$ (or they could be equal but the argument still holds). Hence

$$\mathbb{E}[Y_1] + \mathbb{E}[Y_2] = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]\,.$$

This argument cannot be made without using the linearity of expectation in the first step as this implies that we can take the X_1 and X_2 from a single event.

2. (a) We first make the two explicit calculations. Let X be the number of heads for n flips. We note that X can take on any integer value between 0 and n. Since the coin is fair the probability to get k heads and $n-k$ tails is the same as getting $n-k$ heads and k tails. Hence the PMF is symmetric around $\frac{n}{2}$:

$$\Pr[X = \frac{n}{2} + x] = \Pr[X = \frac{n}{2} - x], \quad x \in \{0, \cdots, \frac{n}{2}\}.$$

We can verify this by plugging into the relevant PMF. From the description we know that X is binomial with PMF

$$\Pr[X = k] = \binom{n}{k}\frac{1}{2}^n = \frac{n!}{(n-k)!k!}\frac{1}{2}^n$$

and thus

$$\Pr[X = \frac{n}{2} + x] = \Pr[X = \frac{n}{2} - x] = \frac{n!}{(\frac{n}{2} - x)!(\frac{n}{2} + x)!}\frac{1}{2}^n.$$

This equation only holds for even n, but in our case this is true. Using this symmetry we can conclude that

$$\Pr[X \geq \frac{n}{2} + k] = \Pr[X \leq \frac{n}{2} - k].$$

(b) If we sum over all potential outcomes the probability should sum up to 1. We can thus conclude that

$$1 = \Pr[X \leq \frac{n}{2} - k] + \Pr[\frac{n}{2} - k < X < \frac{n}{2} + k] + \Pr[X \geq \frac{n}{2} + k].$$

Plugging in our result from (b) we get

$$1 = 2\Pr[X \geq \frac{n}{2} + k] + \Pr[\frac{n}{2} - k < X < \frac{n}{2} + k].$$

We now observe that we can exploit the symmetry again to simplify

$$\Pr[\frac{n}{2} - k < X < \frac{n}{2} + k] = \Pr[X = \frac{n}{2}] + 2\Pr[\frac{n}{2} - k < X < \frac{n}{2}].$$

Combing these two expressions and dividing by 2 yields the desired expression:

$$\Pr[X \geq \frac{n}{2} + k] = \frac{1}{2} - \frac{1}{2}\Pr[X = \frac{n}{2}] - \Pr[\frac{n}{2} - k < X < \frac{n}{2}]$$

(c) Plugging in the specific numbers and provided approximations we get:

$$\Pr[100 - 55 < X_{100} < 55] \approx \frac{1}{1.27 \cdot 10^{30}}(1.01 \cdot 10^{29} + 2 \cdot 3.50 \cdot 10^{29}) \approx 0.631$$

$$\Rightarrow \Pr[X_{100} \geq 55] \approx 0.185,$$

and

$$\Pr[1000 - 550 < X_{1000} < 550] \approx \frac{1}{1.07 \cdot 10^{301}}(2.70 \cdot 10^{299} + 2 \cdot 5.21 \cdot 10^{300}) \approx 0.999065$$

$$\Rightarrow \Pr[X_{1000} \geq 550] \approx 0.000467.$$

(d) Here we do a proper calculation of the tight Chernoff bound as demonstrated during the lectures. If you use the simplified one provided that still gives full points, this would give some extra points beyond that. For calculating the Chernoff bound we use the MGF, which for a binomial variable is

$$M_{X_n}(t) = (1 - p + pe^t)^n$$

and in our specific case (with $p = 0.5$)

$$M_{X_n}(t) = \frac{1}{2^n}(1 + e^t)^n.$$

The Chernoff bound states that for a RV X

$$\Pr[X \geq a] \leq \min_{t>0} \frac{M_X(t)}{e^{ta}}.$$

We let

$$f_n(t) = \frac{1}{2^n}(1+e^t)^n e^{-ta}$$

and calculate the first derivative

$$\frac{\partial f_n}{\partial t}(t) = \frac{1}{2^n}\left(n(1+e^t)^{n-1}e^t e^{-ta} - a(1+e^t)^n e^{-ta}\right) = \frac{1}{2^n}e^{-ta}(1+e^t)^{n-1}\left((n-a)e^t - a\right).$$

Setting this to zero we find that the only real solution is

$$t_{\min} = \ln\frac{a}{n-a} > 0.$$

By calculating the second derivative

$$\frac{\partial^2 f_n}{\partial t^2}(t) = \frac{1}{2^n}\begin{pmatrix} -ae^{-ta}(1+e^t)^{n-1}\left((n-a)e^t - a\right) \\ +e^{-ta}(n-1)(1+e^t)^{n-2}e^t\left((n-a)e^t - a\right) \\ +e^{-ta}(1+e^t)^{n-1}(n-a)e^t \end{pmatrix}$$

and plugging in $t_{\min}$ we find

$$\frac{\partial^2 f_n}{\partial t^2}(t_{\min}) = \frac{1}{2^n}\begin{pmatrix} -0 \\ +0 \\ +e^{-t_{\min}a}(1+e^{t_{\min}})^{n-1}a \end{pmatrix} > 0$$

and thus confirm that $t_{\min}$ is a minimum. We move on to calculating

$$f_n(t_{\min}) = \ldots = \frac{1}{2^n}\left(\frac{n}{n-a}\right)^{n-a}\left(\frac{n}{a}\right)^a$$

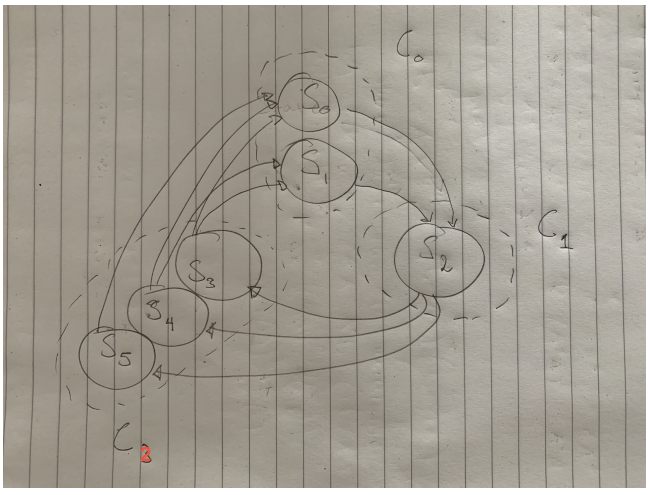
and plug in our cases $(n,a) \in \{(100,55),(1000,550)\}$:

$$f_{100}(t_{\min}) = \frac{1}{2^{100}}\left(\frac{100}{45}\right)^{45}\left(\frac{100}{55}\right)^{55} = \frac{1}{2^{100}}\left(\frac{20}{9}\right)^{45}\left(\frac{20}{11}\right)^{55} = \left(\frac{10}{9}\right)^{45}\left(\frac{10}{11}\right)^{55} \approx 0.606,$$

$$f_{1000}(t_{\min}) = \frac{1}{2^{1000}}\left(\frac{1000}{450}\right)^{450}\left(\frac{1000}{550}\right)^{550} = \left(\frac{1}{2^{100}}\left(\frac{20}{9}\right)^{45}\left(\frac{20}{11}\right)^{55}\right)^{10} = (f_{100}(t_{\min}))^{10} \approx 0.00668.$$

In both cases we see that the bounds hold (sanity check) and that in the latter it is somewhat close (although still a factor 10 off).

3. (a) Here is one simple example satisfying the conditions:



(b) To prove this we consider a state $i \in C_0$. This is arbitrary since we can reorder the classes and still have the same property hold. From the transition probabilities we know that the next state after i must be in C_1, the one after that in C_2 and so on. Let j be the state after $L-1$ transitions and k be the one after L transitions. From the logic above we know that $j \in C_{L-1}$ and thus that $k \in C_0$. In other words, after taking L steps we return to C_0 and start over again. We thus know for certain that the transition probability $P_{ii}(n) = 0$ whenever n is not divisible by L and thus state i has period $d = L$. As mentioned above, we could've picked any state i from any of the classes C_l an simply renamed that classes, so this holds for all states.

4. (a) Let $Y_i = X_i^2$. Then the CDF $F_{Y_i}(y_i)$ of Y_i is

$$F_{Y_i}(y_i) = \Pr[Y_i \leq y_i] = \Pr[X_i^2 \leq y_i] = \Pr[-\sqrt{y_i} \leq X_i \leq \sqrt{y_i}] = \Pr[X_i \leq \sqrt{y_i}] - \Pr[X_i \leq -\sqrt{y_i}]$$

$$\Rightarrow F_{Y_i}(y_i) = F_{X_i}(\sqrt{y_i}) - F_{X_i}(-\sqrt{y_i}),$$

where $F_{X_i}(x)$ is the CDF of X_i. We could plug in the values or just use the fact that $\frac{\partial}{\partial x} F_X(x) = f_X(x)$. Hence:

$$f_{Y_i}(y_i) = f_{X_i}(\sqrt{y_i}) \cdot \frac{1}{2\sqrt{y_i}} - f_{X_i}(-\sqrt{y_i}) \cdot \frac{-1}{2\sqrt{y_i}}.$$

Using the symmetry of the standard normal distribution

$$f_{Y_i}(y_i) = f_{X_i}(\sqrt{y_i}) \cdot \frac{1}{\sqrt{y_i}} = \frac{e^{-\frac{y}{2}} y^{-\frac{1}{2}}}{\sqrt{2\pi}}$$

(b) We use the hint, the MGF of sum of independent RV theorem and the uniqueness of MGF to say the following.
Let $Z_1 \sim \chi^2(k)$ and $Z_2 \sim \chi^2(l)$. Then

$$M_{Z_1+Z_2}(t) = M_{Z_1}(t) \cdot M_{Z_2}(t) = (1-2t)^{-k/2}(1-2t)^{-l/2} = (1-2t)^{-(k+l)/2},$$

which we recognise as the MGF of a χ^2 variable with $k+l$ degrees of freedom, and thus $Z_1 + Z_2 \sim \chi^2(k+l)$.

5. Part 1: (a) The bias can be expressed as the error caused by the simplifying assumptions built into the model. In this example, a very simple model is used, which only takes the height and circumference of a tree into account. The variance explains the stability of the model in response to new training examples. Assume you randomly splits the data set in 2 halves and estimate the linear regression model for each of the 2 halves. Each of them will probably get roughly the same estimated parameter β_0 and hence also the same predicted output for a new output since the data set is large compared to the complexity of the model.

(b) Thus, the bias will be high and the variance will be fairly low.

Part 2:

(a) In this model we will classify a new tree according to the class of the closest tree in the training data. This is highly dependent on the selection of the training data. If we would split the data set in two halves and make a k-nearest neighbor model with $k = 1$ for each of these two data sets, it is likely that we would get very different decision boundaries for the two models since we base the predictions on a single training data point.

(b) This means that we have a high variance in the model. Bias: whether or not this is high or low depends on whether we think that the geographic location alone is sufficiently informative for determining the tree type. If this is the case, then the bias is low, since the 1-NN model can describe very flexible mappings (in this case from "location" to "tree type"). If, however, there is relevant information about the tree type available in the features not used in the model, then this can be viewed as a bias due to under-modeling of the "true" input-output relationship.

6. (a) Since (2) has more flexibility than (1) (Note: if $\beta_2 = \beta_3 = 0$ in (2), you get (1), it will be able to fit to the training data at least as good as (1). Thus, E_{train} for (2) $\leq E_{train}$ for (1).

(b) The argument for (a) is still applicable in the training case, i.e., E_{train} for (2) $\leq E_{train}$ for (1).