MVE137 Exam Solutions

Probability and Statistical Learning Using Python

29 October 2021

Part I

1. Let X_i be outcome of roll i of standard fair die, with $\mathbb{P}[X_i = x] = p = 1/6\,\forall x \in \{1, 2, 3, 4, 5, 6\}$. Let N be the number of rolls for which the first two consecutive sixes have just been rolled, i.e. $N = \min_M \text{s.t.} X_M = X_{M-1} = 6$. We would like to calculate $\mathbb{E}[N]$ and use the provided hint to get

$$\begin{aligned} \mathbb{E}[N] &= \mathbb{E}_{X_1}[\mathbb{E}[N|X_1]] \\ &= \sum_x \mathbb{E}[N|X_1 = x]\,\mathbb{P}[X_1 = x] \\ &= \mathbb{E}[N|X_1 = 6]\,p + \mathbb{E}[N|X_1 \neq 6]\,(1-p) \\ \{\text{memoryless}\} &= \mathbb{E}[N|X_1 = 6]\,p + (\mathbb{E}[N] + 1)(1-p) \end{aligned}$$

Similarly, we can condition $\mathbb{E}[N|X_1 = 6]$ upon $X_2 = 6$ to get

$$\begin{aligned} \mathbb{E}[N|X_1 = 6] &= \mathbb{E}[N|X_1 = 6, X_2 = 6]\,p + \mathbb{E}[N|X_1 = 6, X_2 \neq 6]\,(1-p) \\ &= 2p + (\mathbb{E}[N] + 2)(1-p) \end{aligned}$$

We now have two equations and two unknowns $\mathbb{E}[N|X_1 = 6]$ and $\mathbb{E}[N]$. Solve this equation system and you get

$$\mathbb{E}[N] = 42$$

(which is expected since this is the answer to the ultimate question of life, the universe, and everything :))

2. Let $X = \frac{1}{n}\sum_{i=1}^n X_i$. We then have

$$\mathbb{E}[X] = \mu$$

and

$$\mathbb{V}\text{ar}[X] = \{\text{indep.}\} = \sum_{i=1}^n \mathbb{V}\text{ar}\left[\frac{X_i}{n}\right] = \{\text{ident. dist.}\} = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Chebyshev's inequality for X becomes

$$\mathbb{P}\left[\left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| > \varepsilon\right] < \frac{\sigma^2}{n\varepsilon^2}.$$

Since

$$\frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$$

we get the WLLN.

For the second part what changes is that $\mathbb{V}\text{ar}[X_i] = \sigma_i^2$ is not equal for every i, but we have a bound on the variance i.e. there exists a maximum variance $\sigma^2_{\max}$ s.t. $\sigma_i \leq \sigma_{\max} \forall i \in \mathbb{N}^+$.

We thus get

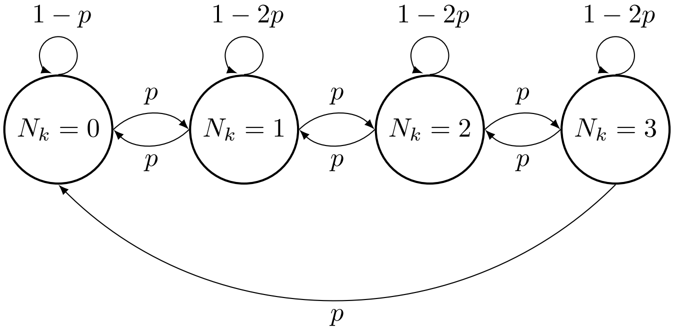
$$\mathbb{V}\text{ar}[X] = \{\text{indep.}\} = \sum_{i=1}^n \mathbb{V}\text{ar}\left[\frac{X_i}{n}\right] \leq n\frac{\sigma^2_{\max}}{n^2} = \frac{\sigma^2_{\max}}{n}$$

and plugging this into Chebyshev we arrive at a similar expression to the first part:

$$\mathbb{P}\left[\left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| > \varepsilon\right] < \frac{\mathbb{V}\text{ar}[X]}{\varepsilon^2} \leq \frac{\sigma^2_{\max}}{n\varepsilon^2}.$$

The final step is identical to the first part, and thus we get the WLLN.

3. The Markov Chain for the number of coins in the bin N_k looks like this:



where $p = 1/5 = \mathbb{P}[Y_k = y]$ and N_k is the number of coins in the bin when customer k is served. The state $N_k = 4$ is not included since the clerk will take this money "immediately after the customer departs, leaving the penny bin empty", i.e. before the next customer. Using **Thm 4** from the formulas we get

$$\begin{aligned}
\mathcal{S} &= \{0,1,2\}, \mathcal{S}' = \{3\} \Rightarrow p\pi_2 = 2p\pi_3 && \Rightarrow \pi_2 = 2\pi_3 \\
\mathcal{S} &= \{0,1\}, \mathcal{S}' = \{2,3\} \Rightarrow p\pi_1 = p\pi_2 + p\pi_3 && \Rightarrow \pi_1 = 3\pi_3 \\
\mathcal{S} &= \{0\}, \mathcal{S}' = \{1,2,3\} \Rightarrow p\pi_0 = p\pi_1 + p\pi_3 && \Rightarrow \pi_0 = 4\pi_3
\end{aligned}$$

Plugging this into $\sum_i \pi_i = 1$ we get $\pi_3 = 1/10$ and thus

$$\boldsymbol{\pi} = \left[\frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10}\right]$$

For the second part we note that the clerk only receives a reward when $Y_k = 4$ and $N_k = 3$. Similarly to what we did in quesiton 1 we get

$$\mathbb{E}[R_k] = \sum_{y,n} \mathbb{E}[R_k | Y_k = y, N_k = n]\, \mathbb{P}[Y_k = y]\, \mathbb{P}[N_k = n] = \mathbb{E}[R_k | Y_k = 4, N_k = 3]\, \mathbb{P}[Y_k = 4]\, \mathbb{P}[N_k = 3].$$

Since Y_k and N_k are independent and the clerk receives 4 pennies when rewarded we get

$$\mathbb{E}[R_k] = 4 \cdot p \cdot \pi_3 = 0.08 \text{ pennies/customer.}$$

4. For calucalting the MGF we use the definition of an MGF to get

$$\begin{aligned}
M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx \\
\{\text{combine exp. and complete the square}\} &= e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-(\mu+t\sigma^2))^2/(2\sigma^2)} dx \\
&= e^{\mu t + \frac{t^2\sigma^2}{2}} \cdot 1,
\end{aligned}$$

where the last step comes from integrating over the pdf of a gaussian RV with mean $\mu + t\sigma^2$ and variance σ^2.

For the second part of calculating the Chernoff bound we let

$$f(t) = \frac{M_X(t)}{e^{t(\mu+\gamma)}} = e^{-\gamma t + \frac{\sigma^2 t^2}{2}}$$

and calculate

$$\frac{df}{dt} = (-\gamma + \sigma^2 t) e^{-\gamma t + \frac{\sigma^2 t^2}{2}}.$$

Since $e^x > 0 \forall x$ we get an extremum point at

$$t_{\text{min}} = \frac{\gamma}{\sigma^2}.$$

We verify that this is a minimum by plugging into second derivative

$$\frac{d^2f}{dt^2} = (\sigma^2)e^{-\gamma t+\frac{\sigma^2 t^2}{2}} + (-\gamma + \sigma^2 t)^2 e^{-\gamma t+\frac{\sigma^2 t^2}{2}}$$

$$\frac{d^2f}{dt^2}(t_{\text{min}}) = \sigma^2 \cdot (>0) + 0 \cdot (>0) > 0.$$

We also note that $t_{\text{min}} = \frac{\gamma}{\sigma^2} > 0$ and thus the bound becomes

$$\mathbb{P}[X \geq \mu + \gamma] \leq \exp\left(-\frac{\gamma^2}{2\sigma^2}\right).$$

5. (a)

$$\beta^* = (X^T X)^{-1} X^T y \tag{1}$$

(b) Ridge regression requires two hyper-parameters. The regularization parameter $\lambda > 0$ and the bandwidth of the guassian kernal $\sigma > 0$. KNN only needs the number of neighbours $K \geq 1$

(c) Several solutions are possible for each of the problems. We only choose here the ones that seem to be most appropriate (i.e., the simplest one). Some methods such as ridge regression may need to be regularized enough.

Problem	R1	R2	R3	R4	R5
Best model among (i-v)	i	iii	No model will be good. i will avoid over fitting.	iii and iv	v

(d) In problem 1, the relation between x and y seem to be linear. Models iii, iv, v might lead to over fitting (although there seem to be sufficiently many points in the data set) if they are not regularized enough.

(e) One solution is to use cross-validation to calibrate the hyper-parameters to regularize enough the methods. Cross validation can also be used to select the best model among (i-v).

6. (a) Logistic regression solves the following optimization problem

$$min_{\beta_0, \beta \in \mathfrak{R}^2} \sum_{i=1}^{n} l(\beta_0 + \beta^T x,\ y) \tag{2}$$

where $\beta_0 \in \mathfrak{R}$ is the intercept (which may be included in the inputs) and $l(\hat{y}, y) \ = \ y \log(1+e^{\hat{y})+(1-y)\cdot\log(1-e^{\hat{y}})}$ is the logistic loss contrary to least squares regression, there is no closed form solution. One needs to use iterative convex optimization algorithms.

(b) Linear discriminant analysis assumes that the data is generated from a mixture of Gaussian. Given a group (label y_i), the variables X_i are independently sampled from the same multi-variate Gaussian distribution. Another common assumption that simplifies the computation of the solution is the homogeneity of variance between groups.

(c)

Problem	C1	C2	C3	C4	C5
Best model among (i-iv)	iii	i, ii, iv	No model will be good. i, ii, iv will avoid over fitting.	i	iv

7. (a) By denoting $y \ = \ \begin{bmatrix} x_1^2 & x_2^2 \end{bmatrix}^T$, one way to find the expression is to analytically solve

$$x\overset{\wedge}{\beta} \ = \ y \tag{3}$$

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} \begin{bmatrix} \overset{\wedge}{\beta}_0 \\ \overset{\wedge}{\beta}_1 \end{bmatrix} = \begin{bmatrix} x_1^2 \\ x_2^2 \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} \overset{\wedge}{\beta}_0 \\ \overset{\wedge}{\beta}_1 \end{bmatrix} = \begin{bmatrix} -x_1x_2 \\ x_{1+}x_2 \end{bmatrix} \tag{5}$$

this gives

$$\begin{aligned} f(x;D) &= \overset{\wedge}{\beta}_0 + \overset{\wedge}{\beta}_1 x \\ &= -x_1x_2 \ + \ (x_1 + x_2)x \end{aligned} \tag{6}$$

(b) Since the only random element of the training data is its two input samples x_1, x_2 which both have a uniform distribution on $[-1, 1]$ the average train model $\overline{f}(x)$ is

$$\overline{f}(x) = E_D\left[f(x;D)\right] = \iint f(x;x_1,x_2)p(x_1,x_2)dx_1x_2 \tag{7}$$

$$\begin{aligned} E_D\left[f(x;D)\right] &= \tfrac{1}{2} \cdot \tfrac{1}{2} \textstyle\iint f(x;x_1,x_2)dx_1x_2 \\ &= \tfrac{1}{4} \textstyle\int_{-1}^{1}\int_{-1}^{1} -x_1x_2 \ + \ (x_1 + x_2)xdx_1dx_2 \\ &= 0 \end{aligned} \tag{8}$$

(c)

$$\begin{aligned} E_x\left[\overline{f}(x) - E_D\left[f(x;D)\right]^2\right] &= E_x\left[\left(0 - x^2\right)^2\right] \\ &= \tfrac{1}{2} \textstyle\int_{-1}^{1} x^4 dx \\ &= \tfrac{1}{5} \end{aligned} \tag{9}$$