Group Assignment 3

Introduction to Data Science H23

Elise Hammarström Theodora Moldovan Ella Schmidtobreick
Georgios Tsouderos Finn Vaughankraska

December 21, 2023

All group members attempted the proofs/exercises individually before meeting. After discussing, we finalized the problems and each of us chose a problem to write up in LaTeX.

1 Exercise 5.20

Exercise 5.20. Show that the relative entropy risk is the same risk as we saw in Section 4.2, it only differs by a constant.

Solution The goal is to show that the relative entropy risk $R(G) = \int \ln\left(\frac{f^*(x)}{g(x)}\right) f^*(x)\, dx$, where f^* and g are probability density functions, is the same as the risk in Section 4.2 (Maximum Likelihood Estimation), $R(\alpha) = \mathbb{E}[\ln(p_\alpha(Z))]$, where p_α is a parametric family of PDFs, and Z follows the distribution p_{α^*}.

Note that the relative entropy risk measures how one distribution G (with PDF g) diverges from another distribution F (with PDF f^*). The risk in Section 4.2 relates to the expectation of the log-likelihood under a parametric family of distributions.

In the relative entropy risk, the term $\ln\left(\frac{f^*(x)}{g(x)}\right)$ compares two distributions. In the maximum likelihood risk, $\ln(p_\alpha(z))$ is the log likelihood for a single distribution p_α evaluated at z.
Rewrite the relative entropy risk in terms of expectation:

$$R(G) = \mathbb{E}_{f^*}\left[\ln\left(\frac{f^*(X)}{g(X)}\right)\right]$$

where X is a random variable following f^*.

Notice that both forms involve the expectation of a log term. The difference is that the relative entropy risk involves a ratio of PDFs, whereas the maximum likelihood risk involves a single PDF.
Consider the Kullback-Leibler divergence (KLD) between f^* and g, which is equivalent to the relative entropy risk:

$$KLD(f^*||g) = \int f^*(x) \ln\left(\frac{f^*(x)}{g(x)}\right) dx$$

.
Assume p_{α^*} in Section 4.2 corresponds to f^* and p_α corresponds to g in the relative entropy risk context. Then the KLD becomes:

$$KLD(p_{\alpha^*}||p_\alpha) = \int p_{\alpha^*}(x) \ln\left(\frac{p_{\alpha^*}(x)}{p_\alpha(x)}\right) dx$$

The maximum likelihood risk $R(\alpha)$ is the expectation of the negative log likelihood under p_α:

$$R(\alpha) = -\mathbb{E}_{p_{\alpha^*}}[\ln(p_\alpha(Z))]$$

Since Z follows p_{α^*}, this can be rewritten as:

$$R(\alpha) = -\int p_{\alpha^*}(x)\ln(p_\alpha(x))dx$$

The KLD (relative entropy risk) involves the log of the ratio of two PDFs, while the maximum likelihood risk involves the log of a single PDF. To reconcile these, we can add and subtract the same term:

$$KLD(p_{\alpha^*}||p_\alpha) = \int p_{\alpha^*}(x)\ln(p_{\alpha^*}(x))dx - \int p_{\alpha^*}(x)\ln(p_\alpha(x))dx$$

The second term is the negative of $R(\alpha)$. The first term $\int p_{\alpha^*}(x)\ln(p_{\alpha^*}(x))dx$ is a constant with respect to α since it only depends on p_{α^*}, the "true" distribution.
Thus, the relative entropy risk $R(G)$ and the risk from Section 4.2 $R(\alpha)$ are equivalent up to a constant term. This constant term is $\int p_{\alpha^*}(x)\ln(p_{\alpha^*}(x))dx$, which is independent of the variable distribution p_α or g and only depends on the fixed distribution p_{α^*} or f^*. This completes the proof that the two risks are essentially the same, differing only by a constant.

2 Exercise 6.11

Lemma 6.10. Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, define $u_i = D(u_{i-1})$ then the sequence $v_i = u_i \mod K$ for $1 \leq K \leq M$ is pseudorandom on $\{0, 1, \ldots, K-1\}$ if M is a multiple of K.

Proof We aim to prove that the sequence v_i is pseudorandom on $\{0, 1, \ldots, K-1\}$. Since M is a multiple of K, there exists an integer n such that $M = nK$. The generator D has a period M, meaning that the sequence $\{u_i\}$ repeats every M elements. Therefore, for any i, $u_i = u_{i+M}$.
Now, consider the sequence $v_i = u_i \mod K$. We will show that this sequence covers all elements in $\{0, 1, \ldots, K-1\}$ and then repeats, thus being pseudorandom on this set.

Let us focus on how often an arbitrarily chosen single value in the sequence v_i appears. Since the period of u_i is M and $K \leq M$, the period of the sequence v_i also needs to be $\leq M$. Due to the fact, that $v_i = u_i \bmod K$ and u_i has period $M = n * K$ we can write the sequence u_i as $u_i = n * k + v_i$. Therefore, each value of the sequence v_i will appear exactly n times if one has a look at period M. This can be seen when looking at the following sequence:

$$\begin{aligned} v_i, \ldots, v_{i+M} &= v_i, \ldots, v_{i+K}, v_{i+K+1}, \ldots, v_{i+2*K}, \ldots, v_{i+n*K} \\ &= \underbrace{u_i \bmod K, \ldots, u_{i+K} \bmod K}_{=v_i,\ldots,v_{i+K}}, \underbrace{u_{i+K+1} \bmod K, \ldots, u_{i+2*K}\ modK}_{=v_i,\ldots,v_{i+k}}, \underbrace{\ldots, u_{i+n*K} \bmod K}_{(n-2)*v_i,\ldots,v_{i+K}} \\ &= v_i, \ldots, v_{i+K}, v_{i+1}, \ldots, v_{i+K}, \ldots, v_{i+K} \\ &= v_i, \ldots, v_{i+K} \quad \text{appears n times} \end{aligned}$$

For any i, we have $v_i = u_i \mod K$. Given that $u_i = u_{i+M}$, it follows that $v_i = u_i \mod K = u_{i+M} \mod K$.

However, since M is a multiple of K, u_{i+M} gives the same remainder as u_i when divided by K. Therefore, $v_i = v_{i+M}$. This implies that the sequence $\{v_i\}$ repeats every M elements, and since M is a multiple of K, the sequence $\{v_i\}$ covers all elements in $\{0, 1, \ldots, K-1\}$ in its period. Since a period is the smallest positive integer, such that $v_{i+T} = v_i$,the final period of sequence v_i will only be K. Thus, $\{v_i\}$ is pseudorandom on $\{0, 1, \ldots, K-1\}$.

3 Exercise 6.19

Theorem 6.18 (Box-Muller). Suppose that $U_1, U_2 \overset{\text{IID}}{\sim} \text{Uniform}([0, 1])$, then

$$Z_0 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$$

$$Z_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$$

are independent random variables, and $Z_0, Z_1 \sim \mathcal{N}(0,1)$.

Proof Consider bivariate normal RV. Z, then the distribution of $Y = |Z|^2$ is χ^2 distributed with 2 degrees of freedom. Furthermore $W = Z/|Z|$, is uniformly distributed on the unit circle. We know that Y, W are independent (see Exercise 6.19). Thus to generate a bivariate normal it is enough to generate from a χ^2 distribution with 2 degrees of freedom and a point from the uniform distribution on the circle. The χ^2 with 2 degrees of freedom is just the exponential distribution with parameter 1/2, as such we can generate it using the inversion sampling method (Theorem 5.38). The rest of the proof is left as an exercise. □

Exercise 6.19. First show that W, Y in the proof above are independent. Then show that W generated using $(\cos(2\pi U_2), \sin(2\pi U_2))$ is uniform on the unit circle. Finally to show that Z_0, Z_1 are independent, since they are Gaussian it suffices to show that their covariance is zero.

Solution

1. Independence of W and Y:

   Proof. We defined

$$Y = z_0^2 + z_1^2,$$

$$W = \left[\frac{z_0}{\sqrt{z_0^2+z_1^2}}, \frac{z_1}{\sqrt{z_0^2+z_1^2}}\right].$$

   We can show they are independent by showing their covariance is 0

$$\text{Cov}(Y,W) = E\left[(Y - E[Y])(W - E[W])\right] = E[YW] - E[Y]E[W]$$

   Since z_0 and z_1 are bivariate normal variables we can then evaluate the following terms in the covariance equation:

$$E[W_i] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{z_i}{\sqrt{z_0^2+z_1^2}} \cdot \frac{1}{2\pi} e^{-(z_0^2+z_1^2)/2}\, dz_0\, dz_1 = 0$$

$$E[YW_i] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_i \cdot \sqrt{z_0^2+z_1^2} \cdot \frac{1}{2\pi} e^{-(z_0^2+z_1^2)/2}\, dz_0\, dz_1 = 0$$

$$\text{Cov}(Y,W) = 0 - E[Y] \cdot 0 = 0$$

   Therefore the covariance of Y and W is 0 and they are independent.

   □

2. Uniform Distribution of W on the Unit Circle:

   Proof. In order to show W generated by (w_0, w_1) $(\cos(2\pi U_2), \sin(2\pi U_2))$ is uniform on the unit circle, we must first convert into polar coordinates as follows:

$$r = \sqrt{w_0^2 + w_1^2} = \sqrt{\cos(2\pi U_2) + \sin(2\pi U_2)} = \sqrt{1} = 1$$

   This proves that the radius is always one for any U_2 generating W. Then

$$\theta = \arctan\left(\frac{w_1}{w_0}\right) = \arctan\left(\frac{\sin(2\pi U_2)}{\cos(2\pi U_2)}\right) = \arctan(\tan(2\pi U_2)) = 2\pi U_2$$

   Shows that U_2, which is a random uniform [0, 1] variable, is scaled uniformly becoming [0, 2π]. Therefore W is is uniformly distributed on the unit circle.

   □

3. Independence of Z_0 and Z_1: To show the independence of two Gaussian random variables, it's sufficient to demonstrate that their covariance is zero.

   Proof. Given $Z_0 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$ and $Z_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$, where $U_1, U_2 \overset{\text{IID}}{\sim}$ Uniform$([0,1])$, we aim to show that Z_0 and Z_1 are independent Gaussian random variables. This can be demonstrated by proving that their covariance is zero.

   The covariance of Z_0 and Z_1 is defined as:

   $$\mathrm{Cov}(Z_0, Z_1) = \mathbb{E}[Z_0 Z_1] - \mathbb{E}[Z_0]\mathbb{E}[Z_1]$$

   Since Z_0 and Z_1 are standard normal, their means are zero, thus simplifying the expression to:

   $$\mathrm{Cov}(Z_0, Z_1) = \mathbb{E}[Z_0 Z_1]$$

   Expanding $\mathbb{E}[Z_0 Z_1]$:

   $$\begin{aligned}\mathbb{E}[Z_0 Z_1] &= \mathbb{E}\left[\sqrt{-2\ln(U_1)}\cos(2\pi U_2)\cdot\sqrt{-2\ln(U_1)}\sin(2\pi U_2)\right]\\ &= \mathbb{E}\left[-2\ln(U_1)\cos(2\pi U_2)\sin(2\pi U_2)\right]\end{aligned}$$

   The term $-2\ln(U_1)$ is independent of $\cos(2\pi U_2)$ and $\sin(2\pi U_2)$ due to the independence of U_1 and U_2. Thus, the expectation can be broken down:

   $$\mathbb{E}\left[-2\ln(U_1)\right]\cdot\mathbb{E}\left[\cos(2\pi U_2)\sin(2\pi U_2)\right]$$

   The critical point is evaluating $\mathbb{E}\left[\cos(2\pi U_2)\sin(2\pi U_2)\right]$. For a uniform distribution over $[0, 2\pi]$, this expectation is zero due to the symmetry and periodicity of the sine and cosine functions:

   $$\mathbb{E}\left[\cos(2\pi U_2)\sin(2\pi U_2)\right] = 0$$

   Therefore, $\mathbb{E}[Z_0 Z_1] = 0$ and consequently, $\mathrm{Cov}(Z_0, Z_1) = 0$. Hence, Z_0 and Z_1 are independent.

   □

4 Exercise 7.12

Exercise 7.12. Prove Lemma 7.11 in a similar way to Lemma 7.7.

Lemma 7.11. For a finite inhomogeneous Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \ldots, \mu_0(s_k)),$$

where $\mu_0(s_i) = \mathbb{P}(X_0 = s_i)$, and transition matrices

$$(P_1, P_2, \ldots), \quad P_t := (P_t(s_i, s_j))_{(s_i,s_j)\in\mathbb{X}\times\mathbb{X}}, t \in \{1, 2, \ldots\}$$

we have for any $t \in \mathbb{Z}_+$ that the distribution at time t given by:

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), \ldots, \mu_t(s_k)),$$

where $\mu_t(s_i) = \mathbb{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t.$$

Proof

At time-step t, we will apply the law of total probability

$$\mathbb{P}^t(X_n = x_n) = \sum_{x_{n-1}} \mathbb{P}^t(X_n = x_n | X_{n-1} = x_{n-1})\mathbb{P}^{t-1}(X_{n-1} = x_{n-1})$$

$$\mathbb{P}^t(X_n = x_n) = \sum_{x_{n-1}} P^t_{x_{n-1}x_n} \mathbb{P}^{t-1}(X_{n-1} = x_{n-1})$$

But $\mathbb{P}^{t-1}(X_{n-1} = x_{n-1})$ can be calculated in the same way as previously. After plugging the result in the previous equation we get :

$$\mathbb{P}^t(X_n = x_n) = \sum_{x_{n-1},x_{n-2}} P^t_{x_{n-1}x_n} P^{t-1}_{x_{n-2}x_{n-1}} \mathbb{P}^{t-2}(X_{n-2} = x_{n-2})$$

Since n is arbitrary we can apply it again until we reach X_0, resulting in :

$$\mathbb{P}^t(X_n = x_n) = \sum_{x_{n-1},x_{n-2},\ldots,x_0} P^t_{x_{n-1}x_n} P^{t-1}_{x_{n-2}x_{n-1}} ... P^1_{x_0x_1} \mathbb{P}(X_0 = x_0)$$

In the equation above, given the initial distribution, $\mathbb{P}(X_0 = x_0) = \mu_0$ and the fact that the rest is just a sequence of matrix multiplications, we can write:

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t.$$

$\square$

5 Exercise 7.17

Exercise 7.17. Do the proof of Theorem 7.16 by using the necessary Definitions.

Theorem 7.16. Let $W_1, \ldots, \overset{\text{IID}}{\sim} F$ such that (ρ_t, W_t) is a RMR for a transition matrix P_t, for all $t \in \mathbb{N}$. Then if $X_0 \sim \mu_0$,

$$X_t := \rho_t(X_{t-1}, W_t), \quad t \in \mathbb{N},$$

is a Markov chain with initial distribution μ_0 and transition matrix P_t at time t.

Proof Given X_0 has distribution μ_0, this establishes the initial state of the Markov chain. The process $\{X_t\}$ is defined recursively as $X_t = \rho_t(X_{t-1}, W_t)$. We need to show $\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1})$. Due to W_t being IID and independent of past X_s for $s < t$, X_t depends only on X_{t-1} and W_t. This implies X_t is conditionally independent of $X_0, X_1, \ldots, X_{t-2}$ given X_{t-1}, satisfying the Markov property.

By the RMR definition, $\mathbb{P}(\rho_t(x, W_t) = y) = P_t(x, y)$. Hence, due to the transition matrix definition, $\mathbb{P}(X_t = y | X_{t-1} = x) = \mathbb{P}(\rho_t(x, W_t) = y) = P_t(x, y)$. This shows that the transition probability from state x to state y at time t is given by $P_t(x, y)$.

The process $\{X_t\}$ with $X_t = \rho_t(X_{t-1}, W_t)$ and established transition probabilities forms a Markov chain. The chain starts with initial distribution μ_0 and follows the transition matrix P_t at each step t.

This completes the proof. The sequence $\{X_t\}$ as defined by the random mapping representation ρ_t and the IID random variables W_t forms a Markov chain with the specified initial distribution and transition matrices.