COMP2610 / COMP6261 — Information Theory: Assignment 1

Robert C. Williamson

Out: 29 August 2018 **Due**: 5pm, 21 September 2018

Instructions

This assignment contributes 20% to your final score for the course.

Marks: The maximum marks available are given in square brackets. Each question is worth 25 marks in total. For questions where you are asked to prove results, if you can not prove a precedent part, you can still attempt subsequent parts of the question assuming the truth of the earlier part.

Assignment Project Exam Help

COMP2610 st *t* expected to answer question 5.

https://eduassistpro.github.io/

COMP6261 st m Q1, Q2, or Q3]. You will only get marks for 4 questions (two of w

Add WeChat edu_assist_pro

Submission: You should submit a paper copy of your assig mission box on the ground floor of the CSIT building by the due date I your answers, but if you use hand writing it must be clearly legible.

You must include your name, UNI-ID and your Tutor group clearly at the top of the first page of your submission.

You must show all your working: it is not sufficient to merely write down an answer!

Cheating and Plagiarism: All assignments must be done individually. *Plagiarism is a university offence* and will be dealt with according to university procedures `http://academichonesty.anu.edu.au/UniPolicy.html`.

1. In the game of Gah! Bad DECAF! a player draws letter tiles out of a bag. Each of the 16 tiles in the bag has a letter $l \in \{A, B, C, D, E, F, G, H\}$ on one side and a value $v \in \{1, 3, 5\}$ on the other side. The number of tiles for each letter that appear in the bag, and the associated value for each letter are shown in the table below:

Letter	A	B	C	D	E	F	G	H
Count	4	2	1	2	4	1	1	1
Value	1	5	3	3	1	5	3	5

Let $d = \{\texttt{yes}, \texttt{no}\}$ indicate whether a tile is a DECAF tile or not. That is, $d = \texttt{yes}$ if $l \in \{D, E, C, A, F\}$ and $d = \texttt{no}$, otherwise. The uppercase symbols L, V, and D will be used to respectively denote the ensembles associated with the letter, value, and DECAF status of a tile randomly drawn from the bag. All tiles in a bag are equally likely to be chosen when a draw is made.

(a) Suppose a single tile is drawn from the bag. Calculate:

i. The information in seeing an A. That is, compute $h(l = A)$. [2 points]

ii. The information in seeing a value of 3. That is, $h(v = 3)$. [2 points]

iii. The conditional information $h(l = A|v = 5)$ of flipping a tile over and seeing an A given that a tile with value 1 was drawn. [2 points]

Assignment Project Exam Help

iv. The conditional information $h(v = 1|l = A)$ of flipping a tile over and seeing a 1 given that an A tile was drawn. Give an intuitive explanation for this answer. [2 points]

v. The en

https://eduassistpro.github.io/

(b) Calculate th $I(L$ V value of a tile, and the mutual information $I(D; V)$ between t f a tile. Using the data processing inequality, explain the relat uantities. [5 points]

Add WeChat edu_assist_pro

(c) Exactly half of all people who play Gah! Bad DECAF! cheat by secretly throwing out all their As and Es, leaving behind a bag with only 8 of the 16 original tiles. We will write $c = \texttt{unfair}$ if a person cheats and $c = \texttt{fair}$ otherwise and let C be the associated ensemble with $p(\texttt{fair}) = p(\texttt{unfair}) = 0.5$.

i. What is the conditional entropy $H(D|C)$? [2 points]

ii. Use the previous result to determine $H(C, D)$. [3 points]

iii. Suppose you were playing against a randomly chosen opponent and she plays a DECAF letter (i.e., $d = \texttt{yes}$) three times in a row, replacing her tile back in her bag after each play. What probability would you assign to her being a cheat? [5 points]

2. Let Y be a random variable with possible outcomes $\{0, 1\}$, and $p(Y = 1) = \frac{1}{2}$. Let X be a random variable with possible outcomes $\mathcal{X} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}\}$. Define

$$\mathbf{p} = (p(X = \mathsf{a}|Y = 1), p(X = \mathsf{b}|Y = 1), p(X = \mathsf{c}|Y = 1))$$
$$\mathbf{q} = (p(X = \mathsf{a}|Y = 0), p(X = \mathsf{b}|Y = 0), p(X = \mathsf{c}|Y = 0)).$$

(a) Suppose that

$$\mathbf{p} = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$$
$$\mathbf{q} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right).$$

Compute $I(X; Y)$. [3 points]

(b) Using the definition of mutual information, show that for any choice of $\mathbf{p}$, $\mathbf{q}$,

$$I(X; Y) = \frac{D_{\mathrm{KL}}(\mathbf{p}||\mathbf{m}) + D_{\mathrm{KL}}(\mathbf{q}||\mathbf{m})}{2},$$

where $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$. [10 points]

Assignment Project Exam Help

(c) Let Z be a random variable with outcomes $x \in \mathcal{X}$, such that for all $x \in \mathcal{X}$ and $y \in \{0, 1\}$, $p(Z = x|Y = y) = p(X = x|Y = 1 - y)$. Using part (b) compute $I(Z; Y)$. Explain i

https://eduassistpro.github.io/

(d) Suppose ample of a random v $; Y) > I(Z; Y)$. Explain your answer in terms of the data-processing ineq [4 points]

Add WeChat edu_assist_pro

(e) Suppose $\mathbf{p}$ and $\mathbf{q}$ are as in part (a). Give an examp Z with possible outcomes $x \in \mathcal{X}$ satisfying $I($ our answer in terms of the data-processing inequality. [4 points]

3. [25 points] Suppose a music collection consists of 4 albums: the album Alina has 5 tracks; the album Bends has 12; the album Coieda has 15; and the album Debut has 12.

(a) How many bits would be required to uniformly code:

i. The index of all the albums? (We are not encoding the actual music, but merely the titles — the metadata). Give an example uniform code for the albums. [2 points]

ii. Only the tracks in the album Alina. Give an example of a uniform code for the tracks assuming they are named "Track 1", "Track 2", etc. [2 points]

iii. All the tracks in the music collection? [2 points]

(b) What is the raw bit content required to distinguish all the tracks in the collection? [2 points]

(c) Suppose every track in the music collection has an equal probability of being selected. Let A denote the album title of a randomly selected track from the collection.

i. Write down the ensemble for A — that is, its alphabet and probabilities. [2 points]

ii. What is the raw bit content of A^4? [2 points]

iii. What is the smallest value of δ such that the smallest δ-sufficient subset of A^4 contains fewer than 256 elements? [3 points]

Assignment Project Exam Help

iv. What is the largest value of δ such that the essential bit content $H_\delta(A^4)$ is strictly gre [4 points]

(d) Suppose the

https://eduassistpro.github.io/

i. Comp o decimal places (you may use a computer or calculator to obtain th t the expression you are approximating). [2 points]

Add WeChat edu_assist_pro

ii. Approximately how many elements are in t or A when $N = 100$ and $\beta = 0.1$? [2 points]

iii. Is it possible to design a uniform code to send large blocks of album titles with a 95% reliability using at most 1.5 bits per title? Explain why or why not. [2 points]

4. Let X be a real-valued random variable with mean μ and standard deviation $\sigma < \infty$. The *median* of X is the real number $m \in \mathbb{R}$ satisfying

$$p(X \geq m) = p(X \leq m) = \frac{1}{2}.$$

If g is some function on $\mathbb{R}$, then $\arg\min_x g(x)$ is the value x^* that minimises $g(x)$: that is $g(x^*) \leq g(x)$ for all x, and $\min_x g(x) = g(x^*)$. The argmin may not be unique, and so we consider it to be set-valued and thus write $x^* \in \arg\min_x g(x)$.

(a) Prove that

$$m \in \arg\min_{c\in\mathbb{R}} \mathbb{E}\left(|X - c|\right)$$

(Hint: break the expectation into two conditional expectations.) [5 points]

(b) Prove that

$$|\mu - m| \leq \sigma.$$

(Hint: The function $x \mapsto |x|$ is convex). [5 points]

(c) The *α-quantile* of X is the real number q_α which satisfies[1]

$$p(X \leq q_\alpha) = \alpha.$$

Assignment Project Exam Help

For $\tau \in (0, 1)$, define the *pinball loss* $\ell_\tau : \mathbb{R} \to \mathbb{R}$ via

https://eduassistpro.github.io/

Show that for any $\alpha \in (0, 1)$,

Add WeChat edu_assist_pro

$$q_\alpha \in \arg\min_{c\in\mathbb{R}} \tag{1}$$

[5 points]

(d) One can show that

$$\mu \in \arg\min_{c\in\mathbb{R}} \mathbb{E}\left((X - c)^2\right) \tag{2}$$

and given (2), by substitution we have

$$\sigma^2 = \min_{c\in\mathbb{R}} \mathbb{E}\left((X - c)^2\right)$$

In light of this, and of part (c), for $\alpha \in (0, 1)$, give an interpretation of

$$Q_\alpha = \min_{c\in\mathbb{R}} \mathbb{E}\left(\ell_\alpha(X - c)\right).$$

Argue that, like σ^2, for $\alpha \in (0, 1)$, Q_α measures the deviation or variability of X. Explain why $Q_\alpha(X) = 0$ only when X is constant. What advantages might Q_α have over σ^2 as a measure of variability? [5 points]

[1] The quantile is not necessarily unique. Observe that $q_{1/2} = m$.

(e) (Don't panic!) A *property* T of a distribution P is a real number that summarises some aspect of the distribution. One often nees to simplify from grappling with an entire distribution to a single number summary of the distribution. Means, variances, and quantiles are all properties, as is the entropy since we can just as well consider the entropy of a random variable X as a property of its distribution P (think of the defintion) and thus write $H(P)$. Expressions such as (1) and (2) are examples of *eliciting* a property of a distribution. In general[2] one might have a function S (called a scoring function) such that for some desired property T of a distribution P one has

$$T(P) \in \arg\min_{c \in \mathbb{R}} \mathbb{E}_{Y \sim P} S(c, Y), \tag{3}$$

where $Y \sim P$ means that Y is a random variable with distribution P. It turns out[3] that not all properties T can be elicited; that is, there is no S such that $T(P)$ can be written in the form (3). For example, the variance can not be elicited. A necessary (and sufficient) condition for a property to be elicitable is that for arbitrary P_0, P_1 such that $T(P_0) = T(P_1)$ we have that for all $\alpha \in (0, 1)$,

$$T((1-\alpha)P_0 - \alpha P_1) = T(P_0).$$

The distribution $(1-\alpha)P_0 + \alpha P_1$ is called a *mixture* of P_0 and P_1.

Assignment Project Exam Help

Construct an example P_0, P_1 such that $H(P_0) = H(P_1)$ but for some $\alpha \in (0, 1)$, the entropy of the mixture distribution satisfies

https://eduassistpro.github.io/

to prove (an be no equation of the form (3) which yields the entrop

(Hint: Easier than it might look. Start simple!) [5 points]

Add WeChat edu_assist_pro

[2] In (1) we have $S(c, x) = |x - c|$ and in (2) we have $S(c, x) = (x - c)^2$.

[3] This is a non-trivial result, which requires several additional technicalities to state precisely. It is proved in [Ingo Steinwart, Chloé Pasin, Robert C. Williamson, and Siyu Zhang, "Elicitation and identification of properties." In *JMLR: Workshop and Conference Proceedings*, 35 (Conference on Learning Theory), pp. 1–45. 2014].

5. Suppose X is a real valued random variable with $\mu = \mathbb{E}(X) = 0$.

   (a) Show that for any $t > 0$,
   $$\mathbb{E}(e^{tX}) \leq e^{g(t(b-a))}$$
   where $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$, with $\gamma = -a/(b-a)$.
   (Hint: write X as a convex combination of a and b, where the convex weighting parameter depends upon X. Exploit the convexity of the function $x \mapsto e^{tx}$ and the fact that inequalities are preserved upon taking expectations of both sides, since expectations are integrals.) [5 points]

   (b) By using Taylor's theorem, show that for all $u > 0$, $g(u) \leq \frac{t^2(b-a)^2}{8}$ and hence
   $$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}.$$
   Furthermore, suppose $\mathbb{E}(X) = \mu \neq 0$. Show that
   $$\mathbb{E}(e^{tX}) \leq e^{t\mu} e^{t^2(b-a)^2/8}.$$
   [5 points]

   (c) Show that for any random variable X and any $t > 0$,
Assignment Project Exam Help
   $$p(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathbb{E}(e^{tX}).$$
   (Hint: Re [illegible]
https://eduassistpro.github.io/
   (d) Assume [illegible] $p(X_i \in [a, b]) = 1$ for $i \in \{1, \ldots, n\}$ and $\mathbb{E}(X_i) = \mu$ for $i \in \{1, \ldots,$ [illegible] X_i be the *empirical mean*. Show that for any $\epsilon \in (0, 1)$,
Add WeChat edu_assist_pro
   $$p\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq [illegible].$$
   [5 points]

   (e) Suppose $X_1, \ldots, X_n$ are iid Bernoulli random variables with parameter $\theta \in (0, 1)$. The above result implies that
   $$p(|\bar{X}_n - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2}$$
   Compare this bound with what one obtains using the Chebyshev and Markov inequalities. For different θ, plot the three bounds for varying n. For $\epsilon = 0.1$, what value n_0 do each of the three bounds tell you that is necessary in order that for $n \geq n_0$, $p(|\bar{X}_n - \theta| > 0.1) \leq 0.1$? [5 points]