ELV832: Special Module in Machine Learning

Mid-term Quiz, Form: $\boxed{\text{A}}$

Maximum marks: 16

(Answer all questions on this question paper. Read all section-specific instructions carefully.)

Name: ______________________________

Entry Number: ______________________________

Section 1. Numerical/Short-answer questions

Instructions: Please write the answer (showing your working/reasoning) immediately following each part of each question.

1. Consider a random variable X which denotes, for a randomly chosen pixel in an RGB image, which of the three colour components has the highest intensity value. For a given data set representing a particular class of images, we estimate that $p(X = \text{'R'}) = 0.5$, $p(X = \text{'G'}) = 0.25$, $p(X = \text{'B'}) = 0.25$.

(a) Calculate the 'volume' of X (as defined in class), based on these estimates. **[2.5]**

$$H(X) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4}\right)$$

$$= -\left(\frac{-1}{2} - \frac{1}{2} - \frac{1}{2}\right) = \frac{3}{2}$$

Volume of $X = 2^{H(X)} = 2^{3/2} = 2\sqrt{2} = 2.828.$

(b) What is the maximum possible value of the volume of a random variable whose range has cardinality $|X|$? Under what circumstance will this be realised? **[1.5]**

Unif. distr. – $H(X) = -\sum_{i=1}^{|X|} \frac{1}{|X|}\log\frac{1}{|X|} = \log|X|$

volume $= 2^{H(X)} = |X|$. (= 3 here.)

(c) What is the minimum possible value of the volume of a random variable of cardinality $|X|$? When will this be realised? **[1.5]**

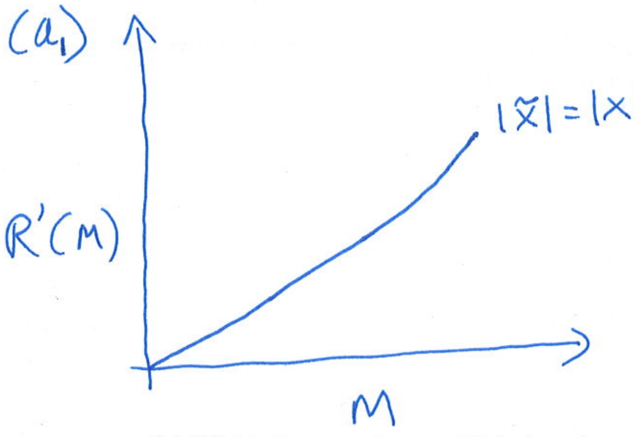
Point distr. – $H(X) = -(1\log 1 + \sum 0\log 0) = 0$

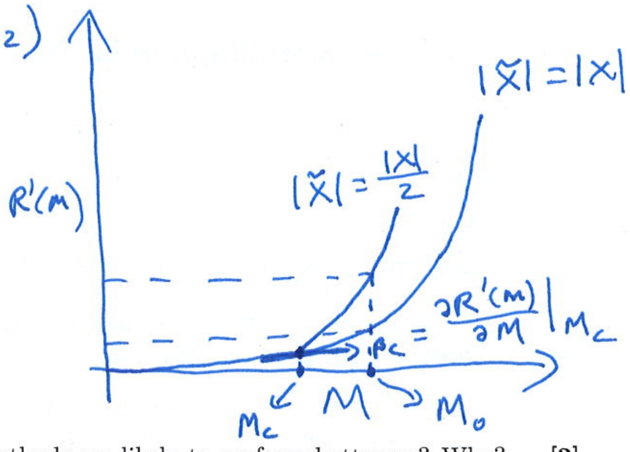
volume $= 2^{H(X)} = 1$.

(d) Consider where the value obtained in (a) lies, on the possible range of values from the minimum to the maximum. Based on this, can you give an intuitive explanation (in words) of what the volume of a random variable is capturing/representing? **[2.5]**

Note that $H(X)$ is avg. no. of bits to represent an outcome of X. For a string of k bits, 2^k is the no. of different possible strings. So $2^{H(X)}$ is like the avg. no. of 'effective' different outcomes of X. This interpretation becomes exact at the two boundary cases mentioned above. For the actual example, 'effective' no. of outcomes, informationally, is slightly less than 3.

2. Consider the below information-plane curves for information bottlenecking, for two different data sets whose underlying joint distribution $p(\mathbf{x}, y)$ has been estimated:





(a) Which data set do you think deep learning methods are likely to perform better on? Why? **[2]**

(a2), because pretty high M can be achieved at pretty low R'(M), so 'bottlenecking' through multiple layers should be possible without much loss of relevant information.

(b) Consider the two curves marked $|\tilde{X}| = |X|$ and $|\tilde{X}| = |X|/2$ on the right-hand plot. Why do these curves overlap for a certain region (*i.e.*, $\beta < \beta_c$)? What exactly happens at the critical value of β_c, beyond which the curves bifurcate? **[4]**

For $\beta < \beta_c$: No more than $|X|/2$ codewords are being used, because for the corresponding required M, additional codewords don't help lower R'(M).

At β_c: we get a 'phase transition': the optimal soft clustering now has $> \frac{|X|}{2}$ clusters. If we restrict codewords, it pushes up the rate R'(M).

(c) Consider the marked value $M = M_0$. At this value, different values of $R'(M)$ (the optimal rate) are obtained for the two curves. Explain why. **[2]**

As β_c has been crossed, optimal rate requires more than $\frac{|X|}{2}$ codewords. For $\frac{|X|}{2}$ codewords, to retain the same value of $M^{(=M_0)}$, each codeword needs to be more 'informative' (on avg.) i.e., pick out a narrower range of values of X. This means $H(X|\tilde{X})$ is lower, and so $I(X;\tilde{X}) = H(X) - H(X|\tilde{X})$ becomes higher.