Introduction to Information Theory and Machine Learning August 5, 2018
(Prof. Permuter Haim, Mr. Oron Sabag and Mr. Ron Shoham)

## Final Exam - Moed B

Total time for the exam: 3 hours!

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1) **Duality bound for memoryless channel (34 Points)**: In this question, you will prove a simple upper bound on the capacity of a memoryless channel. The channel is given by $P_{Y|X}$ and the capacity is denoted by $C$. We will also need a new distribution on channel outputs $Q_Y(\cdot)$ which is an arbitrary distribution. Finally, when we write $I_P(X;Y)$ this means that the mutual information is computed with respect to $P_{X,Y}$.

   a) (4 points) Write the capacity $C$ of a memoryless channel, $P_{Y|X}$ in terms of a divergence (do not use mutual information).

   b) (6 points) Complete $\leq, =, \geq$ between the following expressions (prove you answer):

   $$I_P(X;Y) \quad \text{Vs.} \quad \sum_{x\in\mathcal{X}} \left[P_X(x) D\left(P_{Y|X=x}||Q_Y\right)\right] - D(P_Y||Q_Y).$$

   c) (5 points) Prove the duality bound (justify each step):

   $$C \leq \max_{x\in\mathcal{X}} D\left(P_{Y|X=x}||Q_Y\right).$$

   d) (6 points) Find sufficient and necessary conditions for the tightness of the duality bound.

   e) (7 points) We now define $P_{Y|X}$ to be a binary symmetric channel (BSC) with transition probability $\alpha$. Compute the duality bound when $Q_Y \sim \text{Bernoulli}(0.25)$ and $Q_Y \sim \text{Bernoulli}(0.5)$. Your answers should be simple and without a maximum.

   f) (6 points) Are the two upper bounds from the previous question equal the capacity of the BSC? prove your answers.

   The upper bound that you proved is called the duality upper bound. As seen above, there are conditions for the tightness of the bound, and wise choices of the test distribution $Q_Y$ may give rise to good bounds on the capacity.

2) **Compression using machine learning (30 Points)**
   A source sequence $x_1, x_2, ..., x_n$ is given where the cardinality of the alphabet of $x_i$ is 4, namely, $|\mathcal{X}| = 4$. You observe a noisy version of the sequence, $y_1, y_2, ..., y_n$ where $Y_i = X_i + Z_i$, and $Z_i$ has a Gaussian distribution with zero mean and some variance. You do not know the variance of the noise $Z_i$ nor the explicit alphabet of $X$, but, you do know that the noise is with high probability lower than the minimal difference between the values of $X$.

   a) (5 points) What would you expect the histogram of $y^n$ to be. Draw it.

   b) (5 points) Given the sequence $y^n$, suggest an ML algorithm that estimates $x^n$. (provide a pseudo code).

   c) (5 points) In what category the ML algorithm that you suggested in 2b (previous sub question) is: supervised learning or non-supervised learning.

   d) (5 points) Now, you have the following system that is given in Fig. 1. A new sequence $y^l$ arrives to the encoder and it has a similar distribution as the sequence $y^n$ from 2b. The encoder first estimates $x_i$ from $y_i$ using the inference of the ML algorithm that you have build and trained in 2b and then compress it using variable length coding. Suggest how to build variable length codes using the sequence $y^n$ from 2b. Suggest at least two variable-length codes.

   e) (5 points) Are the variable codes you suggested optimal, and if yes in what sense.

   f) (5 points) Repeat all the previous sub-question where $x_i, y_i$ and $z_i$ are two-dimensional vectors. i.e.

   $$x_i = (x_i^{(1)}, x_i^{(2)})$$
   $$z_i = (z_i^{(1)}, z_i^{(2)})$$
   $$Z_i \sim \mathcal{N}(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}).$$

   *In the two dimensional part you draw a representative contour instead of sketching histogram.

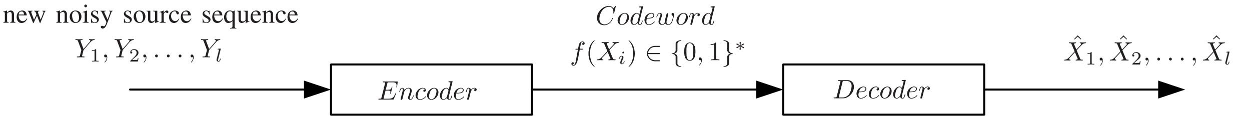

Fig. 1: ML and Source coding problem. The encoder in the figure converts a value $y_i$ into $x_i$ using the ML algorithm you suggested in 2b and then into sequence of bits $\{0,1\}$ of variable length denoted as $\{0,1\}^{l(x)}$. The goal of the decoder is to reconstruct the original signal $X_i$.

3) **True/False (28 Points)**:

a) **Information Theory:** Given is a joint distribution $P_{X,Y}$ and a deterministic function $f : \mathcal{X} \to \mathcal{X}$ that satisfy

$$H(Y|f(X)) \leq H(Y|X).$$

On each of the next statements write True/False.

i) (4 points) There exists a Markov chain $Y - f(X) - X$.

ii) (4 points) The function $f(\cdot)$ is an injective (one to one) function.

iii) (4 points) If $f(X) \sim \text{Unif}(1, \ldots, \mathcal{X})$, then $X \sim \text{Unif}(1, \ldots, \mathcal{X})$.

b) **Machine learning:**

i) (4 points)**True/False:** The log-likelihood of the data will *always* increase through successive iterations of the expectation maximization algorithm.

ii) (4 points) **True/False:** In distribution tree a node can have more than one 'father'.

iii) (4 points) We wish to generate classifier which classify between K classes. In order to do so we train a Neural Net with softmax output layer (with k output neurons). Let us note the output vector as $\hat{Q}_\theta(x)$. We use the cross-entropy cost function to measure the distance between the output distribution and the real labels. **True/False:** By the law-of-large-numbers, the cost is equal to the KL-divergence between the distribution $\hat{Q}_\theta(x)$ and the distribution of the real labels.

iv) (4 points) **True/False:** Decision Tree which was built by ID3 algorithm guarantees 0% training error.

4) **Tree Distribution (23 Points)**: You wish to generate a model to predict if a mushroom is poisonous or not. You have some empirical data:

| Example | Is heavy | Is smelly | Is spotted | Is smooth | Is poisonous |
|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 0 | 1 | 0 | 0 |
| C | 1 | 1 | 0 | 1 | 0 |
| D | 1 | 0 | 0 | 1 | 1 |
| E | 0 | 1 | 1 | 0 | 1 |
| F | 0 | 0 | 1 | 1 | 1 |
| G | 0 | 0 | 0 | 1 | 1 |
| H | 1 | 1 | 0 | 0 | 1 |

a) (9 points) Calculate the empirical mutual information between all couples of features (including *Is poisonous*).

b) (7 points) Build tree distribution for the data according to the maximum-likelihood criteria. You have a constraint that the node of 'Is poisonous' must be the main root (head) of the tree.

c) (7 points) Use the tree you built to determine by the maximum-likelihood criteria whether U,V,W are poisonous or not. If it happens to be that there is a tie, you define it as poisonous.

| Example | Is heavy | Is smelly | Is spotted | Is smooth | Is poisonous |
|---|---|---|---|---|---|
| U | 0 | 1 | 1 | 1 | ? |
| V | 0 | 1 | 0 | 1 | ? |
| W | 1 | 1 | 0 | 0 | ? |

Note:

$$h_b(\frac{1}{8}) = 0.5436, \quad h_b(\frac{1}{4}) = 0.8113, \quad h_b(\frac{3}{8}) = 0.9544, \quad h_b(\frac{1}{7}) = 0.5917, \quad h_b(\frac{2}{7}) = 0.8631,$$
$$h_b(\frac{3}{7}) = 0.9852, \quad h_b(\frac{1}{6}) = 0.6500, \quad h_b(\frac{1}{3}) = 0.9183, \quad h_b(\frac{1}{5}) = 0.7219, \quad h_b(\frac{2}{5}) = 0.9710.$$

Good Luck!