Final Exam - Moed A

Total time for the exam: 3 hours!

Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:_____ "

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, or disprove it, e.g. by providing a counter-example, otherwise.

1) **Entropy rate (36 Points):** The concept of entropy for a stochastic process $\{X_i\}$ can be expressed using the *entropy rate*. This is defined by the following equation, provided that the limit exists:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, ..., X_n). \tag{1}$$

Consider a typewriter with an *m*-letter keyboard. Each letter is distributed i.i.d with equal probability.

a) **(5 points)** Compute the total number of possible sequences that are n letters long.

b) **(6 points)** Determine the *entropy rate* of this typing process.

Entropy Rate for Stationary process - A stochastic process $\{X_t\}$ is said to be **stationary** if for every n, for all $t_1, t_2, ..., t_n$ and for all h, the joint probability distribution function $p(X_{t_1}, X_{t_2}, ..., X_{t_n})$ is equal to $p(X_{t_1+h}, X_{t_2+h}, ..., X_{t_n+h})$, i.e., the joint probability distribution is invariant under time shifts.

For a **stationary** process, answer the following:

c) **(5 points)** (True / False) Does the following equality holds? Explain.

$$H(X_t|X_1, X_2, ..., X_{t-1}) = H(X_{t+i}|X_{1+i}X_{2+i}, ..., X_{t-1+i}), \quad \forall i, t \in \mathbb{Z}. \tag{2}$$

d) **(5 points)** (True / False) Does the following claim correct? Explain.

$$H(X_{n+1}|X_1, ..., X_n) \geq H(X_n|X_1, ..., X_{n-1}) \tag{3}$$

e) **(4 points)** Does the series $a_n = H(X_n|X_1, ..., X_{n-1})$ exhibit monotonicity? If so, what type of monotonicity?

f) **(5 points)** (True / False) The series a_n converge?

g) **(6 points)** Prove that $H(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, ...X_1)$. Utilize the Cesaro Mean theorem: If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$. (Don't forget to show that the series $\{a_i\}$ converges!)

Question Insight - In this question, we gained insights into the concept of entropy rate, which extends the notion of entropy. In the case of independent and identically distributed (i.i.d.) processes, the entropy and entropy rate coincide. However, for stationary stochastic processes, the entropy rate becomes more significant. In class we discovered that entropy serves as the optimal compression rate, but for ergodic and stationary processes, the fundamental limit of compression is determined by the entropy rate.

2) **ML algorithms (32 Points):** For Parts (a) through (c), consider the following problem. In a joint project between ML developers and doctors, a set of features (e.g., temperature, height) have been extracted for each patient. These features will be used to determine whether a new visiting patient has any of three possible diseases: diabetes, heart disease, or Alzheimer's. A patient can have one or more of these diseases.

a) **(8 points)** The ML developers have decided to use a neural network to solve this problem, but they are considering two different approaches:

- Training a separate neural network for each of the diseases.
- Training a single neural network with one output neuron for each disease and a shared hidden layer.

For each approach, draw a neural network that represents it. Under what statistics of the data would the first approach be favored over the second, and vice versa? Justify your answer.

b) **(8 points)** It was decided to train a classifier for **each** disease using a logistic regression learning algorithm. The classifier is trained to obtain **MAP** estimates for the logistic regression trainable weights W, input features X, and decision output Y. Our MAP estimator optimizes the objective

$$W \longleftarrow \arg\max_W \ln \left[P(W) \prod_l P\left(Y^l|X^l, W\right) \right] \tag{4}$$

where l refers to the lth training example. We adopt a Gaussian prior with zero mean for the weights $W = \langle w_1 ... w_n \rangle$ accompanied by a constant weight factor C,

$$W \longleftarrow \arg\max_W \left[C \ln P(W) + \sum_l \ln P\left(Y^l|X^l, W\right) \right] \tag{5}$$

Provide the expression for the equivalent cost function for this setup. Additionally, identify the type of regularization derived from this process.

c) **(8 points)** We re-run the derived learning algorithm with different values of the constant C. Please answer the following true/false question, and explain/justify your answer. **True/False:** The average log probability of the training data is unlikely to increase as we increase the value of C.

d) **(8 points)** Figure 1 illustrates a subset of our training data when we have only two features: X_1 and X_2. Assume that $C = 0$ and our logistic regression model is well-trained. Draw a possible decision boundary of the algorithm. Explain your choice and explain what could happen for choosing a large value of C.

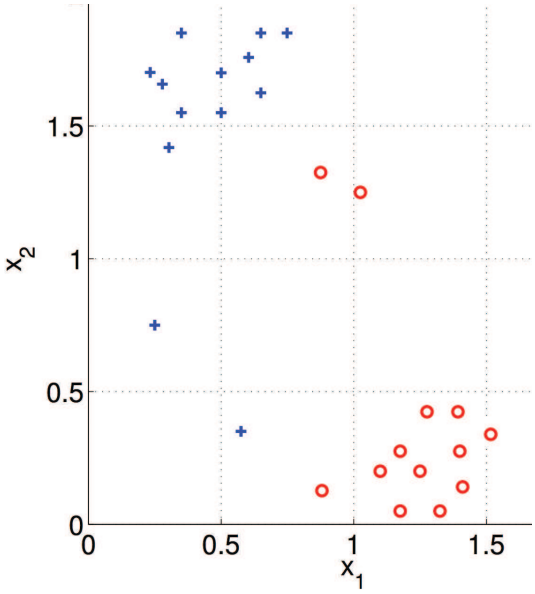

Fig. 1: Classification boundary

3) **Polar codes (36 Points):** Consider a binary erasure channel W with erasure probability p. One step of the polarization process creates a better channel W^+ and a worse channel W^- from two independent copies of W.

a) **(8 points)** The polar code creates 4 effective channels W^{++}, W^{+-}, W^{-+}, W^{--}. Write down the capacities of these 4 channels in terms of information-theoretic quantities.

b) **(12 points)** Compute explicitly the capacities of the 4 channels in terms of the parameter p.

c) **(4 points)** Suppose we would like to send at the rate $3/4$ bits per channel use. Which of the U_i's should be frozen, and which should be set as information bits?

d) **(8 points)** We repeat the polarization process n times to create 2^n different channels from 2^n compies of W. Let $\overline{W}$ and $\underline{W}$ be the best and worst channels among these 2^n channels. Compute explicitly the capacities of $\overline{W}$ and $\underline{W}$ in terms of the parameters p and n.

e) **(4 points)** What happens to the capacities of $\overline{W}$ and $\underline{W}$ as $n \to \infty$?

Good Luck!