Introduction to Information Theory and Machine Learning July 7th, 2022.
(Prof. Permuter Haim, Mr. Bashar Huleihel, and Mr. Omer Luxembourg)

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.
Solution: The total number of such sequences is m^n.

b) **(6 points)** Determine the *entropy rate* of this typing process.
Solution: Since all of the letter are equally likely to be typed,

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) && \text{(2a)}\\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1) && \text{(2b)}\\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) && \text{(2c)}\\ &= \lim_{n\to\infty} \frac{1}{n} n \log m && \text{(2d)}\\ &= \log m. && \text{(2e)} \end{aligned}$$

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.

$$H(X_t|X_1, X_2, ..., X_{t-1}) = H(X_{t+i}|X_{1+i}X_{2+i}, ..., X_{t-1+i}), \quad \forall i, t \in \mathbb{Z}. \tag{3}$$

Solution: True. The conditional entropy is given by the probability function of the process, which is stationary.

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

$$H(X_{n+1}|X_1, ..., X_n) \geq H(X_n|X_1, ..., X_{n-1}) \tag{4}$$

Solution: False,

$$\begin{aligned} H(X_{n+1}|X_1, ..., X_n) &\leq H(X_{n+1}|X_2, ..., X_n) && (5)\\ &= H(X_n|X_{n-1}, ..., X_1), && (6) \end{aligned}$$

where (5) is valid because conditioning reduced entropy, and (6) holds due to the stationarity of the process.

e) **(4 points)** Does the series $a_n = H(X_n|X_1, ..., X_{n-1})$ exhibit monotonicity? If so, what type of monotonicity?
Solution: Monotonic decreasing by previous question (6).

f) **(5 points)** (True / False) The series a_n converge?
Solution: True. $a_n \geq 0, \forall n \in \mathbb{N}$ by definition of entropy, and with (6) and the fact that the series is monotonic decreasing, we get that $\forall n \in \mathbb{N}, \quad 0 \leq H(X_{n+1}|X_1, ..., X_n) \leq H(X_1)$. Thus the limit exists.

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!)

Solution:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, ..., X_n) \tag{7a}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1) \tag{7b}$$

$$= \lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_1), \tag{7c}$$

where (7b) follows by the chain rule, and (7c) originates from the Cesaro Mean theorem, and the fact that the limit exists from the previous question.

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.

Solution:

- Neural networks with a shared hidden layer can capture dependencies between diseases. It can be shown that in some cases, when there is a dependency between the output nodes, having a shared node in the hidden layer can improve the accuracy.
- If there is no dependency between diseases (output neurons), then we would prefer to have a separate neural network for each disease.

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

$$W \longleftarrow \arg\max_{W} \ln \left[P(W) \prod_{l} P\left(Y^l | X^l, W\right) \right] \tag{8}$$

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,

$$W \longleftarrow \arg\max_{W} \left[C \ln P(W) + \sum_{l} \ln P\left(Y^l | X^l, W\right) \right] \tag{9}$$

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

Solution:

We adopt a Gaussian prior with zero mean for the weights $W = \langle w_1 ... w_n \rangle$, making the above equivalent to:

$$W \longleftarrow \arg\max_{W} \left[-C \sum_{i} w_i^2 + \sum_{l} \ln P\left(Y^l | X^l, W\right) \right], \tag{10}$$

This equation can also be equivalently expressed in the form we discussed in class:

$$W \longleftarrow \arg\min_{W} \left[C \sum_{i} w_i^2 - \sum_{l} \ln P\left(Y^l | X^l, W\right) \right] \tag{11}$$

which correspond to an L2 regularization.

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.

Solution:
True. As we increase C, we give more weight to the prior of the predictor and therefore constrain the predictor. Thus it makes our predictor less flexible to fit to the training data (over-constraining the predictor makes it unable to fit to the training data).

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.

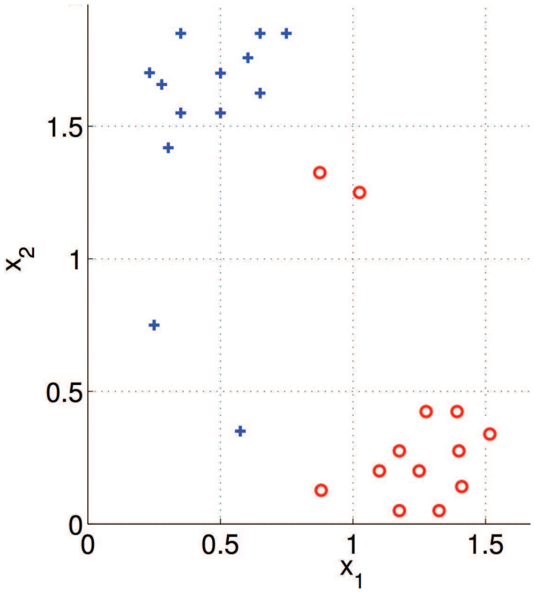

Fig. 1: Classification boundary

Solution:
The decision boundary for logistic regression is linear. One candidate solution which classifies all the data correctly is shown in Figure 2. The decision boundary depends on the value of C, as it is possible for the trained classifier to miss-classify a few of the training data if we choose a large value of C.

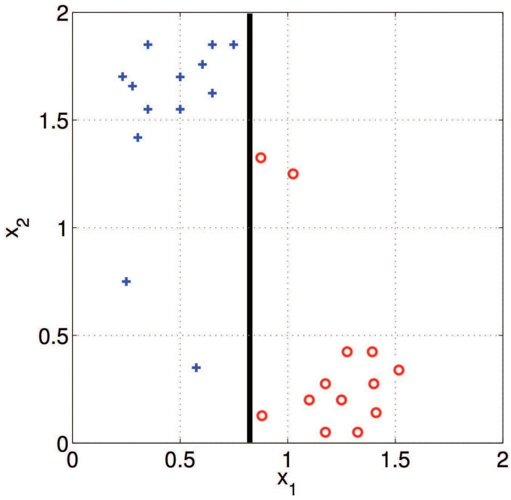

Fig. 2: Classification boundary

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.
Solution:

The capacities in terms of information-theoretic quantities are:

$$\begin{aligned}C(W^{--}) &= I(U_1; Y^4)\\ C(W^{-+}) &= I(U_2; Y^4|U_1)\\ C(W^{+-}) &= I(U_3; Y^4|U^2)\\ C(W^{++}) &= I(U_4; Y^4|U^3).\end{aligned}$$

b) **(12 points)** Compute explicitly the capacities of the 4 channels in terms of the parameter p.
Solution:
As shown in class, the first step of splitting yields two new binary erasure channels:

$$\begin{aligned}W^{-} &= BEC(1-(1-p)^2)\\ W^{+} &= BEC(p^2).\end{aligned}$$

Subsequently, by applying an additional step of splitting, we obtain the following channels:

$$\begin{aligned}W^{--} &= BEC(1-(1-(1-(1-p)^2))^2) = BEC(1-(1-p)^4)\\ W^{-+} &= BEC((1-(1-p)^2)^2)\\ W^{+-} &= BEC(1-(1-p^2)^2)\\ W^{++} &= BEC(p^4).\end{aligned}$$

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?
Solution:
We should freeze U_1 to be 0, and transmit information on U_2, U_3, U_4 since the first channel has the smallest capacity. Accordingly, we transmit 3 information bits out of the 4 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.
Solution:
Following item (a), we can directly deduce that:

$$\overline{W} = BEC(p^{2^n}),$$

which implies that $C(\overline{W}) = 1 - p^{2^n}$.
To compute the capacity of $\underline{W}$, let us denote p_n as the error probability of the worst BEC at stage $n \geq 1$, with $p_0 = p$. Thus, we have $\underline{W} = BEC(p_n)$ and $C(\underline{W}) = 1 - p_n$. Next, we will prove by induction that $p_n = 1-(1-p)^{2^n}$. First, we observe that the following relation holds:

$$p_n = 2p_{n-1} - p_n^2. \tag{12}$$

For $n = 1$, we have $p_1 = 1-(1-p)^2$, which indeed holds. Now, assuming the claim is true for $n = k$, i.e., $p_k = 1-(1-p)^{2^k}$, we will show that it also holds for $n = k+1$.

$$\begin{aligned}p_{k+1} &\overset{(a)}{=} 2p_k - p_k^2\\ &= 2(1-(1-p)^{2^k}) - (1-(1-p)^{2^k})^2\\ &\overset{(b)}{=} 2(1-t) - (1-t)^2\\ &= 1 - t^2\\ &= 1-(1-p)^{2^{k+1}},\end{aligned}$$

where (a) follows by (12), and (b) follows by considering $t \triangleq (1-p)^{2^k}$. To conclude, we obtained that

$$C(\underline{W}) = (1-p)^{2^n}.$$

e) **(4 points)** What happens to the capacities of $\overline{W}$ and $\underline{W}$ as $n \to \infty$?
Solution:
In the following, we analyze the case that $p \neq 1$ (the case of $p = 1$ is straightforward). As $n \to \infty$, we have that $p^{2^n} \to 0$. Accordingly, the capacity $C(\overline{W})$ tends to 1. Furthermore, it is evident that the sequence p_n approaches 1 as n increases. Consequently, the capacity $C(\underline{W})$ tends to 0 since the error probability of the worst binary erasure channel converges to 1.

Good Luck!