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

July 7th, 2022.

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) **Uncertainty about true distribution (24 Points):** Consider a source U with alphabet $\mathcal{U} = \{a_1, \dots a_m\}$ and suppose we know that the true distribution of U is either P_1 or P_2, but we are not sure which.
   a) (8 points) **True/False:** There is a prefix code where the length of the codeword associated to a_i is $l_i = \left\lceil \log_2 \left(\frac{2}{P_1(a_i)+P_2(a_i)} \right) \right\rceil$.
   b) (8 points) Show that the average (computed using the true distribution) length $\bar{l}$ of the code constructed in item (a) satisfies $H(U) \le \bar{l} \le H(U) + 2$.
   c) (8 points) Now assume that the true distribution of U is one of k distributions $P_1, \dots, P_k$, but we don't know which. Show that there exists a prefix code satisfying $H(U) \le \bar{l} \le H(U) + \log_2(k) + 1$.

2) **GMM (18 points):** We will derive the EM update rules for a univariate Gaussian Mixture Model with two mixture components. The mean μ will be shared between the two mixture components, but each component will have its own standard deviation σ_k. The model will be defined as follows:

$$z \sim Bernoulli(\theta),$$
$$p(x|z=k) \text{ is } \mathcal{N}(\mu, \sigma_k).$$

   a) (4 points) Write the density defined by this model (i.e. the probability of x, with z marginalized out)
   b) (4 points) E-step - Compute the posterior probability $w^{(i)} = Pr(z^{(i)} = 1|x^{(i)})$
   c) (5 points) M-Step - Calculate the update rule for μ (for a fixed σ_k)
   d) (5 points) M-Step - Calculate the update rule for σ_k (for a fixed μ)

3) **Linear Regression (26 Points):** You are tasked with solving a fitting a linear regression model on a set of m datapoints where each feature has some dimensionality d. Your dataset can be described as the set $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$. You initially decide to optimize the loss objective:

$$J = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - {x^{(i)}}^T\theta)^2,$$

using Batch Gradient Descent - in which each step involves calculations over the **entire** training set. Here, $\theta \in \mathbb{R}^d$ is your weight vector. Assume you are ignoring a bias term for this problem.

   a) (4 points) Write each update of the batch gradient descent, $\frac{\partial J}{\partial \theta}$ in **vectorized** form. Your solution should be a single vector (no summation terms) in terms of the matrix X and vectors Y and θ, where

$$X = \begin{bmatrix} {x^{(1)}}^T \\ \vdots \\ {x^{(m)}}^T \end{bmatrix}, Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}.$$

   b) (7 points) A coworker suggests you augment your dataset by adding Gaussian noise to your features. Specifically, you would be adding *zero-mean*, Gaussian noise of *known vairance* σ^2 from the distribution

$$\mathcal{N}(0, \sigma^2 I),$$

where $I \in \mathbb{R}^{d\times d}$, $\sigma \in \mathbb{R}$. This modifies your original objective to:

$$J_* = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - (x^{(i)} - \delta^{(i)})^T\theta)^2,$$

where $\delta^{(i)}$ are **i.i.d.** noise vectors, $\delta^{(i)} \in \mathbb{R}^d$ and $\delta^{(i)} \sim \mathcal{N}(0, \sigma^2 I)$.
Express the expectation of the modified objective J_* over the Gaussian noise, $\mathbb{E}_{\delta\sim\mathcal{N}}[J_*]$, as a function of the original objective J added to a term independent of your data. Your answer should be in the form

$$\mathbb{E}_{\delta\sim\mathcal{N}}[J_*] = J + C,$$

where C is independent of points in $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$.

Hint: For a Gaussian random vector δ with zero mean, and convariance matrix $\sigma^2 I$

$$\mathbb{E}_{\delta\sim\mathcal{N}}[\delta\delta^T] = \sigma^2 I, \quad \mathbb{E}_{\delta\sim\mathcal{N}}[\delta] = 0.$$

c) (4 points) What effect would adding noise have on model overfitting/underfitting? Explain why.
Remember that the weights update rule is derived from the loss function, which is the expectation of J_*.

d) (4 points) Is this method similar to a regularization method we studied in class? If so, specify the regularization method and prove it and if not, explain why?

e) (3 points) Consider the limits $\sigma \to 0$ and $\sigma \to \infty$. What impact would these extremes in the value of σ have on model training (relative to no noise added)? Explain why.

f) (4 points) Suggest a cost function and a noise that is related to *Dropout*.

4) **Computable lower bounds (32 Points):** In this question, you will prove a simple lower bound on the capacity of a memoryless channel. Let $p(y|x)$ be a memoryless channel, and let $p(x)$ be a distribution on $\mathcal{X}$. Let $r(x|y)$ be an arbitrary conditional distribution on $\mathcal{X}$ given $\mathcal{Y}$, i.e., for each $x \in \mathcal{X}$ and each $y \in \mathcal{Y}$, $r(x|y) \geq 0$ and $\sum_{\tilde{x}\in\mathcal{X}} r(\tilde{x}|y) = 1$. Define the functional $F(p, r)$ as follows:

$$F(p, r) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y|x) \log_2 \left(\frac{r(x|y)}{p(x)}\right).$$

where p in $F(p, r)$ denotes $P(x)$ and $p(y|x)$ is fixed thought the question. Now, for each input distribution p on $\mathcal{X}$, define the conditional distribution r_p as

$$r_p(x|y) = \frac{p(x)p(y|x)}{\sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x})p(y|\tilde{x})}.$$

That is, r_p is the "true" conditional distribution of $\mathcal{X}$ given $\mathcal{Y}$ when p is the input distribution.

a) (8 points) **True/False:** For all conditional distributions r we have $F(p, r) \leq F(p, r_p)$.

b) (4 points) Show that $I(X; Y) = \max_r F(p, r)$.

c) (8 points) **True/False:** The functional $F(p, r)$ is strictly concave in both p and r.

d) (6 points) In Algorithm 1 below, we introduce an iterative algorithm for maximizing a two-variable function. Following the previous items, suggest such an iterative algorithm to compute the capacity.

Algorithm 1 Alternating maximization procedure

---

input: A function g(x,y) that is concave in both x and y.
output: A global maximum of $g(x, y)$.
initiate x_0 to some value and solve $y_0 = \arg\max_y g(x_0, y)$.
set $i = 1$.
while $g(x_i, y_i)$ not converged **do**
  $x_i = \arg\max_x g(x, y_{i-1})$
  $y_i = \arg\max_y g(x_{i-1}, y)$
  compute $g(x_i, y_i)$
  $i = i + 1$
end
return $g(x_i, y_i)$

Note: The Alternating maximization procedure is known to converge to optimal solution when the function $g(x, y)$ is concave in (x, y).

e) (6 points) For a given memoryless channel, let r^* denote the conditional distribution that should be used to obtain the capacity. Write explicitly r^* for the case of a binary symmetric channel with crossover probability 0.2.

Good Luck!