Homework Set #6
Machine learning and information measures

1. **Transfer Entropy :** Define the Transfer Entropy

$$\mathsf{TE}^{(k)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X^{(k)}_{t-1} \middle| Y^{(k)}_{t-1}\right), \tag{1}$$

where $X^{(k)}_t := (X_t, X_{t-1}, ..., X_{t-k+1})$ is a notation for length-k history of a variable X up to time t.

Let $\{X_t\}$ and $\{Y_t\}$ be stationary and first-order Markov processes taking values from the binary alphabet:

- Process $\{X_t\}$ has a deterministic transitions from 0 to 1 or 1 to 0 each time step, i.e.

$$P(X_t|Y^{t-1}, X^{t-1}) = P(X_t|X_{t-1}), \quad P(X_t = x|X_{t-1} = x \oplus 1) = 1, \tag{2}$$

where $P(X_0) \sim \text{Bern}\left(\frac{1}{2}\right)$.

- Process $\{Y_t\}$ is a noisy observation of the last time step of $\{X_t\}$. Assume $\alpha \neq \frac{1}{2}$ and $0 < \alpha < 1$,

$$P(Y_t|Y^{t-1}, X^{t-1}) = P(Y_t|X_{t-1}), \quad P(Y_t = y|X_{t-1} = x) = \begin{cases} 1-\alpha & \text{if } y = x \\ \alpha & \text{if } y \neq x \end{cases}. \tag{3}$$

Reminder: A stochastic process $\{X_t\}$ is said to be **stationary** if for every t_1, t_2 and h, the joint probability distribution function $P(X_{t_1}, X_{t_1+1}, ..., X_{t_1+h})$ is equal to $P(X_{t_2}, X_{t_2+1}, ..., X_{t_2+h})$, i.e., the joint probability distribution is invariant under time shifts.

(a) **True / False** The described joint process $\{X_t, Y_t\}$ is stationary. Explain your answer.

(b) **True / False** $P(Y_t = y, X_{t-1} = x) \neq P(X_t = x, Y_{t-1} = y)$.

(c) Calculate the Mutual Information between Y_t and X_{t-1}, i.e. $I(Y_t; X_{t-1})$.
Hint: Consider to use the fact that $Y_t = X_{t-1} \oplus Z_{t-1}$, where $\{Z_t\}$ are i.i.d. Bern(α).

(d) **True / False** $I(Y_t; X_{t-1}) = I(X_t; Y_{t-1})$.

(e) Show that the Transfer Entropy for $X \to Y$ with lag $k = 1$ is non-zero, i.e., $\mathsf{TE}^{(1)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X_{t-1} \middle| Y_{t-1}\right) > 0$.
Hint: Utilize the relation $Y_t = X_{t-1} \oplus Z_{t-1}$, and the fact that if $Z_1 \sim \text{Bern}(\alpha)$ and $Z_2 \sim \text{Bern}(\beta)$, then $Z_1 \oplus Z_2 \sim \text{Bern}(\alpha - 2\alpha\beta + \beta)$.

(f) Calculate the Transfer Entropy for $Y \to X$ with lag $k = 1$, i.e., $\mathsf{TE}^{(1)}_{\mathcal{Y}\to\mathcal{X}} = I\left(X_t; Y_{t-1} \middle| X_{t-1}\right)$.

2. **Auto-Encoders** Let $\mathbf{X} \sim \mathcal{N}(\mu_x, \Sigma_x)$. We would like to create a generative model of $\mathbf{X}$.

   (a) Assume we use a simple autoencoder (not variational) and X is some single-dimensional random variable (i.e., the encoder $F : \mathbb{R} \mapsto \mathbb{R}$ and so is the decoder $G : \mathbb{R} \mapsto \mathbb{R}$). What would be the optimal choice of F and G for MSE-minimizing?

   (b) In this section we consider the linear case for a variational autoencoder and we will apply the reparametrization trick. We assume that X is m-dimensional. Let:

$$\mu_z = \begin{pmatrix} w_1^\mathsf{T}\mathbf{X} \\ \vdots \\ w_d^\mathsf{T}\mathbf{X} \end{pmatrix} + b, \qquad \Sigma_Z = A \cdot diag(\mathbf{X}), \qquad diag(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}) = \begin{pmatrix} x_1 & 0 & \dots & \dots & 0 \\ 0 & x_2 & 0 & \dots & 0 \\ \vdots & 0 & x_3 & \dots & 0 \\ 0 & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \dots & x_m \end{pmatrix}$$

   with $w_i \in \mathbb{R}^m$ for $i = 1 \dots d$, $A \in \mathbb{R}^{d\times m}$ and $b \in \mathbb{R}^d$.
   What is the distribution of μ_z? What is the distribution of Z given a realization $X = x$?
   Reminder: if $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ and $Y = AX + b$ then $\mathbf{Y} \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathsf{T})$.

   (c) We now focus on the one-dimensional case. In class we had defined the following loss for the variational-autoencoder:

$$\mathcal{L} := \underbrace{\mathbb{E}_{q(z)}\left[\frac{(x - f(z))^2}{2c}\right]}_{:=\mathsf{MSE}} + \underbrace{D_{KL}(\mathcal{N}(g(x), h^2(x)), \mathcal{N}(0,1))}_{:=\mathsf{D}} \tag{4}$$

   where our goal is to apply $\min_{f,g,h} \mathcal{L}$. We define the functions as follows:

$$f(z) = z, \quad g(x) = \sum_{i=1}^{m} w_i x^i, \quad h(x) = \sum_{j=1}^{m} \exp(-u_j x)$$

   Calculate $\frac{\partial \mathsf{MSE}}{\partial w_i}$ for some general i.

   (d) The KL-divergence between two Gaussians $G_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $G_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ is given by:

$$D_{KL}(G_1, G_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

   Calculate $\frac{\partial \mathsf{D}}{\partial w_i}$ for some general i.

(e) Given some fixed learning rate μ, write down the SGD update rule for the models weights $\{w_i\}_{i=1}^m$.

(f) This section is independent of the previous ones. Assume we want to constraint our latent vector Z to have similar statistical characteristics as some other random vector Y. Propose a modification for the loss in (4) to obtain this request. Suggest a method to control how strongly we want to impose this constraint.

3. **Loss Functions for Logistic Regression Models** : Given two models, we want to select the best model in terms of the loss function. Both of the models are a logistic regression model, but with a different architecture. The models are created by a function $g(\cdot) \to [0,1]$ as follows:

$$\hat{P}(y|x;w) = g\left(w_0 + \sum_{n=1}^{M} w_n \phi_n(x)\right),$$

where $x \in \mathbb{R}$ is the input of the model.

$$\text{Model 1: } \phi_i^{(1)}(x) = \begin{cases} x^2 & i = 1 \\ 0 & i > 1 \end{cases}, \qquad \text{Model 2: } \phi_i^{(2)}(x) = \begin{cases} x & i = 1 \\ cos(x) & i = 2 \\ 0 & i > 2 \end{cases}.$$

(a) Given N training samples $(x_i, y_i), i \in \{1, 2, ...N\}$, we evaluate the MSE risk function score of the two models. Which is better in terms of the risk function score? Model 1, model 2 or neither? Explain your answer.

(b) Define the Baysian Information Criteria (BIC) as follows:

$$BIC = -2 \times LL(N) + \log(N) \times k,$$

where N is the number of samples, $LL(N)$ is the log-likelihood as a function of N, and k is the number of parameters in the model. This criterion measures the trade-off between model fit and complexity of the model.
Let $LL_1(N)$ be the log-probability of the labels that model 1 predicted to N training samples, where the probabilities are evaluated at the maximum likelihood setting of the parameters. Let $LL_2(N)$ be the corresponding log-probability for model 2. We assume here that $LL_1(N)$ and $LL_2(N)$ are evaluated on the basis of the first N training examples from a much larger set.
Our empirical studies has shown that these log-probabilities are related in the next way:

$$LL_2(N) - LL_1(N) \approx 0.001 \times N.$$

How will we select between the two models, when using the BIC score , as a function of the number of training examples? Choose the correct answer.

i. Always select model 1.
ii. Always select model 2.
iii. First select model 1. Then, for larger N, select model 2.
iv. First select model 2. Then, for larger N, select model 1.

(c) Provide an explanation for your last answer.

(d) This section does not depend on the previous ones. Let $g(\cdot)$ be the Sigmoid function and consider the Binary Cross-Entropy loss function, where the labels, $\{y_i\}_{i=1}^N$, are 0 or 1. Suppose you use gradient descent to obtain the optimal parameters $\{w_i\}_{i=0}^M$ for each model. Give the update rule to each parameter for the two models.

4. **MINE**

(a) Suppose we are given with a set of i.i.d. samples from two random variables, $\{(x_i, y_i)\}_{i=1}^n$, where $(x_i, y_i,) \sim P_{X,Y}, \quad x_i, y_i \in \mathbb{R}$ and we attempt to estimate $I(X;Y)$. The method is based on MINE $I_n(X,Y)$ as taught in class.

Which of the following experiment graphs is a possible description of the estimated MI during training (Figure 1)? Write the indices of the correct graphs in your solution and explain your choice.

(b) Now, we wish to estimate the multivariate mutual information between the d-dimensional random vectors, (X^d, Y^d), based on the algorithm and 1-dimensional samples presented in (4a), .

i. What criteria should the elements of $X^d, Y^d \in \mathbb{R}^d$ satisfy if we want to use the dateset given in question (4a)?
ii. What is the relation between the output of MINE and $I(X^d; Y^d)$ under this criteria and $n \to \infty$?

5. **Variant of MINE**

In this question we investigate an algorithm based on the mutual information neural estimator, using the following representation of mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{5}$$

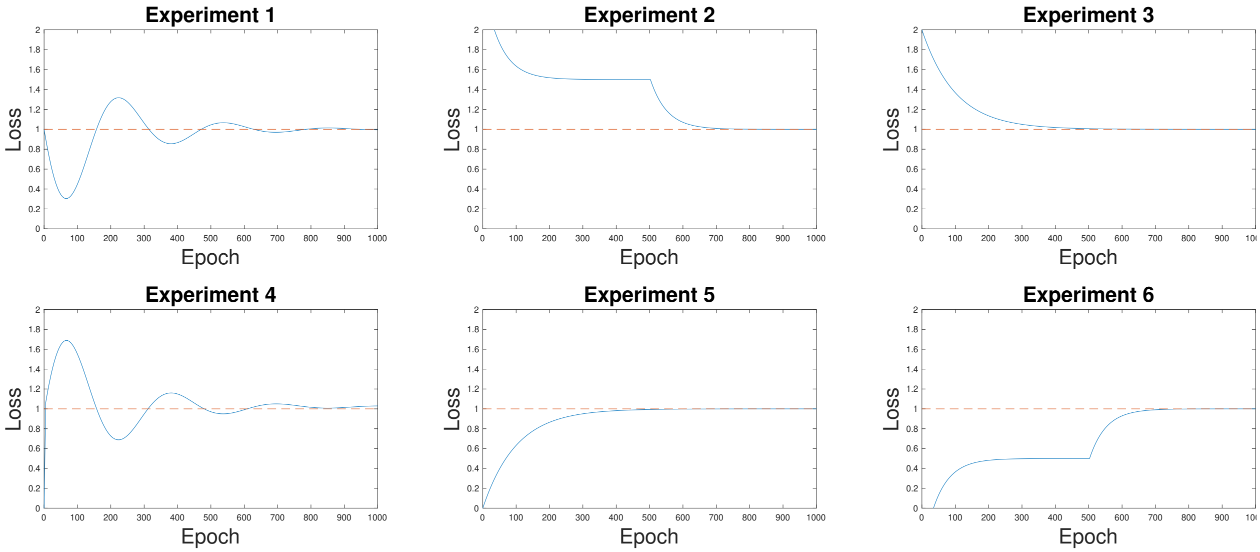

Figure 1: Experiments - filled line represents the model's loss and dashed line represents the true value.

Let $X \sim P_X$, $Y \sim P_Y$ and denote the joint PMF of (X, Y) by P_{XY}. Let U_X be the PMF of the uniform discrete probability measure over $\mathcal{X}$, the alphabet of X (namely, $U_X(x) = \frac{1}{|\mathcal{X}|} \quad \forall x \in \mathcal{X}$).

(a) Prove the following equality:

$$H(X) = H(P_X, U_X) - D_{KL}(P_X \| U_X), \tag{6}$$

where $H(P_X, U_X)$ is the cross-entropy between P_X and U_X.

(b) If we replace the uniform PMF U_X by an arbitrary PMF V_X, does Eq. (6) still hold? Prove or disprove it.

(c) Based on the result of (a), prove the following equation:

$$I(X;Y) = D_{KL}(P_{XY} \| U_{XY}) - D_{KL}(P_X \| U_X) - D_{KL}(P_Y \| U_Y), \tag{7}$$

where U_Y and U_{XY} are defined in the same sense as U_X, on $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$ respectively (assume that $|\mathcal{X} \times \mathcal{Y}| = |\mathcal{X}||\mathcal{Y}|$).

(d) Based on the KL divergence estimation method taught in class, propose an algorithm for the estimation of $I(X;Y)$ from a sample set $\{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$, based on the equality proved in (b). Denote by $\widehat{I}_n^{(H)}(X, Y)$:

i. Write the optimization objective

ii. Give a block diagram of the proposed algorithm for estimating $\widehat{I}_n^{(H)}(X,Y)$. Assume the neural network consists of a single hidden layer with M units.

(e) We now wish to calculate the optimization objective $\widehat{I}_n^{(H)}(X,Y)$. For sufficiently large n, does the following hold? explain.

$$\widehat{I}_n^{(H)}(X,Y) \leq I(X;Y) \tag{8}$$

6. **Linear regression** : Given training set $\{(x_i, y_i)\}_{i=1}^{N}$ where $(x_i, y_i) \in \mathbb{R}^2$.

First, we use a linear regression method to model this data, assume the model has no bias, i.e. b=0. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set. Now let us increase the training set size gradually.

(a) As the training set size increases, what do you expect will happen with the mean training error? explain your answer.

(b) As the training set size increases, what do you expect will happen with the mean test error? explain your answer.

Now we have prior knowledge of our dataset's distribution $y_i \sim \mathcal{N}(\log(wx_i), 1)$.

(c) We now perform a maximum likelihood estimation of w. Which of the following conditions is sufficient and necessary for a maximum likelihood estimation of w:

i. $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

ii. $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

iii. $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

iv. $\sum_i y_i = \sum_i \log(wx_i)$

(d) Provide a pseudocode (or a matlab code) for estimating w.