AI539/ECE599: S/T Statistical Learning - Winter 2025 - Dr. Raviv Raich
Homework assignment 3 (due Feb 17, 2025)

1. Numerical evaluation. Consider the problem of predicting $Y \in \{-1, 1\}$ given $X \in [0, 1]^2$ from n iid training data examples given by $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the training data points and the test data follow the same model: $(x, y) \sim f_X(x)P(y = y|X = x)$ with $f_X(x) = I(0 \leq x_1 < 1)I(0 \leq x_2 < 1)$ (i.e., uniform over the box $[0, 1)^2$) and $P(y = y|X = x)$ is unknown.

   (a) (Setting) Consider the a histogram plug-in classifier, where the probability $P(Y = y|X = x)$ is first estimated using:

   $$\hat{P}(Y = 1|X = x) = \sum_{i=1}^{m} \sum_{j=1}^{m} \hat{P}_{ij} I(x_1 \in S_i) I(x_2 \in S_j) \tag{1}$$

   where $S_i = [\frac{i-1}{m}, \frac{i}{m})$ for $i = 1, 2, \ldots, m$,

   $$\hat{P}_{ij} = \begin{cases} \frac{1}{N_{ij}} \sum_{k=1}^{n} I(y_k = 1) I(x_{k1} \in S_i) I(x_{k2} \in S_j) & N_{ij} > 0 \\ \\ 0.5, & o.w. \end{cases}$$

   and $N_{ij} = \sum_{k=1}^{n} I(x_{k1} \in S_i) I(x_{k2} \in S_j)$. This probability estimate is computing the ratio of the number of positively labeled examples in $[\frac{i-1}{m}, \frac{i}{m}) \times [\frac{j-1}{m}, \frac{j}{m})$ by the total number of examples in that bin for any $x \in [\frac{i-1}{m}, \frac{i}{m}) \times [\frac{j-1}{m}, \frac{j}{m})$. If the bin contains no examples, then a probability of 0.5 is assigned to the bin. Once the probability $P(Y = y|X = x)$ is estimated using the aforementioned $\hat{P}(Y = y|X = x)$ then the following classification rule

   $$\hat{y}(x) = \begin{cases} 1, & \hat{P}(Y = 1|X = x) > 0.5 \\ -1, & \hat{P}(Y = 1|X = x) < 0.5 \\ 2\text{Bernoulli}(0.5) - 1, & \hat{P}(Y = 1|X = x) = 0.5 \end{cases}$$

   is used for prediction.

   (b) (Numerical Experiments) Using training data available (`hw2_data.mat`), compute the risk for the classification rule given as a function of $m \in \{2, 4, 8, 16\}$ and $n \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. Since the risk depends on the classifier which depends on the data, it is a random variable. To visualize it, we will produce two plots. The first, is the average risk (i.e., the empirical equivalent of $E_{\mathcal{D}}[R[\hat{y}]]$) as a function of n for each of the m values (i.e., 4 curves). To do so, for each n, resample from the training data n points 100 times. For each of the of the resampled data (i.e., a Monte-Carlo run), obtain the classifier, and compute its empirical risk over the test data (instead of the risk). Averaging the 100 values you will obtain for $R[\hat{y}]$ should be used to estimate $E[R(\hat{y})]$. This should be repeated for each m. The second plot is going to be a scatter plot of the risk as a function of n. This is done using the same method used before to obtain a classifier and a risk value for each of the 100 Monte Carlo runs. Instead of averaging the risk values, for each n plot the risk of each of the 100 classifiers obtained value of the risk (evaluated over the test set). The result should be a plot of 100 points for each value of n. Sorting the risk values for each of n, also plot the 5th largest risk value and the 5th smallest values. This should allow you to include two curves for each m on top of the scatter plot. Those curves should provide bounds for 90% of the data points and help in understanding the typical values of the risk.