CS 688: Machine Learning
George Mason University, Fall 22 # Homework 1 **Due at 11:59pm, September 13**

Problems marked with **E** are graded on effort, which means that they are graded subjectively on the perceived effort you put into them, rather than on correctness. We strongly encourage you to typeset your solutions in LaTeX.

1. (20 points) Consider the problem of predicting whether a person has a college degree based on age and salary. The follow table contains training data for 10 individuals.

Age	Salary	College Degree
25	41,000	Yes
54	54,000	No
24	25,000	No
26	77,000	Yes
33	48,000	Yes
53	110,000	Yes
23	38,000	Yes
43	44,000	No
53	28,000	No
49	65,000	Yes

   (a) Given a threshold function $\mathbf{1}[\text{Age} \leq 50]$ that outputs 1 (Yes) when Age is less than and equal to 50, compute the zero-one loss on the training data.
   (b) For each of the features (Age and Salary), generate a plot showing the zero-one loss on training data (on the y-axis) as a function of the threshold value (on the x-axis). You will probably want to automate the calculation.
   (c) For each of the two features, report the optimal threshold function that minimizes zero-one loss on the training data.

2. (20 points, cdf and threshold classifier) Consider the sea bass/salmon with length example. Suppose we know the joint distribution of the type of fish and length is

$$\Pr[type = t, length = l] := \begin{cases} \frac{3}{4}\frac{3^l e^{-3}}{l!} & \text{if } t = 1, \text{ and } l = 0, 1, 2, \ldots, \\ \frac{1}{4}\frac{6^l e^{-6}}{l!} & \text{if } t = 0, \text{ and } l = 0, 1, 2, \ldots. \end{cases}$$

where $0! = 1$ and $l! = l \cdot (l-1)!$ for all $l = 1, \ldots$. We use $t = 1$ to denote salmon and sea bass otherwise, and assume the length l to integer.
   (a) Given a random sample from above distribution, what is the probability of the fish being salmon, $\Pr[type = 1]$?
   (b) Compute the likelihood function, $\Pr[length = l | type = 1]$ for all $l = 0, 1, 2, \ldots$.
   (c) Find an optimal threshold τ so that the expected zero one loss of the threshold function $\mathbf{1}[length \leq \tau]$ is minimized.
   (d) If we consider the following loss function: $\ell(0,1) = 1$, $\ell(1,0) = 3$, and $\ell(1,1) = \ell(0,0) = 0$ which costs 1 for labeling salmon as sea bass and 3 for labeling sea bass as salmon. Find the optimal threshold the minimized the expected loss.

3. (10 points, total variational distance and the optimal classifier) Consider the sea bass/salmon with length example. Consider $\Pr[type = 1] = 1/2$, and the likelihood function $\Pr[length = l | type = 1] = p_1(l)$, and $\Pr[length = l | type = 0] = p_0(l)$ for all $l \in \mathbb{N}_0 = \{0, 1, \ldots\}$.
   (a) Find the optimal classifier $f : \mathbb{N}_0 \to \{0, 1\}$ that minimizes the expected zero-one loss (written in terms of p_0 and p_1)
   (b) Show that the optimal expected zero-one loss is $\frac{1}{2} - d_{TV}(p_0, p_1)$ where $d_{TV}(p_0, p_1) := \frac{1}{2}\sum_{l \in \mathbb{N}_0} |p_0(l) - p_1(l)|$.

4. (20 points) In this problem, you will implement linear regression (using polynomial features) to reproduce curves of training and testing loss in class, and explore how complexity of models affect those error. The training and testing data is provided to you, named `train_autopilot.csv` and `test_autopilot.csv`. Every row in these csv files correspond to a single data point. Columns are named id, rolling speed, elevation speed, elevation jerk, elevation, roll, elevation acceleration, controller input. Your task is to predict controller input from rolling speed, elevation speed, elevation jerk, elevation, roll, and elevation acceleration. (You should remove id that is unique to each datapoint and add additional constant feature for the intercept term.)
   (a) Fit the data by applying the psuedo-inverse approach of linear regression using polynomial features x_i^j, $i = 1, \ldots, M$ with $M = 1, 2, \ldots, 6$, where x_j represents the j-th component of vector x. For example, if $x = [x_1, x_2]$ and $M = 2$, then you'd have 4 features, x_1, x_2, x_1^2, x_2^2 along with an additional constant feature 1 for the intercept term. Plot training error and test error as Root Mean Square Error (RMSE) on y-axis against M on x-axis, the order of the polynomial features.
   (b) Fit the data by using psuedo-inverse approach of regularized linear regression. For this, you need to use all polynomial features (i.e. $M = 6$), and choose values of $\ln \lambda$ using a sweep as follows: $-40, -39, \ldots, 18, 19, 20$. Plot training error and test error as Root Mean Square Error (RMSE) on y-axis against $\ln \lambda$, the regularization coefficient on x-axis.

5. (30 points) Consider a linear regression problem with weights. Specifically, given training data with feature $x_1, \ldots, x_N \in \mathbb{R}^d$ and outcome $y_1, \ldots, y_N \in \mathbb{R}$ suppose we want to find $w \in \mathbb{R}^d$ that minimizes

$$L(w) := \frac{1}{2} \sum_{i=1}^{N} r_i (w^\top x_i - y_t)^2$$

   where $r_1, \ldots, r_N > 0$ are positive weights. Note that we worked out the unweighted setting ($r_i = 1$ for all i) in the class. In this problem, we will generalize some of those ideas to the weighted setting where the weight r_i can be different for each of the training examples. (*Hint:* check your answer when $r_i = 1$ for all i)
   (a) Show that $L(w) = (Xw - Y)^\top R(Xw - Y)$ for an appropriate definition of X, Y and R with respect to x_i, y_i, r_i for $i = 1, \ldots, N$.
   (b) Recall that if all r_i's are 1, we showed that $L(w^*) = \min_w L(w)$ when $w^* = (X^\top X)^{-1} X^\top y$. By computing the derivative $\nabla_w L(w)$ and setting that to zero, generalize the above result to the weighted setting and give the new value of w^* that minimizes $L(w)$ in closed form as a function of X, R, and Y.
   (c) Suppose we have a training set $(x_i, y_i)_{i=1,\ldots,N}$ where x_i are fixed and the distribution of y_i is

$$p(y_i \mid x_i; w) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(y_i - w^\top x_i)^2}{2(\sigma_i)^2}.$$

   In other words, y_i has mean $w^\top x_i$ and variance σ_i^2. Show that finding the maximum likelihood estimate of w reduces to solving a weighted linear regression problem. State clearly what r_i's are in terms of σ_i's.