Homework Assignment # 3

Due: Friday, April 13, 2018, 11:59 p.m.
Total Points: 60

This assignment is designed to help you better understand concepts that were presented during class. You may work in small groups (2 or 3 individuals) on the homework assignments, but it is expected that you submit your own work. Therefore, **each student is responsible for submitting their own homework**, whether they work in groups or individually. The submitted homework must include the name of other individuals you worked with (if applicable), as well as all resources that were used to solve the problem (e.g. web sites, books, research papers, other people, etc.). Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Your assignment must be submitted as a single pdf document on Canvas. The questions and your answers must be typed; for example, in Latex, Microsoft Word, Lyx, etc. Images may be scanned and inserted into the document if it is too complicated to draw them properly. Be sure to show all work. **All assignments must be submitted on time to receive credit.** No late work will be accepted, unless you have a prior arrangement with the instructor.

Question 1. [10 POINTS]

The standard linear regression model is: $y = Xw + e$, where X is an $n \times d$ matrix of predictor variables, y is an $n-$dimensional vector of response variables, and $e \sim \mathcal{N}(0, \sigma^2 I)$ is an $n-$dimensional vector of model errors.

(a) What is the PDF of y in terms of X, w, σ^2?

(b) Let the PDF from part (a) be denoted as $f(y|w)$. Suppose also in this case that $w \sim \mathcal{N}(0, \rho^2 I)$. Write an expression for the joint PDF of w and y (e.g. $f(w, y)$).

(c) Show that

$$\hat{w}_{MAP} = \arg\min_w \frac{||w||^2}{\rho^2} + \frac{||y - Xw||^2}{\sigma^2} \tag{1}$$

Question 2. [10 POINTS]

Suppose we have a collection of observations $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i and y_i are real valued. We want to estimate y_i by

$$\hat{y}_i = w_0 + w_1 x_i + w_2 x_i^3 \tag{2}$$

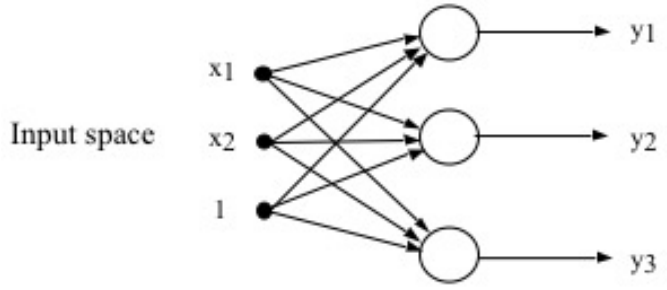
for some coefficients w_0, w_1, w_2. We will choose the coefficients so that $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is minimized. Give an expression for the optimal coefficients w_0, w_1, w_2 as a function of $(x_1, y_1), \ldots, (x_n, y_n)$. Be sure to define precisely any matrices and/or vectors you use. **You do not have to redo any derivations, but clearly justify your answer.**

Question 3. [10 POINTS]

Given the following 3-class classification problem:

C_1: {(4,1), (2,3), (3,5), (5,4), (1,6)}
C_2: {(0,2), (-2,2), (-3,2), (-2,4)}
C_3: {(1,-2), (3,-2)}

and the following single layer perceptron:



(a) Can the net learn to separate the samples, given that you want: if $\mathbf{x} \in C_i$ then $y_i = 1$ and $y_j = -1$ for $j \neq i$. No need to solve for the weights, but justify your answer.

(b) Add the sample (-1,6) to C_1. Repeat part (a).

Question 4. [10 POINTS]

For the following training samples:

$\mathbf{x}_1 = (0,0)^T \in C_1$
$\mathbf{x}_2 = (0,1)^T \in C_1$
$\mathbf{x}_3 = (1,0)^T \in C_2$
$\mathbf{x}_4 = (1,1)^T \in C_2$

(a) Plot them in input space.
(b) Apply the perceptron learning rule to the above samples one-at-a-time to obtain weights that separate the training samples. Set η to 0.5. Work in the space with the bias as another input element. Use $\mathbf{w}(0) = (0,0,0)^T$. Show values for $\mathbf{w}$ after it is updated for each training sample.
(c) Write the expression for the resulting decision boundary from (b).
(d) XOR. For $\mathbf{x}_2, \mathbf{x}_3 \in C_1$ and $\mathbf{x}_1, \mathbf{x}_4 \in C_2$, describe your observation when you apply the perceptron learning rule following the same procedure as in (a)-(c).

Question 5. [10 POINTS]

The $(\boldsymbol{x}, y)$ relationship of a Gaussian-based RBF network is defined by:

$$y(i) = \sum_{j=1}^{K} w_j(n) e^{-\frac{1}{2\sigma^2(n)}||\boldsymbol{x}_i - \boldsymbol{u}_j(n)||^2}, \quad i = 1, 2, \ldots, n \tag{3}$$

where $\boldsymbol{u}_j(n)$ is cluster center of the j^{th} Gaussian, the width $\sigma(n)$ is common to all K Gaussian components, and $w_j(n)$ is the linear weight assigned to the output of the j^{th} Gaussian component. These parameters are measured at time n. The cost function used to train the network is defined by:

$$E = \frac{1}{2} \sum_{i=1}^{n} (d(i) - y(i))^2 \tag{4}$$

where $d(i)$ is the desired output. The cost function is a convex function of the linear weights in the output layer, but non-convex with respect to the centers and width of the Gaussian units.

(a) Evaluate the partial derivatives of the cost function with respect to each of the network parameters $w_j(n)$, $\boldsymbol{u}_j(n)$, and $\sigma(n)$.

(b) Use the gradients obtained in part (a) to determine update formulas for all the network parameters, assuming learning rates of η_w, η_u, and η_σ for the network parameters, respectively. Assume that the underlying update rule is the same for convex and non-convex functions.

Question 6. [10 POINTS]

Consider a Support Vector Machine (SVM) and the following training data from two categories:

w_1:	$(1,1)^T$	$(2,2)^T$	$(2,0)^T$
w_2:	$(0,0)^T$	$(1,0)^T$	$(0,1)^T$

(a) Plot these six training points, and construct by inspection the weight vector for the optimal hyperplane, and the optimal margin.

(b) What are the support vectors?