Homework **2: Linear Classifiers**

140 marks total
This assignment is to be done individually.

Question 1 (30 marks)

Use the perceptron learning algorithm with η = 1 to compute a plane that linearly separates A+ and A- defined in the below:

A+ = {(1,1),(0,2),(3,0)} A-= {(-2,-1),(0,-2)}

(a) Let w(0) be the zero vector. Modify the algorithm so that b is computed together with the components of w. This can be done by adding an extra component of "1" to each training vector, and then computing w' = (w1, w2, b). Draw the obtained separating line along with the dataset.

(b) Solve this problem using the Least Square method. Draw your obtained line in the drawing of part (a) and compare your lines.

(c) Solve this problem using the Fisher method and draw the separating line as explained in (b).

Question 2 (20 marks)

Given a set of data points $\{\mathbf{x}_n\}$, we can define the *convex hull* to be the set of all points **x** given by

$$x = \sum_n \alpha_n x_n$$

Where $\alpha_n \geq 0 \ \ and \ \ \sum_n \alpha_n = 1$.
Consider a second set of points $\{\mathbf{y}_n\}$ together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector **w** and a scalar w_0 such that $w^T x_n + w_0 > 0$_ for all $\mathbf{x}_n$, and _$w^T y_n + w_0 < 0$ r for all $\mathbf{y}_n$.
Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely if they are linearly separable, their convex hulls do not intersect (20 Mark).

Question 3 (30 marks)

In many real-world scenarios, our data has millions of dimensions, but a given example has only hundreds of non-zero features. For example, in document analysis with word counts for features, our dictionary may have millions of words, but a given document has only hundreds of unique words. In this question, we will make l2 regularized (stochastic gradient descent (SGD) efficient when our input data is sparse. Recall that in l2 regularized logistic regression, we want to maximize the following objective (in this problem we have excluded w0 for simplicity):

$$F(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N} l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) - \frac{\lambda}{2}\sum_{i=1}^{d} \mathbf{w}_i^2$$

$$l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) = y^{(j)}(\sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i^{(j)}) - \ln(1 + \exp(\sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i^{(j)}))$$

Where $l(x^{(i)}, y^{(j)}, w)$ is the logistic objective function and the remaining sum is our regularization penalty.
When we do stochastic gradient descent on point $(x^{(i)}, y^{(j)})$ are approximating the objective function as

$$F(\mathbf{w}) \approx l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) - \frac{\lambda}{2} \sum_{i=1}^{d} \mathbf{w}_i^2$$

Definition of sparsity: Assume that our input data has d features, i.e. $x^{(j)} \in R^d$. In this problem, we will consider the scenario where $x^{(j)}$ is sparse. Formally, let s be the average number of nonzero elements in each example. We say the data is sparse when s << d. In the following questions, your answer should take the sparsity of $x^{(j)}$ into consideration when possible.

(a) Let us first consider the case when λ = 0. Write down the SGD update rule for wi when λ = 0, using step size η given the example $(x^{(i)}, y^{(j)})$. (10 Mark)
(b) Now Let us consider the case when λ > 0. Write down the SGD update rule for wi when λ > 0, using step size η given the example $(x^{(i)}, y^{(j)})$. (10 Mark)
(c) If we use a dense data structure, what is the average time complexity to update wi when λ = 0? What if we use a sparse data structure? Justify your answer in one or two sentences. (10 Mark)

Question 4 (60 marks) (implementation)

You were given a dataset "dataset_i.csv" (extracted from the IRIS dataset) including 2 features and 3 classes.

(a) Consider only the first two classes (class 0, and class 1). Perform Fisher, perceptron, and least square classification methods for these two classes. Don't use ML libraries for classifiers and implement them yourself. Draw the obtained separating lines and dataset and compare among 3 resulting lines.
Also, calculate the accuracy of classifiers and compare them in a table. (15 Mark)

(b) Consider two latter classes (class 1, and class 2), Perform Fisher, perceptron, and least square classification methods for these two classes. Don't use ML libraries for classifiers and implement them yourself. Draw the obtained separating lines and dataset and compare among 3 resulting lines.
Also, calculate the accuracy of classifiers and compare them in a table. (15 Mark)

(c) Comparing (a) and (b), analyze the results and conclude in two or three sentences (10 Mark)

(d) Do the parts (a) and (b) for the logistic classifier. In this case, you can use ML libraries that exist for logistic classifiers. Compare the accuracy of this classifier with parts (a) and (b). (10 Mark)

(e) Perform classification for all three classes using the least square method and calculate the accuracy (10 Mark)