STAT 8178/7178

Instructions:

This assignment covers weeks 1, 2 and 3. Each question is worth 15%.

1. Due on 17th March 2023
2. For all the questions please provide the relevant mathematical derivations, the computer programs (only using R software) and the plots.
3. Please submit on iLearn a single PDF file containing all your work (code, computations, plots, etc.). Other file formats (e.g. Word, html) will NOT be accepted.
4. Try to use Rmarkdown through Rstudio. But it is not compulsory to use Rmarkdown even if facilitate to reproduce results. Only upload the pdf file.

1. Question [15 Marks]

Consider the following model for the binary classification task with outcome $Y \in \{0,1\}$. There are p features $X_1, \ldots, X_p$. The model is,

$$\mathbb{P}(Y = 1 \mid X_1, \ldots, X_p) = \sigma(w^T X),$$

where $\sigma(\cdot)$ is the sigmoid function, $w = (w_1, \ldots, w_p)^T$ is the vector containing the regression coefficients and $X = (X_1, \ldots, X_p)^T$ is the vector containing the p features. Note that we do not include any intercept/bias in this model.

Consider we observe m samples $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ with $x^{(i)} \in \mathbb{R}^p$ and $y^{(i)} \in \{0,1\}$. To estimate the parameter of the model, we propose two different loss function:

- (1) Based on the binary cross-entropy loss

$$C_1(w; \text{Data}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

where $\mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -(y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$ and $\hat{y}_i = \sigma(w^T x^{(i)})$

- (2) Based on the squared loss

$$C_2(w; \text{Data}) = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

(a) 2 marks Define the gradient of the loss function $C_1(w; \text{Data})$

(b) 2 marks Define the gradient of the loss functions $C_2(w; \text{Data})$

(c) 2 marks Define the hessian matrix of the loss function $H_1 = \nabla^2(C_1(w; \text{Data}))$

(d) 2 marks Define the hessian matrix of the loss function $H_2 = \nabla^2(C_2(w; \text{Data}))$

(e) 2 marks Show that the function $C_1(w; \text{Data})$ is convex. (Hint: show that the hessian matrix is positive semidefinite)

(f) 2 marks Show that the function $C_2(w; \text{Data})$ is not convex.

(g) 0.5 marks Which loss function you suggest for training your model ?

Consider now only two features from this model and we have observed $m = 5$ samples $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(5)}, y^{(5)})\}$ with $x^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{0,1\}$. The model reduces to:

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \sigma(w_1 X_1 + w_2 X_2).$$

The observed data are: $x^{(1)} = (1, -3)^\top, x^{(2)} = (1, -2)^\top, x^{(3)} = (1, -1)^\top, x^{(4)} = (1, 1)^\top, x^{(5)} = (1, 2)^\top, x^{(6)} = (1, 3)^\top$, and $y^{(1)} = 1, y^{(2)} = 0, y^{(3)} = 0, y^{(4)} = 1, y^{(5)} = 1, y^{(6)} = 0$.

(h) 1 marks Provide a 3D plot representing the loss landscape of $C_1(w; \text{Data})$ for values of $w_1 \in [-10; 10]$ and $w_2 \in [-5; 5]$.

(i) 1 marks Provide a 3D plot representing the loss landscape of $C_2(w; \text{Data})$ for values of $w_1 \in [-10; 10]$ and $w_2 \in [-5; 5]$.

(j) 0.5 marks What do you conclude from these previous plots ?

2. **Question 2:** [15 Marks]

   In this question you will use a logistic regression model for building a classifier. We consider the data from the file `BinaryClassifier.csv` (available on ilearn page). The first 456 rows of this dataset are used for training the model while the reminder rows are used for the test dataset. The first two columns are two features which we will name x_1 and x_2. The third column is the binary outcome y (0 or 1 values).

   (a) 1 marks. Load the dataset (available on Ilearn) and define well the train and test set. Carry out a statistical comparison of your choice between the distribution of the two classes in both the training set and the test set, aiming to show that the sets have been well randomly chosen.

   (b) 2 marks. Fit a logistic model to the training set using a generalized linear model (using `glm` function) to create a binary classifier using the train data.

   (c) 2 marks. Evaluate the performance of your classifier on the test set. You should provide the confusion matrix as well as the F_1 score.

   (d) 5 marks. Now using first principles (not using any specific packages), build a binary classifier using the sigmoid function on the linear combination of the features (including a bias term). You will estimate your parameter by exploiting the cross-entropy loss. Remember that it is equivalent to the logistic model. You will use your own batch gradient descent algorithm for optimizing your cost function. Provide at least two R functions:

      i. A first function for getting the estimates of your model. Some arguments of your function might be: the initial start values of the parameters, a data matrix containing features and response variable, the tolerance for your stoping rule, the maximum number of iterations, the learning rate, ...

      ii. A second function for classifying new data points.

   (e) 2 marks. Train your model using the training data. Provide a plot of the loss function during training to illustrate convergence of your model. You might try different learning rate. To help the convergence of your model it is recommended to *standardize* your train data (see page 32 of "Notes-on-Machine-Learning-Chapt" in ilearn).

   (f) 2 marks. Evaluate the performance of your classifier on the test set. You should provide the confusion matrix and F_1 score and compare with the results of item 3 above. **Be aware that if you *standardize* the train set, you have to *standardize* the test set as well using the sample means and the standard deviations computed from the training set**.

   (g) 1 marks. Compare your previous results with those found in (c).