Assignment: Logistic Regression

Write jupyter notebook scripts for the following questions. Use print out statements and markdown cells to show and explain your results. Submit your notebooks to canvas.

1. **Logistic regression**

   (a) (10 points) Do natural log transform of the PFOS variable in file `pfas.csv` and store the results as a new variable log_PFOS in the data file. Standardize the variables x=[log_PFOS, age, gender, BMI].

   (b) (35 points) Use y=`disease` and the standardized x=`[PFOS, age, gender, BMI]` to write and debug your own gradient descent algorithm for logistic regression. Your algorithm should export the learned parameters in the θ vector. Note that you can modify the gradient descent algorithm that you have written for the linear regression algorithm to achieve logistic regression.

   (c) (10 points) Apply your own algorithm to the standardized data and provide the values of the learned θ.

   (d) (10 points) Apply LogisticRegression in sklearn to the y and the standardized x. What are the θ values you get from sklearn? Information about how to apply LogisticRegression in sklear can be found at

   `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

   (e) (10 points) Add constant to the standardized x using the function `add_constant`. Instructions about how to use `add_constant` can be found at:

   `https://www.statsmodels.org/dev/generated/statsmodels.tools.tools.add_constant.html`

   Apply Logit in statsmodels to the data with constant 1 added. What θ do you get? Instructions about how to use statsmodels to do logistic regression can be found at:

   `https://www.statsmodels.org/stable/generated/statsmodels.formula.api.logit.html`

   (f) (25 points) Compare θ from your own algorithm, θ from LogisticRegression in sklearn, and θ from statsmodel. Do you get very similar results? If not, what could you do to make the θ values similar?