



PHYS 467- Homework 1

In this Homework, you will perform linear classification on the SUSY dataset. The aim is to predict whether a process (characterized in the dataset by 18 features) corresponds to a background process, or a signal process which produces supersymmetric particles. Return **a single Jupyter notebook** with all the plots *already generated*. The answers to the theoretical questions should also be in the notebook, as Mark-down cells.

1. Download the SUSY dataset, for example from <http://mlphysics.ics.uci.edu/data/susy/>. The dataset is quite large, so it may take some time.
2. (1pt) Using the `pandas.read_csv` function ( put the options `index_col, header` to `None`), open the dataset and print it. How many features are there? How many data points? Which column corresponds to the label?
3. (1pt) As you notice, the labels take value 0 or 1. To connect with the lecture, transform the labels so that they take value $-1, +1$.
4. (1pt) Split the dataset into train, test and validation sets. You can use `train_test_split` from `sklearn.model_selection`. Each set should correspond to respectively 60%, 20%, 20% of the total data.
5. (1pt) Using the `sklearn.preprocessing.StandardScaler` method, preprocess the data. You can refer to the online documentation or Exercise 3 on Moodle for further instruction on how to use it. Fit the scaler on the train set. Why don't we fit it on the test set? On the validation set? Justify briefly in less than two lines.
6. (2pt) We will consider two linear classifiers : the **ridge classifier** (square loss function) and **logistic regression**. In both cases, we do *not* fit an intercept and we add some ℓ_2 regularization of strength λ . For the ridge classifier, implement yourself a function which takes the train set and the train labels, and returns the weights of the classifier, using the closed-form formula seen in class.
7. (3pt) For $\lambda = 0.1$, run ridge classification on the dataset, and print the accuracy, defined as the fraction of the test data whose label is correctly predicted.
8. (2pt) Now turn to logistic regression. You will use the sklearn implementation https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. The model is fully explained in the User Guide https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression, section 1.1.11.1. Is the expression the same as the one you saw in class? Why is that? Show that the two are actually equivalent. *Hint*: look in the documentation: which values should the input train labels take for the sklearn implementation?
9. (4pt) For $\lambda = 0.1$, run logistic regression on the dataset, and print the accuracy.  Be careful how the regularization is defined in sklearn!
10. (5pt) We will now select the best value of λ . For λ in `np.logspace(-1, 4, 10)`, print the train and validation accuracies of both ridge classification and logistic regression with regularization λ . Which regularization and loss would you choose?