AI6012: Machine Learning Methodologies & Applications Assignment

Important notes: to finish this assignment, you are allowed to look up textbooks or search materials via Google for reference. NO plagiarism from classmates is allowed.

The submission deadline is by 11:59 pm, Oct 2, 2020 (19 days). And the file to be submitted is a single PDF (no source codes are required to be submitted). Multiple submission attempts are allowed, and the last one will be graded. A submission link is available under "Assignments" of the course website in NTULearn.

Question 1: Consider a multi-class classification problem of C classes. Based on the parametric forms of the conditional probabilities of each class introduced on the 54th Page ("Extension to Multiple Classes") of the lecture notes of L4, derive the learning procedure of logistic regression for multi-class classification problems.

Hint: define a loss function by borrowing an idea from binary classification, and derive the gradient descent rules to update $\{\mathbf{w}^{(\mathbf{c})}\}$'s.

Question 2: This is a hands-on exercise to use the LinearSVC API of scikit-learn[1] to train a linear SVM on a binary classification dataset. The details of instructions are described as follows.

1. Download the a5a dataset from the LIBSVM Dataset page.

   This is a preprocessed dataset of the Adult dataset in the UCI Irvine Machine Learning Repository[2], which consists of a training set (available here) and a test set (available here).

   Each file (the train set or the test set) is a text format in which each line represents a labeled data instance as follows:

label index1:value1 index2:value2 ...

   where "label" denotes the class label of each instance, "indexT" denotes the T-th feature, and valueT denotes the value of the T-th feature of the instance. This is a sparse format, where only non-zero feature values are stored for each

[1] Read Pages 65-66 of the lecture notes of L5 for reference

[2] The details of the original Adult dataset can be found here.

instance. For example, suppose given a data set, where each data instance has 5 dimensions (features). If a data instance whose label is "+1" and the input data instance vector is [2 0 2.5 4.3 0], then it is presented in a line as

```
+1  1:2  3:2.5  4:4.3
```

Hint: sciki-learn provides an API ("sklearn.datasets.load_svmlight_file") to load such a sparse data format. Detailed information is available here.

2. Given a set of 5 candidate values of the parameter C in LinearSVC, $\{0.01, 0.1, 1, 10, 100\}$, use 3-fold cross-validation to determine which is the best value of C in terms of classification accuracy on the a5a **training set**. Generate the following table based on your experimental results. Note that for all the other parameters in LinearSVC, you can simply use the default values. If you set them to be other values, specify them in your submitted PDF file.

Table 1: The 3-fold cross-validation results of varying values of C in LinearSVC on the a5a training set (in accuracy).

	$C = 0.01$	$C = 0.1$	$C = 1$	$C = 10$	$C = 100$
Accuracy of linear SVMs	?	?	?	?	?

Hint: there are no specific functions of cross-validation for SVMs in sciki-learn. However, you can use some APIs under the category "Model Selection $\rightarrow$ Model validation" to implement it. Some examples can be found here.

3. Set the parameter value of C to the best one found in the previous step, use LinearSVC to train a model using the whole a5a training set, and make predictions on the a5a **test set**. List the result in terms of accuracy in the following table, where c^* is the best value of C based on cross-validation in the previous step.

Table 2: Test results of LinearSVC with the best value of C on the a5a test set (in accuracy).

	$C = c^*$
Accuracy of linear SVMs	?

Question 3 (optional): Using the kernel trick introduced in L5 to extend the regularized linear regression model to solve nonlinear regression problems. Derive a closed-form solution.