CS 5350/6350: Machine Learining Fall 2017

Max Taggart, Homework 5

November 29, 2017

1 Logistic Regression

1. [5 points] What is the derivative of the function $g(\mathbf{w}) = \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$ with respect to the weight vector?
   The derivative with respect to w is

$$\frac{1}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}(-y_i \mathbf{x}_i) \exp(-y_i \mathbf{w}^T \mathbf{x}_i)$$

2. [5 points] The inner most step in the SGD algorihtm is the gradient update where we use a single example instead of the entire dataset to compute the gradient. Write down the objective where the entire dataset is composed of a single example, say $(\mathbf{x}_i, y_i)$. Derive the gradient with respect to the weight vector.
   The objective, $J(\mathbf{w})$, for a single example is

$$J(\mathbf{w}) = \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{1}{2\sigma^2} \mathbf{w}^T \mathbf{w}$$

   The gradient with respect to the weight vector is

$$\frac{dJ}{d\mathbf{w}} = \frac{1}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}(-y_i \mathbf{x}_i) \exp(-y_i \mathbf{w}^T \mathbf{x}_i) + \frac{1}{2\sigma^2} \mathbf{w}$$

3. [10 points] Write down the pseudo code for the stochastic gradient algorithm using the gradient from previous part.

   for i in testSet:
   gradient = $\frac{dJ}{d\mathbf{w}}(y_i, \mathbf{x}_i, \mathbf{w})$
   $\mathbf{w}$ = $\mathbf{w}$ - gradient

2 Experiments

2.1 Algorithm Implementation

All of the algorithms were implemented using python2.7. The feature matrices were represented using sci-kit's csrmatrix() class to minimize the memory footprint and the number of

cpu-operations required for each matrix-vector operation.
Each model was represented using a python class implementing a fit() and predict() method similar to sci-kit learn's API. Cross validation splits were generated using sci-kit learn's StratifiedKFold class with the random state fixed at 0. The grid search was executed using simple nested for-loops where the parameters that produced the top average accuracy were maintained and used to re-fit the model on the whole training set.
The number of training epochs for the SVM and logistic regression models was selected as 3 after evaluating the performance for epochs in the range 1-5.

2.2 Results

1. Results:

	Best hyper-parameters	Average cross-validation accuracy	Training accuracy	Test Accuracy
SVM	Epochs: 3, C: 10, rate: 0.0001	0.77599	0.86267	0.81702
Logistic regression	Epochs: 3, Tradeoff: 100.0, rate: 0.001	0.77634	0.84847	0.80000
Naive Bayes	Smoothing: 0.5	0.68985	0.72746	0.67659
Bagged Forests	Depth: 10	NA	0.79382	0.68510
SVM over trees	Epochs: 3, Depth: 10, C: 10, rate: 0.001	0.60320	0.72924	0.66383
Logistic regression over trees	Depth: 10, rate: 0.001, Tradeoff: 1000.0	0.54964	0.59439	0.57446

Table 1: Result table