CMSC726 : Machine Learning *Ravi Bhadeshiya*

HW2 - LR

1. What is the role of the learning rate?

   In order to work SGD (Stochastic Gradient Descent) learning rate must be set to appropriate value. The large value of learning rate cause the oscillation in accuracy and small value of it cause to slow convergence.Effect of learning rate over the test data, $0.5 \rightarrow 93.2331\%$, $0.1 \rightarrow 94.7368\%$, $0.01 \rightarrow 91.7293\%$.

2. How many passes over the data do you need to complete?

   For the multiple passes testing accuracy are almost same. So, the evidence seems to provide that multiple passes over the data is unnecessary.

3. What words are the best predictors of each class? How (mathematically) did you find them?

   The best feature can be found using the weight values. The max positive values of weights are the best for the positive class and the min negative values of weights are the best for the negative class.It can be found by sorting of weight arrays.

   (a) **Best features for Baseball class:** ['runs', 'baseball', 'hit', 'pitching', 'catcher', 'ball', 'book', 'rickert', 'stance', 'bat']

   (b) **Best features for Hockey class:** ['hockey', 'playoffs', 'golchowy', 'goals', 'pick', 'next', 'ice', 'biggest', 'names', 'playoff']

4. What words are the poorest predictors of classes? How (mathematically) did you find them?

   The worst feature are not significant,thus it values are almost zero. Mathematically, it can be found by sorting of absolute value in weight arrays.

   - **Worst features:** ['riel', 'broad', 'racist', 'vintage', 'blasted', 'intermissions', 'memoriam', 'deceased', 'rode', 'hesitate']

Extra Credits:

5. Effect of using a schedule to update the learning rate.

   Adapting the learning rate for each iteration seems to improve the performance. Following accuracy on testing dataset, • Time based decay : 0.924812 • Exponential decay : 0.932331 • Asymptotically decay : 0.939850
   This type of learning rate seems to reduce over-fitting on training data with multiple passes.

6. Effect of modifying the feature values to tf-idf.

   Use of tfidf seems to reduce over-fitting with multiple passes but it changes the best and worst predictors. However, convergence is slow.