**Name: Your Name**

**Andrew ID: Your Andrew Id**

**Nickname: Your Nickname on the Leaderboard**

**Machine Learning with Graphs**

**Homework 2 – Template**

1. **Statement of Assurance**

*You must certify that all of the material that you submit is original work that was done only by you. If your report does not have this statement, it will not be graded.*

1. **Data Preprocessing**
2. **[5 pts]** After your finishing the data preprocessing, report the top 9 frequent tokens and corresponding counts in the report.

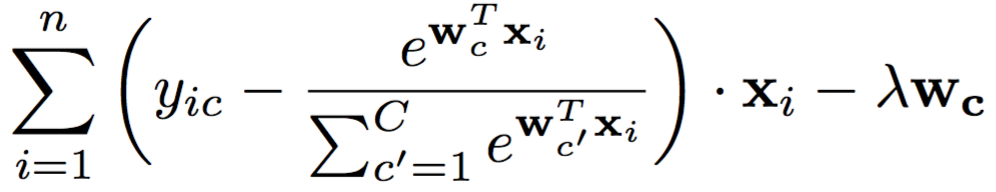
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rank | Token | Count | Rank | Token | Count | Rank | Token | Count |
| No. 1 |  |  | No. 2 |  |  | No. 3 |  |  |
| No. 4 |  |  | No. 5 |  |  | No. 6 |  |  |
| No. 7 |  |  | No. 8 |  |  | No. 9 |  |  |

1. **[5 pts]** Before continuing to the next step, another interesting problem is to check the star distribution of training samples. Report the count of training samples for each star (i.e., 1 to 5).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Star | 1 | 2 | 3 | 4 | 5 |
| # of training data |  |  |  |  |  |
| Percentage |  |  |  |  |  |

Do you find something unexpected from the distribution (e.g., whether the dataset is balanced)? Will this be a problem in training the model? If so, could you give some idea about how to address it and explain why your idea should work?

1. **Model Design**
2. **[5 pts]** Show that the gradient of regularized conditional log-likelihood function with respect to the weight vector of class (i.e., ) is equal to



Notice that the gradient of log-likelihood function with respect to a vector is itself a vector, whose -th element is defined as , where is the -th element of vector .

1. **[5 pts]** Let the learning rate be , outline the algorithm (Batched-SGD) for implementation. You should cover how would you like to update the weights in each iteration, how to check the convergence and stop the algorithm and so on.
2. **[10 pts]** After implementing your model, please use these two types of prediction to calculate and report the Accuracy and RMSE (See definition in Evaluation part) on the entire training set with the two features designed in Task 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | CTF | | DF | |
| Dataset | Training | Development | Training | Development |
| Accuracy |  |  |  |  |
| RMSE |  |  |  |  |
| Parameters Setting | Learning Rate alpha=? Regularization Parameter lambda=?  How many iterations used? Batch size=? | | | |

**[10 pts] Multi-class Support Vector Machine**

After you figure them out, report only the accuracy on the training and development set using the two features designed in Task 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | CTF | | DF | |
| Dataset | Training | Development | Training | Development |
| Accuracy |  |  |  |  |
| Parameters Setting | All parameters you used to run SVM. If you run SVM in terminal, include your command line here. | | | |

1. **Feature Engineering**

**[10 pts]** Describe in details your most satisfying design and the corresponding considerations, use formula to illustrate your idea if necessary. Besides, report the evaluation results on training and development set here (The reported result here should match the record on the leaderboard).

1. **One sentence of your feeling for this homework**

Is that good or not? Why?