**Mini-Lab**

**Create Models (50 points)**

*“Create a logistic regression model and a support vector machine model for the classification task involved with your dataset. Assess how well each model performs (use 80/20 training/testing split for your data). Adjust parameters of the models to make them more accurate. If your dataset size required the use of stochastic gradient descent, then linear kernel only is fine to use. That is, the SGD classifier is fine to use for optimizing logistic regression and linear support vector machines. For many problems, SGD will be required in order to train the SVM model in a reasonable timeframe.”*

\*\*This 50 points appears that it’s all based on the code – doesn’t look like it needs any write-up (other than the commented code).\*\*

**Model Advantages (10 points)**

*“Discuss the advantages of each model for each classification task. Does one type of model offer superior performance over another in terms of prediction accuracy? In terms of training time or efficiency? Explain in detail.”*

Logistic Regression is more time efficient. SVM was very time intensive and with a larger dataset, would be unrealistic. The SVM took approximately 25 minutes to run.

**Interpret Feature Importance (30 points)**

*“Use the weights from the logistic regression to interpret the importance of different features for the classification task. Explain your interpretation in detail. Why do you think some variables are more important?”*

Logistic Regression:

SVM: The weights of the two models are similar. The SVM model would probably include MaxTemp (0.2715) in the model. The logistic regression did not (MaxTemp=0.1097). (I don’t have a reason why, just something notably different)

**Interpret Support Vectors (10 points)**

*“Look at the chosen support vectors for the classification task. Do these provide any insight into the data? Explain. If you used stochastic gradient descent (and therefore did not explicitly solve for support vectors), try subsampling your data to train the SVC model – then analyze the support vectors from the subsampled dataset.”*

\*\*Here’s where I was trying find info. Did it choose all of our variables to report on? How do we know which ones are the ‘chosen support vectors’. :-/ \*\*