Kernel Project Report

Team: Practically Catastrophic Accuracy
Avetik Karagulyan (MVA), Aria Akhavan-Foomani (MASH)
and Amir-Hossein Bateni (MASH)

We measured the performance of each method by a 5 fold cross validation before submitting its result over test data on Kaggle.

We started by working on the prepared extracted data. We simply applied linear logistic kernel and linear kernel SVM, and the result was near 50 percent. we implemented the linear kernel SVM by an optimization solver from the library *cvxopt*.

Then, we decided to implement a new kernel. Our first choice was the Spectrum kernel. The classification method applied on this kernel was linear kernel SVM by the same implementation as before. We tried it with different length (l) of sub-string. The best performance was achieved for $l = 7$ with a submitting score around 70 percent.

We continued with this method by working more on shrinkage of the features and regularization parameter of SVM by which we could get a submitting score near 74 percent. In this case, $l = 6$ was the optimized length.

So far, we had merged the three data sets, and we were applying our methods on this collection of the three data sets. But, we figured out separating the data sets and processing each of them individually improve the performance significantly which yielded to 78.266 as submitting score with same length $l = 6$.

Furthermore, we tried to implement the Sub-string Kernel. But, the slowness of the method did not allow us to achieve a clear conclusion on this kernel.

At the end, we also tried dimension reduction techniques such as kernel PCA, by which we did not obtain better results.