10701 Project Proposal

Teammates: Joshua Brakensiek, Jacob Imola, Sidhanth Mohanty

The title of our proposed project is “Semi-Supervised Learning of Pen-Based OCR.”

We plan to utilize the following two data sets “Pen-Based Recognition of Handwritten Digits Data Set” [AA98] and “UJI Pen Characters (Version 2) Data Set” [LPM$^+$08] from the UCI Machine Learning Repository [Lic13]. The data set consists vector-based representations of symbols drawn by users on a table-like input devise. Each data set has over 10,000 data points. The first data set consists entirely of drawings of the digits $0, \ldots, 9$, while the second data set uses a much richer variety of characters.

Let $\operatorname{Seq}\mathbb{R}^2$ be the set of sequences of points in $\mathbb{R}^2$, and let $\operatorname{Seq}(\operatorname{Seq}\mathbb{R}^2)$ be sequences of sequences (representing that some curves have multiple connected components). Let $\mathcal{S}$ be the set of symbols we seek to learn. The function we seek to learn is $f : \operatorname{Seq}(\operatorname{Seq}\mathbb{R}^2) \to \mathcal{S}$. The main goal of our project is too see how well this function can be learned with various semi-supervised learning methods, particularly Transductive Support Vector Machines and ANOTHER METHOD. We will partition the data into three groups, a labeled training set of size 20%, a labeled validation set of size 5%, and an unlabeled portion of size 75% for which we will disregard the given labels. We plan to implement our learning algorithms in Python 2.x, building on the NumPy and SciPy libraries (and possibly other libraries as we see fit).

Our goal by April 6, 2016, is to implement a parser for the data sets, partition the data into the categories, decide on features of the pen-based data to test with various learning algorithms, and implement a basic transductive supper vector machine, and test it with cross-validation. By the final deadline, we hope to have implemented additional semi-supervised learning methods we may also explore the tradeoff of what proportion of the data is unlabeled versus classifier accuracy. [WORK DIVISION]

We plan to read the [Zhu05] to understand various semi-supervised learning methods. We will also read [**?**] to get the basis of our TSVMs.

References

[AA98] E Alpaydin and Fevzi Alimoglu. Uci pen-based recognition of handwritten digits data set. 1998.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[LPM$^+$08] David Llorens, Federico Prat, Andrés Marzal, Juan Miguel Vilar, María José Castro, Juan-Carlos Amengual, Sergio Barrachina, Antonio Castellanos, Salvador España Boquera, JA Gómez, et al. The ujipenchars database: a pen-based database of isolated handwritten characters. In *LREC*, 2008.

[Zhu05] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.