# Midway Report: Semi-Supervised Learning of Pen-Based OCR.

**Joshua Brakensiek, Jacob Imola, Sidhanth Mohanty**

## 1 Background

Handwriting recognition is a classic machine learning problem. Generally, an image of a digit is inputted to an algorithm and classified. Instead of taking this approach, we used 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], which consist of pixel coordinates that the writers' pens took at certain time intervals. 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. We seek to explore if adding the extra information of time while sacrificing some detail from the shape of the digits can produce a viable semi-supervised model. We implement our learning algorithms in Python 2.x, using on the NumPy, SciPy, and Scikit-learn libraries (and possibly other libraries as we see fit). The models we try are Transductive Support Vector Machines and graph-based kernel methods and regularization. Throughout this paper, let $X = \{x_1, x_2, \ldots, x_k\}$ be the labeled training data, $Y = \{y_1, y_2, \ldots, y_k\}$ be the corresponding labels, $X^* = \{x_1^*, x_2^*, \ldots, x_n^*$ be the unlabeled training data, and $Y^* = \{y_1^*, y_2^*, \ldots, y_n^*$ be variables representing the unlabeled training data's labels. Now, we will give a description of the models:

Recall that an SVM reduces to the following primal problem:

$$\arg\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{k} \zeta_i$$

$$\begin{aligned} \text{subject to} \quad & \forall_{i=1}^{k} : y_i(wx_i + b) \geq 1 - \zeta_i \\ & \forall_{i=1}^{k} : \zeta_i > 0 \end{aligned}$$

A TSVM's primal problem allows the unlabeled data to take on any label, and hence reduces to a similar problem [Joa99]:

$$\arg\min_{w,b,\zeta,\zeta^*,Y^*} \frac{1}{2} w^T w + C \sum_{i=1}^{k} \zeta_i + C^* \sum_{i=1}^{n} \zeta_i^* \tag{1}$$

$$\begin{aligned} \text{subject to} \quad & \forall_{i=1}^{k} : y_i(wx_i + b) \geq 1 - \zeta_i \\ & \forall_{i=1}^{n} : y_i^*(wx_i^* + b) \geq 1 - \zeta_i^* \\ & \forall_{i=1}^{k} : \zeta_i > 0 \\ & \forall_{i=1}^{n} : \zeta_i^* > 0 \end{aligned}$$

Here, $C$ and $C^*$ are parameters that control the type of fit we get.

## 2 Related Work

(Probably could cite some of our Bibliography here)

## 3 Methods

One TSVM method we used is an SVM with local search. The user inputs $C$ and $C^*$ defined in (1). First, the $Y^*$ are initialized to be

## 4 Experiments

## 5 Future Plans

## References


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

[Joa99] T Joachims. Transductive inference for text classifcation using support vector machines. 1999.

[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.