Maximum Margin Semi-Supervised Learning for Structured Variables

Y. Altun, D. McAllester
TTI at Chicago
Chicago, IL 60637
altun,mcallester@tti-c.org

M. Belkin
Department of Computer Science
University of Chicago
Chicago, IL 60637
misha@cs.uchicago.edu

Abstract

Many real-world classification problems involve the prediction of multiple inter-dependent variables forming some structural dependency. Recent progress in machine learning has mainly focused on supervised classification of such structured variables. In this paper, we investigate structured classification in a semi-supervised setting. We present a discriminative approach that utilizes the intrinsic geometry of input patterns revealed by unlabeled data points and we derive a maximum-margin formulation of semi-supervised learning for structured variables. Unlike transductive algorithms, our formulation naturally extends to new test points.

1 Introduction

The discriminative methods, such as Boosting and Support Vector Machines have significantly advanced the state-of-the-art for classification. However, traditionally these methods do not exploit dependencies between class labels where more than one label is predicted. Many real-world classification problems, on the other hand, involve sequential or structural dependencies between multiple labels. For example labeling the words in a sentence with their part-of-speech tags involves sequential dependency between part-of-speech tags; finding the parse tree of a sentence involves a structural dependency among the labels in the parse tree. Recently, there has been a growing interest in generalizing kernel methods to predict structured and inter-dependent variables in a supervised learning setting, such as dual perceptron [6], SVMs [2, 14, 13] and kernel logistic regression [1, 9]. These techniques combine the efficiency of dynamic programming methods with the advantages of the state-of-the-art learning methods. In this paper, we investigate classification of structured objects in a semi-supervised setting.

The goal of semi-supervised learning is to learn from both labeled and unlabeled data. This idea has recently attracted a considerable amount of interest due to ubiquity of unlabeled data. In many applications from data mining to speech recognition it is easy to produce large amounts of unlabeled data, while labeling is often manual and expensive. That is also true for many structured classification problems. A variety of methods ranging from Naive Bayes [11], Cotraining [4], to Transduc-

tive SVM [7] to Cluster Kernels [5] and graph-based approaches [3] and references therein, have been proposed. The intuition behind many recent methods is that the classification/regression function should be smooth with respect to the geometry of the data. The labels of two input patterns x and $\bar{x}$ are likely to be the same if x and $\bar{x}$ are similar with respect to the intrinsic geometry of the set of input patterns. This idea is often represented as the *cluster assumption* or the *manifold assumption*. The unlabeled points reveal the intrinsic geometry, which is then utilized by the classification algorithm. A discriminative approach to semi-supervised learning was proposed in [3], providing a principled geometric method of incorporating unlabeled data into the standard kernel framework by using the Laplacian operator associated to the point cloud as an additional penalty on the space of functions in a Reproducing Kernel Hilbert Space.

In this paper, we generalize that approach to problems that involve structured and inter-dependent outputs. We derive a maximum-margin formulation that utilizes labeled and unlabeled data and leads to a quadratic program similar to the standard SVM. We present experimental results showing the validity of our methods. It is important to note that unlike some recent transductive algorithms, which only produce labelings for the unlabeled part of the data set, we do not encounter an out-of-sample extension problem. Our classification function is an outcome of a Representer theorem, similar to that for an ordinary SVM and is defined naturally on new data points.

Related Work: We pointed out the extensive literature on semi-supervised learning and the growing number of studies on learning structured and inter-dependent variables. Smola et. al. [12] investigate structured learning with missing variables and propose marginalizing over the missing nodes of a graph which is prohibitive for a semi-supervised learning scenario. The most relevant previous work is the transductive structured learning proposed by Lafferty et. al. [9]. We give a detailed comparison of this work and ours in Section 5. Here, we emphasize that the main distinction is the limitation of [9] to in-sample observations, whereas our approach extends naturally to new test points.

2 Supervised Learning for Structured Variables

In structured learning, the goal is to learn a mapping $h : \mathcal{X} \to \mathcal{Y}$ from *structured* inputs to *structured* response values. For example, in parsing, h maps a sentence x to a parse tree y. There exists a feasible set $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$. We write $\mathcal{Y}(x)$ for the set of $y \in \mathcal{Y}$ such that $\langle x, y \rangle \in \mathcal{Z}$. For simplicity, let us assume that $\mathcal{Y}(x)$ is finite for all $x \in \mathcal{X}$, which is the case in many real world problems and in all our examples.

The common approach to learn h is to construct a discriminant function $F : \mathcal{Z} \to \Re$ which maps the feasible input-output pairs to real values (measuring the compatibility of the pair) and to maximize F over $\mathcal{Y}(x)$ to make a prediction for x,

$$h(x) = \operatorname*{argmax}_{y \in \mathcal{Y}(x)} F(x, y). \tag{1}$$

In order to construct F, we use a part formulation similar to the formulation in [10]. We let $\mathcal{P}$ be a set of "parts" of all possible structured input-output pairs and define a function P such that for $x \in \mathcal{X}$ and $y \in \mathcal{Y}(x)$, $P(\langle x, y \rangle)$ is the set of parts of $\langle x, y \rangle$ and is a finite subset of $\mathcal{P}$. To avoid confusion with probabilities, with a slight abuse of notation, we write $\mathcal{P}(\langle x, y \rangle)$ instead of $P(\langle x, y \rangle)$. For a Mercer kernel $k : \mathcal{P} \times \mathcal{P} \to \Re$ on $\mathcal{P}$, there is an associated RHKS $\mathcal{H}_k$ of functions $f : \mathcal{P} \to \Re$, where f measures the *goodness* of a part p. For any $f \in \mathcal{H}_k$, we define a function

F_f on $\mathcal{Z}$ as

$$F_f(x,y) = \sum_{p \in \mathcal{P}(\langle x,y \rangle)} f(p), \tag{2}$$

where the compatibility of $\langle x,y \rangle$ is given by the total goodness of its parts.

In the supervised learning scenario, we are given a sample S of ℓ pairs $\langle x^1, y^1 \rangle$, ..., $\langle x^\ell, y^\ell \rangle$ drawn i. i. d. from an unknown but fixed probability distribution P on $\mathcal{Z}$. The goal is to learn a function f with small expected loss $E_P[\mathcal{L}(x,y,f)]$ where $\mathcal{L}$ is a prescribed loss function. This is commonly realized by learning f that minimizes the regularized loss functional

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_k} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda \|f\|_k^2, \tag{3}$$

where $\|.\|_k$ is the norm corresponding to $\mathcal{H}_k$ measuring the complexity of f.

A variety of loss functions $\mathcal{L}$ have been considered in the literature. In kernel conditional random fields (CRFs) [9], the loss function is given by

$$\mathcal{L}(x,y,f) \quad = \quad -F_f(x,y) + \log \sum_{\hat{y} \in \mathcal{Y}(x)} \exp(F_f(x,\hat{y}))$$

In structured Support Vector Machines (SVM), the loss function is given by

$$\mathcal{L}(x,y,f) \quad = \quad \max_{\hat{y} \in \mathcal{Y}(x)} \Delta(x,y,\hat{y}) + F_f(x,\hat{y}) - F_f(x,y), \tag{4}$$

where $\Delta(x,y,\hat{y})$ is some measure of distance between y and $\hat{y}$ for a given observation x. A natural choice for Δ is to take $\Delta(x,y,\hat{y})$ to be the indicator $1_{[y \neq \hat{y}]}$ [2]. Another choice is to take $\Delta(x,y,\hat{y})$ to be the size of the symmetric difference between the sets $\mathcal{P}(\langle x,y \rangle)$ and $\mathcal{P}(\langle x,\hat{y} \rangle)$ [13].

Let $\mathcal{P}(x)$ be the union of all sets of the form $\mathcal{P}(\langle x,y \rangle)$ for $y \in \mathcal{Y}(x)$ and $\mathcal{P}(S)$ be the union of all sets $\mathcal{P}(x^i)$ for x^i in the sample. Then, we have following straightforward variant of the Representer Theorem [8], which was also presented in [9].

Definition: A loss $\mathcal{L}$ is *local* if $\mathcal{L}(x,y,f)$ is determined by the value of f on the set $\mathcal{P}(x)$, i.e., for $f, g : \mathcal{P} \to \Re$ we have that if $f(p) = g(p)$ for all $p \in \mathcal{P}(x)$ then $\mathcal{L}(x,y,f) = \mathcal{L}(x,y,g)$.

Theorem 1. *For any local loss function $\mathcal{L}$ and sample S there exist weights α_p for $p \in \mathcal{P}(S)$ such that f^* as defined by (3) can be written as follows.*

$$f^*(p) = \sum_{p' \in \mathcal{P}(S)} \alpha_{p'} k(p', p) \tag{5}$$

Chains: We now give a simple standard example. In a simple chain, $\mathcal{X}$ is the set of all finite sequence of observations and $\mathcal{Y}$ is the set of all finite strings over a hidden state alphabet Σ. The feasible set $\mathcal{Z}$ is the set of all $\langle x,y \rangle$ pairs where x and y have the same length. $|\mathcal{Y}(x)|$ grows exponentially in $|x|$, where $|x|$ is the length of x. For $1 \leq t \leq |x|$, let x_t be the observation token at position t in x. We define $|y|$ and y_t similarly. The set of parts of $\langle x,y \rangle$, $\mathcal{P}(\langle x,y \rangle)$, consists of two sets: the set of parts capturing dependencies between the label y_t and the observation x_t and the set of parts capturing interdependencies of consecutive labels y_t, y_{t+1} for all positions t in the sequence.

$$\mathcal{P}(\langle x,y \rangle) = \{\langle t, y_t, x_t \rangle : \ 1 \leq t \leq |x|\} \cup \{\langle t, y_t, y_{t+1} \rangle : \ 1 \leq t \leq |y| - 1\}.$$

Note $\mathcal{P}(\langle x, y\rangle)$ grows polynomially in $|x|$. $\mathcal{P}$ is the set of all triples $\langle t, s, \bar{s}\rangle$ and triples $\langle t, s, u\rangle$ where $t \geq 1$, $s, \bar{s} \in \Sigma$ and u ranges over observations. In general $u \in \Re^d$. $k(p, p')$ is commonly defined as

$$k(\langle t, s, \bar{s}\rangle, \langle t', s', \bar{s}'\rangle) = \delta(s, s')\delta(\bar{s}, \bar{s}'), \quad (6)$$

$$k(\langle t, s, u\rangle, \langle t', s', u'\rangle) = \delta(s, s')\tilde{k}(u, u'), \quad (7)$$

where $\delta(w, w')$ denotes the Kronecker-δ. $k(p, p')$ is 0 if p and p' are of different types, i.e., if neither of the above equations apply. The parts in this example correspond to the cliques of the dependence graph. However, the part notion is more general than the clique notion. For example, it also includes parsing.

3 A Semi-Supervised Learning Approach to Structured Variables

In semi-supervised learning, we are given a sample S consisting of l input-output pairs $\{(x^1, y^1), \ldots, (x^\ell, y^\ell)\}$ drawn i. i. d. from the probability distribution P on $\mathcal{Z}$ and u unlabeled input patterns $\{x^{\ell+1}, \ldots, x^{\ell+u}\}$ drawn i. i. d from the marginal distribution $P_\mathcal{X}$, where usually $l < u$. Let $\mathcal{X}(S)$ be the set $\{x^1, \ldots, x^{\ell+u}\}$ and let $\mathcal{Z}(S)$ be the set of all pairs $\langle x, y\rangle$ with $x \in \mathcal{X}(S)$ and $y \in \mathcal{Y}(x)$.

If the smoothness assumption is valid, one can use the unlabeled data points to determine functions that respect that assumption. Belkin et. al. [3] present a formalism to learn a linear discriminant function that varies smoothly over the intrinsic structure revealed by the marginal distribution of the input patterns, by minimizing an objective function consisting of three terms: a loss function over labeled training examples, a penalty term that controls the complexity of the function f and a penalty term that controls the smoothness of f over the intrinsic marginal structure,

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_k} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda_1 ||f||_k^2 + \lambda_2 ||f||_{k_S}^2 \quad (8)$$

where k_S is a kernel representing the intrinsic measure of the marginal distribution. Since, in general, the marginal distribution of the input patterns is not know, the smoothness penalty term is estimated using the labeled and unlabeled input patterns.

Here, we generalize this framework to structured variables. The smoothness assumption in the structured setting states that f should be smooth on the underling density on the parts $\mathcal{P}$. We approximate this constraint by enforcing f to assign similar *goodness* scores to two parts p and p', if p and p' are *similar*, for all parts of $\mathcal{Z}(S)$. Let $\mathcal{P}(S)$ be the union of all sets $\mathcal{P}(z)$ for $z \in \mathcal{Z}(S)$ and let W be symmetric matrix where $W_{p,p'}$ represents the similarity of p and p' for $p, p' \in \mathcal{P}(S)$.

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_k} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda_1 ||f||_k^2 + \lambda_2 \sum_{p,p' \in \mathcal{P}(S)} W_{p,p'}(f(p) - f(p'))^2 \quad (9)$$

Note that the last term depends only on the value of f on the parts in the set $\mathcal{P}(S)$. For any local loss $\mathcal{L}(x, y, f)$, we immediately have the following Representer Theorem for the semi-supervised structured case where now S includes the labeled and the unlabeled data.

$$f^* = \sum_{p' \in \mathcal{P}(S)} \alpha_{p'} k(p', p) \quad (10)$$

For simplicity in notation, we rewrite the optimization problem in (9) using vectors and matrices indexed by $\mathcal{P}(S)$. We let α range over vectors with a component α_p for each $p \in \mathcal{P}(S)$. Let G be the Gram matrix of the base kernel on the set $\mathcal{P}(S)$ defined by $G_{p,p'} = k(p,p')$ and f_α be the function on $\mathcal{P}$ represented by α, i.e., $f_\alpha(p) = \sum_{p' \in \mathcal{P}(S)} \alpha_{p'} k(p',p)$. The Laplacian matrix L of a similarity graph represented by a similarity matrix W is $D - W$ where D is a diagonal matrix defined by $D_{p,p} = \sum_{p'} W_{p,p'}$. Plugging (10) in (9), we get the following optimization problem.

$$\alpha^* = \operatorname*{argmin}_{\alpha} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f_\alpha) + \lambda_1 \alpha^T G \alpha + \lambda_2 \alpha^T GLG\alpha \tag{11}$$

f_α, as a vector in the linear space H_k, is a linear function of the vector α. If $\mathcal{L}(x,y,f)$ is convex in f, as in the case for logistic or hinge loss, then (11) is convex in α.

In semi-supervised learning literature, the similarity matrix is defined by building a k-nearest neighbor graph on the set $\mathcal{X}(S)$ to "smooth" the classification function. Similarly, we build the adjacency matrix W over $\mathcal{P}(S)$ from a K-nearest neighborhood graph N, such that for $p, p' \in \mathcal{P}(S)$, $W_{p,p'}$ is the inverse weight of the edge between p and p' in N if it exists and 0 otherwise. The metric of the K-nearest neighbor algorithm is defined with respect to the nature of the parts. For instance, we assign $d(p,p') = \infty$ if p or p' does not capture any dependence on an observation, since such a part does not convey any information on the underlying geometric structure of the observations.

4 Maximum margin semi-supervised structured learning

We now investigate optimizing the hinge loss as defined by (4). Defining $\beta^{x,y}$ to be the vector where $\beta_p^{x,y} = 1$ if $p \in \mathcal{P}(\langle x,y \rangle)$ and 0 otherwise, the linear discriminant on $\mathcal{Z}$ is given by

$$F_{f_\alpha}(x,y) = \alpha^T G \beta^{x,y}.$$

Replacing the max in (4) by inequalities on slack variables, we can rewrite (10) for margin maximization as

$$\alpha^* = \operatorname*{argmin}_{\alpha} \min_{\xi} \sum_{i=1}^{l} \xi_i + \alpha^T Q \alpha \tag{12}$$

$$\xi_i \geq \triangle(\hat{y}, y^i) - \alpha^T G \left(\beta^{x^i, y^i} - \beta^{x^i, \hat{y}} \right) \quad \forall \hat{y} \in \mathcal{Y}(x^i),\ \forall i \leq l \tag{13}$$

where $Q = \lambda_1 G + \lambda_2 GLG$. This gives a convex quadratic program over the vectors indexed by $\mathcal{P}(S)$. The main difficulty is that the number of constraints in (13) is typically very large. In the chain example, $|\mathcal{Y}(x)|$ is exponential in $|x|$. Fortunately we can solve the optimization problem without enumerating the entire set of constraints.

Introducing Lagrange parameters $\theta_{(x^i,y)}$ to enforce the margin constraints for all labeled observations x and $y \in \mathcal{Y}(x)$ and using the standard Lagrangian duality techniques, we get the following dual Quadratic program:

$$\theta^* = \operatorname*{argmin}_{\theta} \theta^T dR\, \theta - \Delta^T \theta \tag{14}$$

$$\theta_{(x^i,y)} \geq 0,\ \sum_{y \in \mathcal{Y}(x)} \theta_{(x^i,y)} = 1, \quad \forall y \in \mathcal{Y}(x^i), \quad \forall i \leq l, \tag{15}$$

where Δ is a vector of $\triangle(y, \hat{y})$ for all $y \in \mathcal{Y}(x)$ of all labeled observations x, $d\beta$ is a matrix whose (x^i, y)th column $d\beta_{.,(x^i,y)} = \beta^{x^i,y^i} - \beta^{x^i,y}$ and

$$dR = d\beta^T \, G^T Q^{-1} G d\beta. \tag{16}$$

Now, the linear discriminant is given by

$$F_{f_\theta}(x, y) = \theta^T \, d\beta^T \, G^T Q^{-1} G \, \beta^{x,y}$$

Although this is an exponential sized QP, efficient optimization methods have been proposed for similar problems in the supervised structured learning setting [2, 14]. We adopt the algorithm proposed in [14] to solve Eq. (14) by simply replacing the kernel on structured objects with dR. This algorithm has been proven to converge in polynomial time in the size of the output of each observation. Note that the number of parameters in θ is determined by the number of labeled observations which is generally small in semi-supervised structured learning.

Let us examine the quadratic term in more detail. It is easy to see that the Eq. (14) can be rewritten in terms of $R = \beta^T \, G^T Q^{-1} G \beta$. Now, we consider a special case where if p and p' are of different part types, then their kernel value $k(p, p') = 0$ and their distance in the nearest neighborhood graph $d(p, p') = \infty$. Then, the Gram matrix G and the Laplacian L can be presented as block diagonal matrices, where each block refers to a type of part. and the kernel R decomposes into contributions from each part type. Such decomposition significantly reduces the complexity of the computation of R, in particular the computation of Q^{-1}.

Our chain example is an instance of the special case considered above. The kernel R for this structure decomposes into a contribution from inter-label dependencies and a contribution from label-observation interactions.

$$R((x, y), (x', y')) = R^1((x, y), (x', y')) + R^2((x, y), (x', y')) \tag{17}$$

$$R^1((x, y), (x', y')) = \sum_{t,t'} \delta(y_t, y'_{t'}) \delta(y_{t+1}, y'_{t'+1}) \tag{18}$$

$$R^2((x, y), (x', y')) = \sum_{t,t'} \delta(y_t, y'_{t'}) \tilde{k}_{x_t}^T (\lambda_1 \tilde{K} + \lambda_2 \tilde{K} L_x \tilde{K})^{-1} \tilde{k}_{x'_{t'}} \tag{19}$$

where $\tilde{K}$ is the gram matrix of each observation in both labeled and unlabeled observation sequences using the kernel $\tilde{k}$ in Eq. (7), $\tilde{k}_u$ is the vector $\tilde{k}(u, \bar{u})$ for all observations $\bar{u}$ in labeled and unlabeled observation sequences and L_x is the Laplacian of the similarity graph over the observation tokens. Thus, in simple chains, unlabeled data simply augments the kernel $\tilde{k}$ over observation tokens in order to take the intrinsic structure of the marginal distribution of the observations into account.

5 Semi-Supervised vs Transductive Learning

Since one major contribution of this paper is learning a classifier for structured objects that is defined over the complete part space $\mathcal{P}$, we now examine the differences of semi-supervised and transductive learning in more detail. The most common approach in semi-supervised learning to realize the smoothness assumption is to construct a data dependent kernel k_S derived from the graph Laplacian on a nearest neighbor graph on the labeled and unlabeled input patterns in the sample S. Thus, k_S is not defined on observations that are out of the sample. Given k_S, one can construct a function $\tilde{f}^*$ on S defined as

$$\tilde{f}^* = \operatorname*{argmin}_{f \in \mathcal{H}_{k_S}} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda ||f||^2_{k_S}. \tag{20}$$

This formalism has been extended to the structured learning setting in [9], where k_S is defined over $\mathcal{P}(S)$.

It is well known that kernels can be combined linearly to yield new kernels. This observation in the transductive setting leads to the following optimization problem, when the kernel of the optimization problem is taken to be a linear combination of a graph kernel k_S and a standard kernel k restricted to $\mathcal{P}(S)$.

$$\begin{aligned} f^*_{\text{in}} &= \operatorname*{argmin}_{f \in \mathcal{H}_{(\mu_1 k + \mu_2 k_S)}} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda ||f||^2_{(\mu_1 k + \mu_2 k_S)} \\ &= \operatorname*{argmin}_{f \in \mathcal{H}_{(\mu_1 k + \mu_2 k_S)}} \sum_{i=1}^{\ell} \mathcal{L}(x^i, y^i, f) + \lambda \mu_1 ||f||^2_k + \lambda \mu_2 ||f||^2_{k_S} \end{aligned} \tag{21}$$

The possibility of the combination of the kernel derived from the graph Laplacian and the standard kernel over parts of a structure was pointed out by [9]. An important distinction between f^*_{in} and f^* in (8), the optimization performed in this paper, is that f^*_{in} is only defined on $\mathcal{P}(S)$ (only on observations in the training data) while f^* is defined on all of $\mathcal{P}$ and can be used for novel (out of sample) inputs x. We note that in general $\mathcal{P}$ is infinite. Out-of-sample extension is already a serious limitation for transductive learning, but it is even more severe in the structured case where parts of $\mathcal{P}$ can be composed of multiple observation tokens.

Lemma 1. *If $\mathcal{H}_k$ contains all functions on the finite set $\mathcal{P}(S)$ then f^*_{in} as defined in* (21) *and f^* as defined in* (8) *agree on all points in $\mathcal{P}(S)$.*

This lemma applies whenever k is a universal kernel. Note that the universality of k is not necessary for the representer theorem (10).

6 Experiments

We evaluate the validity of our approach to semi-supervised structured learning by performing experiments using a simple chain model for pitch accent prediction. The goal in pitch accent prediction, a sub-task of speech recognition, is to detect more prominent words in an utterance We model this problem as a sequence annotation problem, where $\mathcal{Y}(x) = \{0, 1\}^T$ with $T = |x|$. and where $x_t \in \Re^{31}, \forall t$.

We ran experiments comparing the performance of the semi-supervised SVM approach in [3] which ignores the structural dependency of problem and the performance of structured semi-supervised SVM. For simplicity, we refer to structured semi-supervised SVM as STR SVM and to unstructured semi-supervised SVM simply as SVM. For both SVM and STR SVM, we used RBF kernel as the base kernel $\tilde{k}$ in Eq. (7) and a 5-nearest neighbor graph to construct the Laplacian. We report the average results of experiments with 5 random selection of labeled sequences in Table 1. Here, we varied the number of labeled sequences (L:X) and report the average number of labeled observation tokens (Lo:X), and similarly for unlabeled observations (U:X , Uo:X). We report per-label accuracy of test data as well as in-sample unlabeled data.

The results show the advantage of a sequence model over a non-structured model. These two models perform the same only when 4 labeled sequences are available, where both of the models simply select the most common label. We also observe the usefulness of unlabeled data both in the structured and unstructured models. The small difference between the accuracy of in-sample unlabeled data and the test data indicates the natural extension of our framework to new data points. One

	L:4 U:0	L:4 U:80	L:40 U:0	L:40 U:80	L:40 U:200
	Lo:34.2 Uo:0	Lo:33.0 Uo:641	Lo:355 Uo:0	Lo:298 Uo:641	Lo:301 Uo:1605
SVM					
test	65.92	68.83	70.34	71.27	73.68
in-sample	-	69.94	-	72	73.11
STR SVM					
test	65.81	70.28	72.15	74.92	76.37
in-sample	-	70.72	-	75.66	77.45

Table 1: Per-label accuracy of Pitch Accent Prediction.

surprising result is the large improvement of STR SVM from L:40 U:0 to L:40 U:80, even though the ratio of labeled and unlabeled data points is low. We believe this is in fact to due the low accuracy on L:40 and U:0 setting, which is also reflected in the difference between SVM and STR SVM for the same setting.

7 Conclusions

We presented a discriminative approach to semi-supervised learning of structured and inter-dependent response variables. In this framework, we derived a maximum margin formulation and presented experiments for a simple chain model. Our approach naturally extends to the classification of unobserved structured inputs and this is supported by our empirical results which showed similar accuracy on in-sample unlabeled data and out-of-sample test data. We are currently working on models with more complicated parts than the simple chain.

References

[1] Y. Altun, T. Hofmann, and A. Smola. Gaussian process classification for segmenting and annotating sequences. In *ICML*, 2004.

[2] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *ICML*, 2003.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. Technical Report 06, University of Chicago CS, 2004.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] O. Chappelle, J. Weston, and B. Scholkopf. Cluster kernels for semi-supervised learning. In *(NIPS)*, 2002.

[6] M. Collins and N.l Duffy. Convolution kernels for natural language. In *(NIPS)*, 2001.

[7] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *(ICML)*, pages 200–209, 1999.

[8] G. Kimeldorf and G. Wahba. Some results on tchebychean spline functions. *Journal of Mathematics Analysis and Applications*, 33:82–95, 1971.

[9] John Lafferty, Yan Liu, and Xiaojin Zhu. Kernel conditional random fields: Representation, clique selection, and semi-supervised learning. In *(ICML)*, 2004.

[10] D. McAllester, F. Pereira, and M. Collins. Case-factor diagrams for structured probabilistic modeling. In *(UAI)*, 2004.

[11] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proceedings of AAAI-98*, pages 792–799, Madison, US, 1998.

[12] Alex Smola, S. V. N. Vishwanathan, and Thomas Hofmann. Kernel methods for missing variables. In *Proceedings of AISTAT*, 2005.

[13] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2004.

[14] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *(ICML)*, 2004.