Optimally Extracting Discriminative Disjunctive Features for Dimensionality Reduction

Amrita Saha
IBM Research
Bangalore, India
amrsaha4@in.ibm.com

Naveen Nair
Indian Institute of Technology,
Bombay, India
naveennair@cse.iitb.ac.in

Ganesh Ramakrishnan
Indian Institute of Technology,
Bombay, India
ganramkr@cse.iitb.ac.in

ABSTRACT

Dimension Reduction is one popular approach to tackle large and redundant feature spaces as seen in most practical problems, either by selecting a subset of features or by projecting the features onto a smaller space. Most of these approaches suffer from the drawback that the dimensionality reduction objective and the objective for classifier training are decoupled. Recently, there have been some efforts to address the two tasks in a combined manner by attempting to solve an upper-bound to a single objective function. But the main drawback of these methods is that they are all parametric, in the sense that the number of reduced dimensions needs to be provided as an input to the system. Here we propose an integrated non-parametric learning approach to supervised dimension reduction by exploring a search space of all possible disjunctions of features and discovering a sparse subset of (interpretable) disjunctions that minimise a regularised loss function. Here, in order to discover good disjunctive features, we employ algorithms from hierarchical kernel learning to simultaneously achieve efficient feature selection and optimal classifier training in a maximum margin framework and demonstrate the effectiveness of our approach on benchmark datasets.

1. INTRODUCTION

In building machine learning models using features, it can so happen that several features might be either irrelevant or contain redundant information, which could befuddle the model learner or lead to over-fitting and consequently, a less effective model. Therefore, a small set of relevant and non-redundant features that effectively discern classes is desired. Identifying the best feature subspace for classification comes under the broad area of dimensionality reduction (DR) techniques, which can be divided into methods for feature subset selection and methods for feature extraction. Here we pose dimensionality reduction as discovering a small set of interpretable disjunctions of basic features as reduced dimension in a supervised setting and propose an integrated non-parametric learning approach that minimize a regularized loss function to solve the objective

Feature subset selection (FSS) is the process of selecting a subset of features that embodies relevant and non-redundant information for use in model construction. Some approaches like Relief, FOCUS and wrapper methods [14] (often greedily) select the subset of features based on some local relevance criteria such as information gain, chi-squared test, *etc.*, or some global objective such as the $\mathcal{L}_1$ norm regularization in an SVM classifier.

On the other hand, feature extraction approaches attempt to discover a lower-dimensional embedding of the feature space that will approximately retain the statistical relation between the instances and the class label as in the original space. Any approach for dimensionality reduction can be adjudged parametric or non-parametric respectively depending on whether the number of reduced dimensions of the embedding is considered as an input parameter or whether it is estimated within the approach. Further, each line of work can be classified as supervised, weakly-supervised or unsupervised. Unsupervised parametric methods include projective methods like Principal Component Analysis (PCA), Kernel PCA and its variants, manifold methods like Multi-Dimensional Scaling, Laplacian EigenMaps, discriminant analysis techniques such as Linear Discriminant Analysis, Kernel Discriminant Analysis, Hybrid Discriminant Analysis [29] (a combination of Principal Component Analysis and Linear Discriminant Analysis which is claimed to lead to more robust models), and Continuous Latent Variable methods like Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation *etc.*

There has been considerable amount of work on adapting dimension reduction methods in supervised or weakly-supervised settings, for example, supervised latent variable models like supervised latent dirichlet allocation (sLDA) [7], hierarchical supervised LDA (HsLDA) [22] which extends sLDA to the case of hierarchical supervision, labeled-LDA [24] which focuses on multi-labeled supervision for multi-labeled collections.

There have also been a few attempts at building discriminative frameworks for supervised dimensionality reduction, for example discLDA [15] motivated by the observation that the parameter estimates obtained in the generative counterparts by maximum likelihood or as Bayesian posterior do not necessarily lead to optimum models for predictive tasks. Some of these methods make assumptions that may not be appropriate in reality. For example sLDA assumes a normal distribution for the response variable and further as-

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

sumes it to be linearly dependent on its empirical mixture proportions. On the other hand, discLDA assumes that the mixture proportions of each class after a linear transformation should be close to each other. This assumption seems very restrictive and also appears to go directly against classification requisites. The approach generally adopted for dimensionality reduction in nonparametric settings is to employ stochastic processes instead of distributions, which are flexible in the sense of accommodating infinite number of variables and hence able to estimate the number of reduced dimensions implicitly. Two such stochastic variants of the unsupervised approaches are Hierarchical Dirichlet Processes (HDP) [26] and hierarchical LDA [6]. Other than this, there have been some approaches on determining the size of learned ontologies, in the area of topic-modeling, by studying the change in average cosine distances between topics with respect to the increase in number of topics. Arun *et. al* [3] additionally consider information from the topic-document matrix and propose a measure for the estimation of the 'correct' number of topics based on Kullback-Leibler divergence of the singular value distributions of these matrices.

The above methods that treat dimensionality reduction as an isolated problem, allow for the use of any classification or regression model building on the discovered subspace. However, since the dimensionality reduction and classifier training are decoupled from each other, these approaches cannot generally guarantee optimality of feature selection with respect to the classifier objective. There have been some attempts [18] to develop models that integrate dimensionality reduction with model building, and which have shown the ablility to discover predictive topic representations that are more suitable for supervised prediction tasks. Maximum Entropy Discrimination Latent Dirichlet Allocation (medLDA) [30] is a maximum margin variant of maximum-entropy discrimination LDA which integrates maximum margin criterion with LDA by optimizing a single objective function with a set of expected margin constraints. A more recent approach, Multi-Modal Probabilistic Latent Semantic Analysis (MMpLSA) [28], that evolved along the same lines, has integrated PLSA (in place of its Bayesian version of LDA) with maximum margin criterion and has shown to perform better than its predecessors [7, 15, 30]. Further, these studies have shown that building another classifier using the induced dimensions does not introduce much performance gain. The main limitation of these methods is that they are parametric – they need the number of reduced dimensions as input and have no intrinsic mechanism for estimating this number within the system.

Parallely there has been work aimed at feature selection embedded within a max-margin classifier, popularly SVM. [27] uses zero normed regularizer in the SVM objective, where approximate minimization of the zero-norm leads to sparse feature selection. But their feature selection is parametric, since the desired number of features to be selected explicitly given as input, and its value is decided by cross-validation. Empirically they show the model performance does not degrade as much as the baseline SVM as desired number of features given as input gets smaller, but there is no way to determine the optimum number of features without actually enumerating. On the other hand, [21] proposes a convex energy-based framework to jointly perform feature selection and SVM parameter learning for linear and non-linear kernels. Theoretically they have shown a tight connection of their formulation with that of L_1 norm regularized SVM and empirically their classification accuracies are found to be comparable to vanilla L_2 norm SVM and show some reduction in number of features selected over this non-sparse variant of SVM.

There has not been much discussion on the optimality of the models built subsequently from the reduced set of features produced by dimensionality reduction techniques. Gal Chechik [9] solves the maximum margin objective in the dual but does not guarantee optimality, while medLDA [30] and MMpLSA [28] have both solved a tight bound approximation of the original objective in the interest of tractability.

1.1 Our Contribution

Our work falls in the league of integrated maximum margin linear dimensionality reduction approaches for model building in a supervised classification setting[1] setting. For simplicity, we assume that all our basic features (attributes themselves) are boolean. For example, presence or absence of a word in the dictionary can be a basic boolean feature. Neverthless, our approach can be applied to settings with nominal features by booleanizing nominal or numerical features. We intend to learn fewer features that capture the redundant information present in basic features by grouping them and representing each group as a single disjunctive feature. In the case of text classification, the disjunction of (nearly) synonymous or strongly correlated words can be treated as a single feature. For example, words such as *beautiful* and *gorgeous* could convey similar sense about an entity and therefore a single feature that is a disjunction of these could be sufficient. The preferred disjunctive features should be maximum in the sense that we try to include as many synonymous basic features as possible and exclude any non-synonymous basic feature in a disjunction. For example, *ugly* is not a synonym of *beautiful* and therefore we would generally not expect *ugly* and *beautiful* to co-occur in the same disjunctive feature. Our objective is to construct an optimal set of relevant and non-redundant features for classification, by minimizing the hinge loss. As noted above, any approach with the dimensionality reduction approach decoupled from the classifier training has limitations in finding optimum models. In this paper, we propose an integrated supervised approach for dimensionality reduction in a maximum margin framework. Since the dimensions are disjunctions of basic features, the induced features can be expected to be interpretable.

To the best of our knowledge, there has not been any approach in discriminative learning that integrates non-parametric dimensionality reduction with optimal model building for classification. There has been some work on employing maximum margin based nonparametric dimensionality reduction in a multi-task setting [2] and in the area of learning underlying shared structures amongst classes [1] in a multi-class setting. Although these methods solve their objectives optimally, they are not directly comparable to our work, since in the case of binary classification or 1-task case, their dimensionality reduction approach reduces to a trivial feature selection using 1-norm regularization. Our main contribution in this paper is an integrated optimal and efficient clas-

[1]While our approach can be very naturally extended to the regression setting by changing the loss function, we have not empirically studied supervised dimensionality reduction for regression in this work.

sifier learning and dimensionality reduction technique based on the Hierarchical Kernel Learning (HKL) setting [4]. We conclude this section by briefly introducing the hierarchical kernel learning framework.

Hierarchical kernel learning (HKL) [4] approaches have gained interest recently due to their ability to learn kernels in a large kernel space [4]. Bach, 2009 has introduced HKL framework that efficiently explores an exponential kernel space where individual kernels can be decomposed into base kernels [4]. This approach selects kernels from the space of all possible kernels embedded in a directed acyclic graph using a graph based sparsity inducing norm. The complexity of HKL is polynomial in the number of selected kernels. The employed regularizer discourages complex kernels, thereby encouraging a small set of simple kernels to be selected. In this paper, we leverage the HKL framework to simultaneously perform dimensionality reduction and classifier training. We prune away irrelevant features and group redundant features in the form of logical disjunctions. The sparsity inducing hierarchical regularizer used in HKL selects a sparse set of disjunctions. We evaluate our approach on standard datasets and compare with other approaches. From our experiments, we observe that the disjunctions discovered by our method are more fine-grained than the topics identified by competing methods, while also yielding significant improvements in test accuracies.

The rest of our paper is organized as follows. In Section 2, we discuss the proposed approach and algorithm for learning disjunctions for dimensionality reduction in a hierarchical kernel learning setting. Section 3 is dedicated to experiments and empirical observations. We conclude our paper in Section 4.

2. OPTIMAL NON-PARAMETRIC MAX MARGIN DIMENSIONALITY REDUCTION

We now formally define our problem of simultaneously performing dimensionality reduction and classifier training and present an efficient polynomial time algorithm to solve the objective optimally. We consider features that do not discern classes as irrelevant and therefore can be discarded. For example, in sentiment classification, a set of similar-meaning words such as *method, algorithm, etc.* might not help in discerning classes and can be omitted. On the other hand, multiple features capturing the same information (synonyms) are redundant and might result in an ineffective classifier. For example, words such as *beautiful, exquisite, gorgeous, charming, etc.* capture similar information about the entity being discussed and should probably be clubbed together in the same dimension. For effectively capturing the meaning of a group of synonymous basic features without redundancy, we explore the space of disjunctive features that are disjunctions ($\vee$) of basic features. For instance, in document classification, synonymous words *beautiful*, *exquisite*, and *gorgeous* can be used to construct a disjunctive feature and the feature is instantiated when any one or more of the component features are active. We refer to such features as DisjunctProjs (**Disjunct**ive **Proj**ections).The space of all possible DisjunctProjs can be visualized as a lattice, with a structure similar to the subset lattice, where the top node is the empty node, the nodes at the next level are the individual basic features and so on. The bottom node in the lattice is the disjunction of all the basic features. Upward and downward refinements of a node can be defined in terms of deletion or addition of a basic feature from or to the node respectively.

We aim at automatically selecting good maximal DisjunctProjs from the ordering. A *good* DisjunctProj is a disjunction which does not contain any statistically different feature. A maximal DisjunctProj is a disjunction of maximum number of basic features capturing (statistically) similar information about the classes being discriminated against each other. Therefore, a good and maximal DisjunctProj corresponds to a disjunction of synonymous basic features in which no more basic features can be added without affecting its meaning. With this understanding, if a DisjunctProj is not effective for classification, we assume that the feature will not become more effective by the addition of a new basic feature to the disjunction. For example, if *beautiful* $\vee$ *ugly* is not good, then *beautiful* $\vee$ *ugly* $\vee$ *gorgeous* may not be good in general. Therefore, in the ordering, if a node is not selected, we expect that none of its descendants be selected either. Now let us assume that we have a good DisjunctProj in the form of *beautiful* $\vee$ *exquisite* $\vee$ *gorgeous*. This is maximal if adding a new word to the disjunction results in a bad DisjunctProj for classification. Therefore, if *ugly* is added, in the new DisjunctProj formed by this addition, *ugly* can be considered as noise. Next, we formally define our problem.

2.1 Formal Specification of the Problem

We pose our requirement as a maximum margin optimization problem which is expected to select a sparse set of good DisjunctProjs from the ordering and learn their optimal feature weights simultaneously. Let each elements of the vector $\boldsymbol{\psi}$ correspond to a node in the disjunction lattice and $\mathbf{w}$ the corresponding weights. $\mathcal{V}$ is the set of indices to nodes in the lattice. A node $\psi_v(.)$ in the ordering is a disjunction of a set of basic features and can be represented as $\vee_{\hat{v}\in v}\ \psi_{\hat{v}}(.)$, where $\hat{v}$ stands for a basic feature present in v.

To select a sparse set of DisjunctProjs from the exponential feature space, we employ a hierarchical regularizer on the exponential feature space and present the SVM formulation for binary classification as,

$$\min_{\mathbf{w},b,\boldsymbol{\xi}}\ \frac{1}{2}\Big(\sum_{v\in\mathcal{V}}\delta_v\parallel \mathbf{w}_{D(v)}\parallel_p\Big)^2 + C\ \mathbf{1}^\top\boldsymbol{\xi} \qquad (1)$$

$$s.t.\ \forall i:\ y_i\Big(\sum_{v\in\mathcal{V}}\langle w_v,\psi_v(\mathbf{x}_i)\rangle - b\Big)\geq 1-\xi_i,\ \boldsymbol{\xi}\geq 0$$

where $\mathbf{w}_{D(v)}$ is the vector of feature weights corresponding to the elements in the descendant nodes $D(v)$ of node v including the node v itself, $p\in(1,2]$, δ_v is the prior parameter that can be interpreted as the usefulness of node v, C is the regularization parameter, ξ_i is the slackness in the margin for i^{th} example, $\mathbf{x}_i$ is the input vector of dimension n (where n is the number of basic features) corresponding to the i^{th} example, $y_i\in\{0,1\}$ is the predicted output value for the i^{th} example, b is the bias term, w_v is the feature weight corresponding to v^{th} node and $\psi_v(\mathbf{x}_i)$ is the truth value of the v^{th} node for the i^{th} training example. To discourage very large and potentially ineffective DisjunctProjs, we define δ_v as $\beta^{|v|-k}$, where β and k are some parameterized constants and $|v|$ is the size of the node v. Many of $\parallel \mathbf{w}_{D(v)}\parallel_p$ are expected to be zero due to the 1-norm which will force w_u, $\forall u\in D(v)$, to reduce to zeros. This effectively

discourages selection of large number of DisjunctProjs. Additionally, as illustrated by Szafranski *et al.* [25], p-norm, where $p \in (1, 2]$, induces further sparsity among the nodes. Therefore, for $\| \mathbf{w}_{D(v)} \|_p$ that are not reduced to zero by 1-norm, the p-norm forces many of the descendants of node v to zero and thus encourages a sparse solution.

The kernel for a node v can be defined as, $\mathbf{K}_v(\mathbf{x}_i, \mathbf{x}_j) = \langle \psi_v(\mathbf{x}_i), \psi_v(\mathbf{x}_j) \rangle = (1 - \prod_{\hat{v} \in v} \overline{\psi_{\hat{v}}}(\mathbf{x}_i))(1 - \prod_{\hat{v} \in v} \overline{\psi_{\hat{v}}}(\mathbf{x}_j))$. This enables the sum of the kernels over a sub-lattice $\mathcal{V}$ to be computed efficiently. For instance, the sum of kernels over the entire lattice is,

$$\sum_{v \in \mathcal{V}} \mathbf{K}_v(\mathbf{x}_i, \mathbf{x}_j) = 2^n + \prod_{l=1}^{n}(1 + \overline{\psi_l}(\mathbf{x}_i)\overline{\psi_l}(\mathbf{x}_j)) - \prod_{l=1}^{n}(1 + \overline{\psi_l}(\mathbf{x}_i)) - \prod_{l=1}^{n}(1 + \overline{\psi_l}(\mathbf{x}_j)).$$

This is consistent with the requirement of polynomial time summability of descendant kernels in HKL [4] and thus the active set algorithm can be employed to iteratively select a sparse set of features, since the optimality condition check (which has been discussed afterwards) in it culminates into a more efficient computation with the exponential number of summations being reduced to polynomial number of products.

We now discuss the solution to the problem defined in equation (1).The solution to equation (1) is expected to yield a sparse set of features with non-zero weights. Therefore, as illustrated in [4], the solution to equation (1) when solved with the entire set of features is the same when solved with the optimum set of features.

As the latter has lesser computational complexity, an active set algorithm can be employed, which starts with a small subset of DisjunctProjs and iteratively adds nodes that violate a sufficiency condition. The primal optimization problem with an active set of features $\mathcal{A}$ (restricted primal) can be represented as

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\Big(\sum_{v \in \mathcal{A}} \delta_v \| \mathbf{w}_{D(v) \cap \mathcal{A}} \|_p \Big)^2 + C \ \mathbf{1}^\top \boldsymbol{\xi} \quad (2)$$

$$s.t. \ \forall i: \ y_i \Big(\sum_{v \in \mathcal{A}} \langle w_v, \psi_v(\mathbf{x}_i) \rangle - b \Big) \geq 1 - \xi_i, \quad \boldsymbol{\xi} \geq 0 \quad (3)$$

To efficiently solve, we apply the variational characterization proposed in lemma 26 of [20] on the regularization term and represent equation (1) as

$$\min_{\boldsymbol{\gamma} \in \Delta_{|\mathcal{V}|,1}} \ \min_{\boldsymbol{\lambda}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} \ \min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \sum_{u \in \mathcal{V}} \sigma_u^{-1}(\boldsymbol{\gamma}, \boldsymbol{\lambda}) \| w_u \|_2^2 \ + \ C \ \mathbf{1}^\top \boldsymbol{\xi} \quad (4)$$

$$s.t. \ \forall i: \ y_i \Big(\sum_{v \in \mathcal{V}} \langle w_v, \psi_v(\mathbf{x}_i) \rangle - b \Big) \geq 1 - \xi_i, \ \boldsymbol{\xi} \geq 0 \quad (5)$$

where $\Delta_{d,r} = \{ \boldsymbol{\theta} \in \mathbb{R}^d \mid \boldsymbol{\theta} \geq 0, \ \sum_{i=1}^{d} \theta_i^r = 1 \}$, $\sigma_u^{-1}(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{v \in A(u)} \frac{\delta_v^2}{\gamma_v \lambda_{vu}}$ as defined in the lemma 26 of [20] and $A(u)$ denotes ancestors of u which includes the node u itself. By applying the representer theorem [23] on the variation characterization of the regularizer term, we can derive the partial dual of the above primal form with respect to $\mathbf{w}, b, \boldsymbol{\xi}$ alone as,

$$\min_{\boldsymbol{\gamma} \in \Delta_{|\mathcal{V}|,1}} \ \min_{\boldsymbol{\lambda}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} \ \max_{\boldsymbol{\alpha} \in \tau(y, C)} \ G(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \quad (6)$$

where

$$G(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^\top \left(\sum_{u \in \mathcal{V}} \sigma_u(\boldsymbol{\gamma}, \boldsymbol{\lambda}) \mathbf{K}_u \right) \boldsymbol{\alpha} \quad (7)$$

and $\tau(y, C) = \{ \boldsymbol{\alpha} \in \mathbb{R}^m \mid 0 \leq \boldsymbol{\alpha} \leq C, \ \sum_{i=1}^{m} y_i \ \alpha_i \ = \ 0 \}$

The final dual of equation (1) is derived as,

$$\min_{\eta \in \Delta_{|\mathcal{V}|,1}} g(\eta) \quad (8)$$

where

$$g(\eta) = \max_{\boldsymbol{\alpha} \in \tau(y, C)} \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2} \Big(\sum_{v \in \mathcal{V}} \zeta_v(\eta) (\boldsymbol{\alpha}^\top K_v \boldsymbol{\alpha})^{\bar{p}} \Big)^{\frac{1}{\bar{p}}} \quad (9)$$

and $\zeta_v(\eta) = \Big(\sum_{u \in A(v)} \delta_u^p \eta_u^{1-p} \Big)^{\frac{1}{1-p}}$ $\bar{p} = \frac{p}{2(p-1)}$.

The solution to the final dual, with $\mathcal{V}$ restricted to the active set $\mathcal{A}$ gives the solution to the restricted primal problem. To solve the problem efficiently, we employ an active set algorithm.

The active set algorithm [4] starts with an initial set of features and at every iteration, solves the dual problem with the current active features, checks a sufficiency condition on the nodes that are sources of the complement set of current active set ($sources(\mathcal{A}^c) = \{ w \in \mathcal{A}^c | Ancestor(w) \bigcap \mathcal{A}^c = \{w\} \}$, where $\mathcal{A}^c$ is the complement of $\mathcal{A}$ in $\mathcal{V}$) and adds the violating nodes to the active set. The process is continued until no new node violates the sufficiency condition. A mirror descent algorithm [5] is employed to solve the dual.We now derive the sufficiency condition that determines whether a given active set of features will yield an optimal model.

The sufficiency condition for the solution to the primal, is essentially obtained by restricting the duality gap by a threshold ϵ and is specified below. The duality gap is given by

$$\begin{aligned}
&\max_{\alpha \in \tau(y,C)} \min_{\boldsymbol{\gamma} \in \Delta_{|\mathcal{V}|,1}} \min_{\boldsymbol{\lambda}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} G(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&\qquad - \min_{\boldsymbol{\gamma} \in \Delta_{|\mathcal{V}|,1}} \min_{\boldsymbol{\lambda}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} \max_{\alpha \in \tau(y,C)} G(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&\leq \frac{1}{2}\Big(\sum_{v \in V} \| \mathbf{w}_{D(v)} \|_p^2 + C \mathbf{1}^\top \boldsymbol{\xi} - \min_{\hat{\boldsymbol{\gamma}} \in \Delta_{|\mathcal{V}|,1}} \min_{\hat{\boldsymbol{\lambda}}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} G(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}, \boldsymbol{\alpha}) \Big) \\
&\qquad = \sum_{v \in V} \| \mathbf{w}_{D(v)} \|_p^2 + C \mathbf{1}^\top \boldsymbol{\xi} - \mathbf{1}^\top \boldsymbol{\alpha}) + \frac{1}{2} \Bigg(\sum_{v \in V} \| \mathbf{w}_{D(v)} \|_p^2 \\
&\qquad\qquad - \max_{\hat{\boldsymbol{\gamma}} \in \Delta_{|\mathcal{V}|,1}} \max_{\hat{\boldsymbol{\lambda}}_v \in \Delta_{|D(v)|,\hat{p}} \forall v \in \mathcal{V}} \sum_{u \in V} \sigma_u(\gamma, \lambda) \boldsymbol{\alpha}^\top \mathbf{K}_u \boldsymbol{\alpha} \Bigg)
\end{aligned}$$

Taking the Lagrange dual and applying the Lemma 26 of [20] we derive the final form of the sufficiency condition from the upper bound of the duality gap as,

$$\max_{u \in sources(\mathcal{A}^c)} \boldsymbol{\alpha}_{\mathcal{A}}^\top Q(u) \boldsymbol{\alpha}_{\mathcal{A}} \leq \ \Big(\sum_{v \in \mathcal{A}} \delta_v \| \mathbf{w}_{D(v)} \|_p \Big)^2 + \epsilon$$

Input: Training data D, Maximum tolerance ϵ
1. Initialize Active set $\mathcal{A} = Top\ node$ in the lattice
2. Compute $\eta, \boldsymbol{\alpha}$ by solving (8)
3. **while** sufficiency condition is not satisfied, **do**
4. Add nodes from the set $sources(\mathcal{A})$ violating sufficiency condition to $\mathcal{A}$
5. Recompute $\eta, \boldsymbol{\alpha}$ by solving (8)
6. **end while**
7. **Output:** active-set $\mathcal{A}, \eta, \boldsymbol{\alpha}$

Figure 1: Active set algorithm

The $(i,j)^{th}$ component of $Q(u)$ can be simplified to

$$Q(u)_{ij} = \frac{\left(\beta^{\frac{2k}{n}}(1+\frac{1}{(1+\beta)^2})\right)^n}{\left(\beta^2(1+\frac{1}{(1+\beta)^2})\right)^{|u|}} - \prod_{\hat{u}\in u}\frac{\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_i)}{\beta^2}}{\left(1+\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_i)}{(1+\beta)^2}\right)}\prod_{l=1}^{n}\left(\beta^{\frac{2k}{n}}\left(1+\frac{\overline{\psi_l}(\mathbf{x}_i)}{(1+\beta)^2}\right)\right) - \prod_{\hat{u}\in u}\frac{\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_j)}{\beta^2}}{\left(1+\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_j)}{(1+\beta)^2}\right)}\prod_{l=1}^{n}\left(\beta^{\frac{2k}{n}}\left(1+\frac{\overline{\psi_l}(\mathbf{x}_j)}{(1+\beta)^2}\right)\right) + \prod_{\hat{u}\in u}\frac{\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_i)\overline{\psi_{\hat{u}}}(\mathbf{x}_j)}{\beta^2}}{\left(1+\frac{\overline{\psi_{\hat{u}}}(\mathbf{x}_i)\overline{\psi_{\hat{u}}}(\mathbf{x}_j)}{(1+\beta)^2}\right)}\prod_{l=1}^{n}\left(\beta^{\frac{2k}{n}}\left(1+\frac{\overline{\psi_l}(\mathbf{x}_i)\overline{\psi_l}(\mathbf{x}_j)}{(1+\beta)^2}\right)\right)$$

We now discuss the mirror descent approach to solving equation (8). For a given $\boldsymbol{\eta}$, let $\bar{\boldsymbol{\alpha}}$ be the solution to (9). Then the v^{th} sub-gradient of $g(\eta)$ can be obtained as

$$(\bigtriangledown g(\eta))_v = -\frac{\delta_v^p \eta_v^{-p}}{2\bar{p}}\Big(\sum_{u\in\mathcal{V}}\zeta_u(\eta)(\bar{\boldsymbol{\alpha}}^\top\mathbf{K}_u\bar{\boldsymbol{\alpha}})^{\bar{p}}\Big)^{\frac{1}{\bar{p}}-1}\Big(\sum_{u\in D(v)}\zeta_u(\eta)^p(\bar{\boldsymbol{\alpha}}^\top\mathbf{K}_u\bar{\boldsymbol{\alpha}})^{\bar{p}}\Big)$$

The updated $\boldsymbol{\eta}$ is then used to solve (9) in an sequential minimal optimization (SMO) style till convergence. The active set algorithm thus returns a sparse set of DisjunctProjs and their optimal weights. For settings where some background knowledge is available, the feature space can be explored more efficiently. We discuss this next.

2.1.1 Incorporating Background Knowledge

In domains such as sentiment classification, where some background knowledge about the features is available, it is possible to control the number of nodes to be considered for inclusion in each iteration of the active set algorithm and thus speed up the learning process. For instance, in sentiment classification, the background information is often derived from user-preferences on terms by modifying the Dirichlet prior or by adding some prior knowledge of word sentiments as available in SentiWordNet, HowNet *etc.* [13, 17, 19]. This helps to explore lexical properties of words. For example, words in a synonymous wordset are likely to possess similar polarities.

Since words representing similar meanings are likely to have the same part-of-speech, we can restrict our space of DisjunctProjs to disjunctions of words belonging to the same part-of-speech. This reduces the number of potential features to explore which in turn effectively speeds up the learning process. Prior information about the polarity of words can be embedded in the prior parameter δ_v of node v, in our integrated dimensionality reduction approach. This effectively helps the system to select features that are strongly related to the classification task.

To incorporate the sentiment prior of a node v, we can heuristically set δ_v to the product of the absolute sentiment score of the individual words in the disjunction corresponding to that node. *i.e.* $\delta_v = \prod_{\hat{v}\in v} SS(\hat{v})$ where $SS(\hat{v}) \in [0,1]$ is the absolute sentiment score of the word $\hat{v}$, measured as $|$*positive sentiment score - negative sentiment score*$|$, and where the sentiment score can be obtained from the SentiWordNet.

The parameter δ_v encourages selection of the more strongly polar features over the weaker ones. The motivation behind taking absolute score is that when biasing the system towards selecting sronger features, the polarity of the features need not be considered. On this account, the $Q(u)_{ij}$ term in the sufficiency condition for optimality is modified as,

$$Q(u)_{ij} = \frac{1}{\prod_{\hat{u}\in u}(\frac{SS(\hat{u})}{1+SS(\hat{u})})}\Bigg(\frac{\prod_{l=1}^{n}(1+\frac{1}{(1+SS(l))^2})}{\prod_{\hat{u}\in u}(1+(1+SS(\hat{u}))^2)} - \frac{\prod_{l=1}^{n}(1+\frac{\overline{\psi_l}(x_i)}{(1+SS(l))^2})}{\prod_{\hat{u}\in u}(1+\frac{(1+SS(\hat{u}))^2}{\overline{\psi}_{\hat{u}}(x_i)})} - \frac{\prod_{l=1}^{n}(1+\frac{\overline{\psi_l}(x_j)}{(1+SS(l))^2})}{\prod_{\hat{u}\in u}(1+\frac{(1+SS(\hat{u}))^2}{\overline{\psi_{\hat{u}}}(x_j)})} + \frac{\prod_{l=1}^{n}(1+\frac{\overline{\psi_l}(x_i)\overline{\psi_l}(x_j)}{(1+SS(l))^2})}{\prod_{\hat{u}\in u}(1+\frac{(1+SS(\hat{u}))^2}{\overline{\psi_{\hat{u}}}(x_i)\overline{\psi_{\hat{u}}}(x_j)})}\Bigg)$$

While we have presented the feasibility of incorporating background knowledge in our dimensionality reduction approach, we leave experimental validation of this idea (*e.g.*, on sentiment datasets) as future work.

3. RESULTS AND DISCUSSION

In this section, we discuss our experimental setup and compare our results with the state-of-the-art methods for dimensionality reduction. Our approach is implemented in Java and we performed our experiments on a 12-core (2.66 GHz) 64 bit AMD machine with 8 GB RAM and running Ubuntu 11.04. We report our results on two publicly available datasets, (i) a subset of datasets from the UCI data repository [11] and (ii) the 20 *Newsgroups* dataset [16].

UCI data

We performed our first set of experiments on the following data-sets: *Breast-cancer, Winconsin breast-cancer, Hepatitis, Monk-1, Monk-2, Monk-3, Transfusion, Tic-Tac-Toe* and *Vote*, from the UCI repository. Each of these datasets corresponds either to a binary or a multi-class classification problem. We performed experiments on each of the above datasets with all the wrapper-based dimensionality reduction approaches available in weka [12], a machine learning toolbox. Out of all the wrapper methods provided by weka,

		Wilcoxon Signed Rank Test(Level of Significance)	
Subset Evaluator	Search Method	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM
Correlation	BestFirst	Significant(0.05)	Significant(0.01)
	GreedyStep	Significant(0.05)	Significant(0.01)
	LinearFwd	Significant(0.05)	Significant(0.01)
	Rank	Significant(0.05)	Significant(0.01)
	SubsetSize	Significant(0.05)	Significant(0.01)
Consistent	BestFirst	Not Significant	Significant(0.01)
	GreedyStep	Significant(0.05)	Significant(0.01)
	LinearFwd	Significant(0.05)	Significant(0.01)
	Rank	Significant(0.05)	Significant(0.01)
	SubsetSize	Significant(0.05)	Significant(0.01)
Filtered	BestFirst	Not Significant	Significant(0.01)
	GreedyStep	Not Significant	Significant(0.01)
	LinearFwd	Not Significant	Significant(0.01)
	Rank	Significant(0.05)	Significant(0.01)
	SubsetSize	Not Significant	Significant(0.01)
Baseline $\mathcal{L}_2$		Significant(0.05)	Significant(0.005)
Baseline $\mathcal{L}_1$		Significant(0.05)	Significant(0.005)

Table 2: Wilcoxon Signed Rank Test (with level of significance) to indicate whether *IntegratedDim.Red.* is significantly better than Weka's Feature Subset Selection Wrapper Methods

Approach	Accuracy
$\mathcal{L}-1$ SVM	93.14%
$\mathcal{L}-2$ SVM	91.38%
MMpLSA	84.7%
DiscLDA	83.0%
MedLDA	73.12%[1]
Integ. Dim. Red.	**94.55**%

Table 3: Comparison of accuracies of different approaches on 20 *Newsgroups* **dataset.**

only those which give comparable results are reported. For each dataset, and for each choice of the wrapper, we considered two choices for the classifier: 1) a 2-norm SVM ($\mathcal{L}_2$) [8] and 2) a 1-norm SVM ($\mathcal{L}_1$) [10]. The comparison of the aforementioned methods with our dimension reduction approach is provided in Table 1.

Among the approaches that we compare against, we report the accuracies of only those (namely, Correlated Subset Evaluator, Consistency Subset Evaluator and Filtered Subset Evaluator) which perform comparably or better. For each of the subset selection approaches, various search strategies such as, BestFirst, GreedyStep, LinearFwd, Rank and SubsetSizeFwd have been employed and the results reported.

Further all accuracies (except for *Monk-1, Monk-2* and *Monk-3*), are presented as averages over 4-fold cross validation. For *Monk-1, Monk-2* and *Monk-3*, the train and test splits provided in the UCI repository have been used.

We observe that our approach (Integrated Dim. Red.) performs consistently better than most of the other approaches we compared against. On each dataset, our results are comparable with the best results among all the approaches and better in many cases, as shown by the Wilcoxon Signed Rank Test. This significance test was performed in order to compare each of the Feature Selection wrappers provided by Weka, when LibSVM (2-norm regularized SVM $\mathcal{L}_2$) and LibLinear (1-norm regularized SVM $\mathcal{L}_1$) are used for model building, the results of the same reported in Table 2. The results show that our approach is significantly better, at *0.01* level of significance, than each of the 15 Feature Subset Selection methods provided by Weka, when LibLinear is used as a model builder Correspondingly when LibSVM is used for model building, our model is found to be signifcantly better, at *0.05* level of significance, than a majority (*i.e.* 10 out of the 15) Feature Subset Selection methods, while for the remaining methods, our approach is found to have comparable performance. This indicates that our approach can leverage feature selection and model learning better than 2-norm SVM as well as 1-norm SVM, and with minimum over-fitting. Some wrappers that use ChiSquare, GainRatio, InfoGain, LSA-based, PCA-based or Relief-based attribute evaluator showed significantly worse performance and are not included in the table.

20 Newsgroups data

In order to compare against the current state-of-the-art supervised dimensionality reduction-cum-classification techniques, we evaluated our approach on the 20 *Newsgroups* dataset that contains postings to Usenet newsgroups. We apply our approach on the binary classification problem of distinguishing postings from two newsgroups *alt.atheism* and *talk.religion.misc*, which is considered to be a hard task, owing to the content similarity between them.In Table 3, we present a comparison of accuracy achieved by our approach with the best values reported by the existing approaches such as DiscLDA, MedLDA and MMpLSA on this dataset and in Table 1, the comparison has been reported on the feature selection wrappers provided by Weka.

We note that the proposed integrated dimensionality reduction approach outperforms other approaches. The number of disjunctions discovered (automatically) by our approach is 170. Some of the disjunctions reported are as follows:
$\{religion, sandvik, benedikt\}$
$\{religion, kent, benedikt\}$
$\{biblical, islam\}$
$\{atheism, historical\}$
$\{reading, writes\}$
$\{fax, run, mode\}$
$\{data, mode, graphics\}$
$\{version, order, directory\}$
$\{version, works, help, mail\}$
$\{use, interested, need\}$
$\{book, bill\}$
$\{images, mode\}$
$\{mode, algorithm\}$
$\{works, run\}$
$\{god, beliefs\}$
$\{book, edu\}$

MedLDA [30] is reported to have a best improvement ratio of 0.2 at 20 topics, over its baseline which is a two-step LDA + SVM approach as well as the baseline used in [28] which is a two-step pLSA + SVM approach. Whereas MMpLSA [28], which gives best accuracy of 84.7% at 3 topics, shows a 0.39 relative improvement ratio over its baseline *i.e.* pLSA + SVM and a 2% relative improvement over DiscLDA and claims to perform better than MedLDA consistently. DiscLDA itself has the best accuracy of 83.0% which is achieved

[1]MedLDA accuracy is not exactly reported in [30] and therefore, it has been calculated from the relative improvement ratios reported in [28]. The results of the competitor approaches are the best ones obtained by the authors by crossvalidating on the parameter, number of topics.

		Breast-cancer		Wisconsin		Hepatitis		Transfusion		Vote		Monk-1		Monk-2		Monk-3		Tic-Tac-Toe		20NewsGroups	
Subset Evaluator	Search Method	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM	$\mathcal{L}_2$SVM	$\mathcal{L}_1$SVM
Correlation	BestFirst	74.64	70.65	93.84	94.72	**95.0**	93.75	91.04	91.04	96.28	96.28	69.37	74.94	63.34	59.63	**97.22**	**97.22**	80.88	73.88	93.67	89.10
	GreedyStep	74.64	67.75	93.84	94.72	**95.0**	93.75	91.04	91.04	96.28	96.28	69.37	74.94	63.34	59.63	**97.22**	**97.22**	80.88	73.88	93.76	89.10
	LinearFwd	74.64	67.75	93.84	94.72	**95.0**	93.75	91.04	91.04	96.28	96.28	69.37	74.94	63.34	59.63	**97.22**	**97.22**	80.88	73.88	92.09	89.98
	Rank	74.64	70.65	93.84	94.72	94.15	93.75	91.04	91.04	96.28	96.28	69.37	74.94	63.34	59.63	**97.22**	**97.22**	80.88	73.88	92.26	91.38
	SubsetSize	74.64	70.65	93.84	94.72	94.15	93.75	91.04	91.04	96.28	96.28	69.37	74.94	63.34	59.63	**97.22**	**97.22**	80.88	73.88	92.09	89.98
Consistent	BestFirst	67.39	72.46	95.31	95.89	88.75	86.25	**92.37**	90.78	95.26	96.28	**100.0**	83.29	**93.27**	59.63	93.03	**97.22**	**100.0**	76.49	89.98	92.97
	GreedyStep	70.29	67.75	95.75	95.01	87.0	86.25	91.18	71.92	94.40	93.10	**100.0**	74.94	87.93	59.63	93.5	**97.22**	99.58	98.33	89.98	92.97
	LinearFwd	70.65	71.01	94.72	94.43	88.75	87.5	91.44	90.78	**97.41**	95.69	**100.0**	83.3	**93.27**	59.63	**97.22**	**97.22**	99.68	75.44	87.34	89.28
	Rank	68.48	68.48	94.57	92.08	92.5	91.25	91.57	89.84	94.6	93.1	**100.0**	74.94	91.87	59.63	93.27	**97.22**	99.68	80.45	93.67	91.91
	SubsetSize	70.65	71.01	94.72	94.43	88.75	88.75	91.04	91.04	96.28	96.28	83.3	66.59	**93.27**	59.63	**97.22**	**97.22**	99.68	75.44	87.34	89.28
Filtered	BestFirst	77.54	70.29	94.43	94.14	91.25	91.25	90.1	90.1	96.28	96.28	74.94	74.94	62.41	59.63	**97.22**	**97.22**	70.01	70.01	93.14	91.56
	GreedyStep	77.54	70.29	94.43	94.14	91.25	91.25	90.1	90.1	96.28	96.28	74.94	74.94	62.41	59.63	**97.22**	**97.22**	70.01	70.01	93.14	91.56
	LinearFwd	77.54	70.29	94.43	94.14	91.25	91.25	90.1	90.1	96.28	96.28	74.94	74.94	62.41	59.63	**97.22**	**97.22**	70.01	70.01	90.51	87.34
	Rank	77.14	70.29	94.43	94.14	88.75	92.50	91.04	91.04	96.28	96.28	74.94	74.94	62.41	59.63	**97.22**	**97.22**	70.01	70.01	84.88	85.86
	SubsetSize	**77.89**	70.29	94.43	94.14	91.25	91.25	90.1	90.1	96.28	96.28	74.94	74.94	62.41	59.63	**97.22**	**97.22**	70.01	70.01	90.51	87.34
Integrated Dim. Red.		75.36±0.49		**96.34**±0.19		91.25±0.29		91.04±0.30		96.28±0.17		**100.0**±0.0		85.15±0.38		**97.22**±0.16		**100.0**±0.0		**94.55**±0.23	
Baseline $\mathcal{L}_2$		71.01		95.89		87.50		91.17		94.40		100.0		87.93		93.50		99.58		91.38	
Baseline $\mathcal{L}_1$		70.65		96.04		86.25		89.57		93.10		74.94		61.25		97.21		98.32		93.14	

Table 1: Comparison of accuracies (in percentage) of different dimensionality reduction approaches on the UCI dataset and the 20Newsgroups *alt.atheism vs. talk.religion.misc* **problem.**

at 60 topics.

In addition to the improvement in performance, unlike other approaches that require the number of topics to be learned as an input, our approach automatically learns the number of disjunctive features. Since other methods do not discover the number of topics, they often have to resort to enumerating the classifier model's performance for different values of this parameter and have to report the number of topics that leads to the best performance in classification. Moreover, parametric approaches to determine the number of topics may not yield an optimum result, especially if there is no integrated learning of the topic detection parameters and the classification parameters.As a result, inappropriate number of topics may be used by such systems and thus can result in over-fitting, as hypothesized by the authors of [28]. We overcome this limitation by our nonparametric approach and learn an optimum number of disjunctive projections. On the other hand, since our model assimilates topic selection within the classifier and handles over-fitting by regularization in a unified manner, our approach guarantees an optimum model that performs efficient dimension reduction without compromising on the classifier performance.

We now compare our method with the other state-of-the-art embedded feature selection methods, namely the Zero-Norm SVM based method in [27] and Weighted SVM method in [21], both of which tackle the problem of feature selection in an integrated manner during classifier training. In order to do this comparion, four datasets are used, namely the Ionosphere and Wisconsin Dataset from UCI repository [11] which have been used in [21] and the two-class MicroArray datasets on Colon Cancer and B-Cell Lymphoma detection that have been used in [27]. These datasets have numerical features (except for Wisconsin dataset which has nominal features) and have been consequently booleanized to serve our purpose.

The Tables 4 and 5 comparing our method with Zero Norm SVM and Weighted SVM respectively, shows that our method is comparable in classifier performance with these previous approaches but in our case the same classifier performance is achieved by selecting much lesser number of features. In the tables 4 and 5 $Fraction\ of\ Features\ Selected$ is the fraction of the features out of the entire set that is chosen parametrically or selected non-parametrically by the algorithm. The Zero-Norm based SVM has a parameterized method for feature selection, with the number of features selection directly being the parameter, and the experiment results on each dataset are tabulated against a range of values for this parameter. their main observation is that unlike other SVM-based feature selection methods where the performance degrades, their classifier performance actually improves as the parameter value signifying the number of features is decreased till a limit, after which again increases, as indicated in the table.Whereas, in our method the value for the optimal number of features (or more appropriately disjunctions) to be selected, is obtained as a by-product of the classification algorithm which essentially allows better leveraging between feature selection and classifier training as seen in Table 4.

Similar trend is seen in Table 5 where the comparison is made against a Weighted Sparse version of SVM which also performs non-parametric feature selection integrated with model building. On both the datasets presented, the classifier performance is very similar but the notable difference is that our method again achieves comparable performance by selecting significantly lesser number of features (where each feature is disjunction of any number of basic features). The primary difference between our method and all of the previous works discussed here is the kind of feature space that is explored in the algorithm, namely the exponential-sized space of disjunctive features in the former as compared to the simple space of basic features in the latter.

4. CONCLUSION

Most existing approaches to dimensionality reduction for classification decouple the dimensionality reduction and classification phases. Some approaches are greedy while some others are parameterized, imposing a restriction on the learning system. In this paper we pose the requirement of optimal dimensionality reduction as an integrated non-parametric supervised max-margin optimization problem. We project the original features into the space of disjunctions and present algorithms inspired by the hierarchical kernel learning approach to select a sparse set of important disjunctions. We have shown analytically and empirically that our integrated approach learns optimal features in the form of interpretable

Approach	Fraction of Features Selected	Colon-Cancer	Fraction of Features Selected	B-Lymphoma
Zero-Norm-SVM	1.0	86.11%	1.0	92.13%
	0.5	85.83%	0.49	92.13%
	0.25	85.83%	0.124	93.24%
	0.125	86.39%	0.0621	93.89%
	0.0625	88.06%	0.0248	94.07%
	0.03125	88.89%	0.0124	93.24%
	0.015625	85.83%	0.006	92.13%
Integrated Dim. Red.	**0.007**	**88.71%**	**0.0065**	**92.71 %**

Table 4: Comparison of Feature Selection and Classifier Accuracies of Zero-Norm SVM and Integrated Dim.Red on Binary class microArray Datasets.

Approach	Fraction of Features Selected	Ionosphere	Fraction of Features Selected	Wisconsin Breast Cancer
Weighted SVM	0.7333	88.5%	1.0	96.31%
Integrated Dim. Red.	**0.6323**	**88.6%**	**0.7**	**96.04%**

Table 5: Comparison of Feature Selection and Classifier Accuracies of Weighted SVM and Integrated Dim. Red on UCI datasets

disjunctions of features capturing similar discriminative information for classification and leads to accurate models.

5. REFERENCES

[1] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 17–24, New York, NY, USA, 2007. ACM.

[2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.

[3] R. Arun, V. Suresh, C. Veni Madhavan, and M. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*, volume 6118 of *Lecture Notes in Computer Science*, pages 391–402. Springer Berlin / Heidelberg, 2010.

[4] Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. *Technical report, INRIA, France*, 2009.

[5] Aharon Ben-Tal and Arkadiaei Semenovich Nemirovskiaei. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[6] David M. Blei, Thomas Griffiths, Michael Jordan, and Joshua Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003.

[7] David M. Blei and Jon D. Mcauliffe. Supervised topic models. 2007.

[8] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[9] Gal Chechik. Max margin dimensionality reduction. Technical report, 2008.

[10] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[11] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[13] Yulan He. Incorporating sentiment prior knowledge for weakly-supervised sentiment analysis. *ACM Transactions on Asian Language Information Processing*, 2011. to appear.

[14] Ron Kohavi and George H. John. Wrappers for feature subset selection. *ARTIFICIAL INTELLIGENCE*, 97(1):273–324, 1997.

[15] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, pages 897–904, 2008.

[16] K. Lang. 20 newsgroups data set.

[17] Fangtao Li, Minlie Huang, and Xiaoyan Zhu. Sentiment analysis with global topics and local dependency. In Maria Fox and David Poole, editors, *AAAI*. AAAI Press, 2010.

[18] Haifeng Li, Tao Jiang, and Keshu Zhang. Efficient and robust feature extraction by maximum margin criterion. In *In Advances in Neural Information Processing Systems 16*, pages 157–165. MIT Press, 2003.

[19] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

[20] Charles Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 2005.

[21] Minh Hoai Nguyen and Fernando de la Torre. Optimal feature selection for support vector machines. *Pattern Recogn.*, 43(3):584–591, March 2010.

[22] Adler J. Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. Hierarchically supervised latent dirichlet allocation. In *NIPS*, pages 2609–2617, 2011.

[23] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *JMLR*, 9:2491–2521, 2008.

[24] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[25] Marie Szafranski and Alain Rakotomamonjy. Composite kernel learning. *ICML*, 2008.

[26] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[27] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, March 2003.

[28] Wanhong Xu. Supervising latent topic model for maximum-margin text classification and regression. In Mohammed Zaki, Jeffrey Yu, B. Ravindran, and Vikram Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6118 of *Lecture Notes in Computer Science*, pages 403–414. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-13657-3_44.

[29] Jie Yu, Qi Tian, Ting Rui, and T.S. Huang. Integrating discriminant and descriptive information for dimension reduction and classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3):372–377, march 2007.

[30] Jun Zhu, Amr Ahmed, and Eric Xing. MedLDA: Maximum Margin Supervised Topic Models. *Journal of Machine Learning Research*, 1:1–48, 2010.