Efficient Discriminative Training Method for Structured Predictions

Huizhen Yu JANEY.YU@CS.HELSINKI.FI
HIIT, University of Helsinki

Dimitri P. Bertsekas DIMITRIB@MIT.EDU
Dept. of EECS, M.I.T.

Juho Rousu JUHO.ROUSU@CS.HELSINKI.FI
Dept. of Computer Science, University of Helsinki

Keywords: large margin discriminative training, dual optimization, data-dependent linear parametrization

Abstract

We propose an efficient discriminative training method for generative models under supervised learning. In our setting, fully observed instances are given as training examples, together with a specification of variables of interest for prediction. We formulate the training as a convex programming problem, incorporating the SVM-type large margin constraints to favor parameters under which the maximum a posteriori (MAP) estimates of the prediction variables, conditioned on the rest, are close to their true values given in the training instances. The resulting optimization problem is, however, more complex than its quadratic programming (QP) counterpart resulting from the SVM-type training of conditional models, because of the presence of non-linear constraints on the parameters. We present an efficient optimization method, which combines several techniques, namely, a data-dependent reparametrization of dual variables, restricted simplicial decomposition, and the proximal point algorithm. Our method extends the one for solving the aforementioned QP counterpart, proposed earlier by some of the authors.

1. Overview

We consider discriminative training of parameters for generative models with directed acyclic graphs and with discrete-valued variables. We assume a supervised learning setting, in which fully observed instances are given as training examples, together with a specification of variables of interest to prediction. These will be called hidden variables and they may vary from instance to instance. As to which variables are considered as hidden, it is sometimes naturally determined by the model or task, as in the case of a hidden Markov model (HMM) or a classification task. However, the selection can also be made only for the purpose of discriminative training: for example, in a way emulating the idea of the "coding technique" or "pseudo-likelihood" of Besag (1974), we can select a subset of nodes of a complex Bayesian network (BN) such that their edges cover many parts of the graph, and given the rest of the nodes, the inference on the graph is relatively easy. In this work we focus primarily on the algorithmic aspects of efficient training.

We take the log-probabilities associated with the edges of the generative model as model parameters. (They can be shared across edges as in HMM and relational Bayesian networks.) We formulate the discriminative training problem as a convex programming problem. Its objective function has KL divergence terms to control the degree of deviation of the model parameters from certain given distributions, and penalty terms for the SVM-type margin violation to favor parameters under which the maximum a posteriori (MAP) estimates of the hidden variables, conditioned on the rest, are close to their true values given in the training instances. Large margin type of training criteria have been used for conditional models in recent structured prediction works (e.g., (Collins, 2002; Altun et al., 2003; Taskar et al., 2004)); there, the resulting optimization problems are primarily convex quadratic pro-

Appearing in the 6^{th} *International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

gramming (QP) problems with a large number of linear constraints due to the margin penalty. Here, we consider optimization problems that can be more complex and have non-linear constraints on the parameters, in addition to the margin constraints.

We present an efficient optimization method; its idea applies to solving a class of problems resulting from the enforcement of large margin constraints. Our method is an extension of the one proposed earlier by two of the authors (Yu & Rousu, 2007), which deals with the aforementioned QP problems resulting from non-kernelized versions of large margin training formulations. Our method differs from those of (Taskar et al., 2004; Tsochantaridis et al., 2005; Rousu et al., 2006) in its technique of handling the large number of margin constraints by a *data-dependent linear reparametrization* of dual variables. This technique reduces the dimension of the dual problem so that it is independent of the size of the prediction space, and is amenable to the use of efficient optimization methods of the feasible-direction type. For problems with additional parameter constraints, a simple example has been demonstrated in (Yu & Rousu, 2007), where the additional constraints are simple sign constraints. As the problems considered here are more complex, we enhance these techniques by combining several ideas.

In broad terms, the method we propose in this work operates at two levels. At the top level, we apply the proximal point algorithm and solve a sequence of regularized primal problems, which have nicer properties than the original problem and whose solutions converge to that of the latter. At the bottom level, we solve each regularized primal problem by dual optimization. In particular, by reparametrizing the multipliers associated with the large number of linear margin constraints, we derive an equivalent size-reduced dual problem which has an implicit polyhedral set constraint. We then use the restricted simplicial decomposition (RSD) method (Hearn et al., 1987) to deal with the set constraint, while we deal with the rest of the constraints directly. Due to space limit, in this extended abstract, we will describe only the principal aspects of our method, leaving details, variants and extensions, as well as experiments on HMMs and some UCI data sets for our full paper.

2. Formulation and Algorithm

Let $\theta_i, i \in \mathcal{I}$ be vector-valued variables, each of which corresponds to a vector of log-conditional probabilities of some variable given its parents. Let $\mathcal{K}$ index the training examples, and for $k \in \mathcal{K}$, let $\mathcal{S}_k$ denote the space of all possible value assignments of the hidden variables in the k-th example. In the case of HMM, for instance, θ_i corresponds to the log state transition/observation probabilities, and $\mathcal{S}_k$ is the space of the hidden state sequences for the k-th training trajectory. We define some shorthand notation. For $x \in \Re^d$ with components x_j, e^x denotes the vector in $\Re^d$ with components e^{x_j}. The shorthand $\mathbf{1}'x$ will be used for $\sum_{j=1}^{d} x_j$, where we treat $\mathbf{1}$ as a vector of all ones with however a varying dimension depending on x. We use θ to denote the vector consisting of the collection of $\theta_i, i \in \mathcal{I}$, and we adopt similar notation for other variables.

2.1. Primal Problem

We formulate the training problem as solving the following convex program:

$$\text{(P)} \quad \min_{\theta,\epsilon} \quad -\sum_{i\in\mathcal{I}} c_i'\theta_i + \eta \sum_{k\in\mathcal{K}} \epsilon_k \tag{1}$$

$$\text{subj.} \sum_{i\in\mathcal{I}} a_{i,k}(s)'\theta_i + b_k(s) \le \epsilon_k, \quad \forall s \in \mathcal{S}_k,\ k \in \mathcal{K} \tag{2}$$

$$\mathbf{1}'e^{\theta_i} \le 1, \quad \forall i \in \mathcal{I} \tag{3}$$

$$\theta_i \le 0, \quad \forall i \in \mathcal{I}, \quad \epsilon_k \ge 0, \quad \forall k \in \mathcal{K} \tag{4}$$

Here, ϵ denotes the collection of scalar slack variables $\epsilon_k, k \in \mathcal{K}$, and η is some positive number. For each example $k \in \mathcal{K}$, the linear constraints in (2) correspond to the SVM-type margin constraints:

$$\ln P(s, o\,;\theta) - \ln P(s^*, o\,;\theta) + l_k(s, s^*) \le \epsilon_k,$$

where (s^*, o) denotes the true values of the hidden and non-hidden variables (respectively) given by the example, s denotes a possible value assignment of the hidden variables, and l_k denotes the loss function. The term $\sum_{i\in\mathcal{I}} a_{i,k}(s)'\theta_i$ corresponds to $\ln P(s, o\,;\theta) - \ln P(s^*, o\,;\theta)$ (notice that the log joint probability is a linear function of θ), while the term $b_k(s)$ corresponds to $l_k(s, s^*)$. The constraint $\mathbf{1}'e^{\theta_i} \le 1$ in (3) *only requires the sum of the associated probabilities to be no greater than* 1. Apart from the pure convexity/algorithmic-related reason, we interpret the missing probability mass $1 - \mathbf{1}'e^{\theta_i}$, if $\mathbf{1}'e^{\theta_i} < 1$, to be the probability of the variable taking an "unknown" value, given its parents. Each term in the summation $-\sum_{i\in\mathcal{I}} c_i'\theta_i$ in the objective function (1) is derived from the KL-divergence $D(p \,\|\, q)$, where for each i, we let $q = e^{\theta_i}$, while we let $p = c_i$ be some fixed distribution, e.g., the uniform distribution or the ML estimates. (Our optimization method applies to more general choices of objective terms, which are not necessarily linear.) These KL divergence terms ensure that the solution of (P) is non-degenerate.

2.2. Reparametrization

To effectively deal with the margin constraints (2), we reparametrize the associated multipliers by a data-dependent linear transformation that makes the dimension of the dual function independent of the size of assignment space $\mathcal{S}_k$.[1] For each $k \in \mathcal{K}$, let β_k with components $\beta_k(s), s \in \mathcal{S}_k$ be the multipliers associated with the constraints (2), and let λ_i be the multipliers associated with the constraints (3) for each $i \in \mathcal{I}$. We write the dual problem equivalently in terms of reparametrized variables (μ, ω, λ) with an implicit polyhedral set constraint:

$$\text{(D)} \qquad \max_{\mu,\omega,\lambda} \ \omega - \sum_{i\in\mathcal{I}} \lambda_i + \sum_{i\in\mathcal{I}} q_i(\mu_i, \lambda_i) \qquad (5)$$

$$\text{subj. } \lambda \geq 0, \quad (\mu, \omega) \in \mathcal{D} \qquad (6)$$

where the set $\mathcal{D}$ is determined by a data-dependent linear transformation of β:

$$\mathcal{D} = \Big\{ (\mu,\omega) \,\Big|\, \mu_i = \sum_{k\in\mathcal{K}, s\in\mathcal{S}_k} \beta_k(s) a_{i,k}(s),$$
$$\omega = \sum_{k\in\mathcal{K}, s\in\mathcal{S}_k} \beta_k(s) b_k(s),$$
$$\beta_k \geq 0, \quad \mathbf{1}'\beta_k \leq \eta, \ \forall k \in \mathcal{K} \Big\} \qquad (7)$$

and the functions $q_i, i \in \mathcal{I}$ are defined by

$$q_i(\mu_i, \lambda_i) = \min_{\theta_i \leq 0} \ \Big[(\mu_i - c_i)'\theta_i + \lambda_i \mathbf{1}' e^{\theta_i} \Big]. \qquad (8)$$

At an optimal dual solution $(\mu^*, \omega^*, \lambda^*)$, an optimal primal solution θ^* can be determined from (8).

2.3. A Dual Proximal Point Algorithm

While for QP problems RSD can be applied directly to solve the reduced dual problem after reparametrization (Yu & Rousu, 2007), for our problem there is a difficulty related to the domain of the dual function (5), so we actually solve a sequence of regularized primal problems by dual optimization and reparametrization. Our algorithm can be viewed as a dual proximal point algorithm. Instead of solving (D) directly, we solve a sequence of (D_n) which is identical to (D) except that the functions q_i are replaced by q_i^n defined by:

$$q_i^n(\mu_i, \lambda_i) = \min_{\theta_i \in \Re^{d_i}} \ \Big[(\mu_i - c_i)'\theta_i + \lambda_i \mathbf{1}' e^{\theta_i} + \tfrac{\gamma_n}{2} \|\theta_i - \theta_i^n\|^2 \Big]. \qquad (9)$$

[1] We derive our reparametrization essentially from the Lagrangian function. Based on the same idea, one can derive alternative reparametrizations, including for formulations with loss-rescaled slacks and quadratic penalties (Tsochantaridis et al., 2005), as well as for partitioning the training set into working sets; these can be found in (Yu & Rousu, 2007) and our full paper.

Here, for all n, θ^n satisfies the constraints (3), with the initial θ^0 being any point, and $\gamma_n \in (0, \gamma]$ for some arbitrary positive γ. The functions q_i^n and the dual function of (D_n) are everywhere real-valued functions, so RSD can be applied directly to solve (D_n). A flexible rule for changing θ^n can be given, based on proximal point algorithm theory, for the solutions to converge to the optima of (P) and (D).

Let us add two remarks relating to efficient computation, without giving details: (i) RSD operates by making successive inner approximations of $\mathcal{D}$ and optimizing the dual function on them. The complexity of the latter step is independent of the size of the original problem. The former step corresponds to solving loss-augmented inference problems. The set $\mathcal{D}$ and its approximations are unaffected by a varying θ^n, thus, the overhead of the dual proximal point algorithm is not as high as might seem. (ii) The value, gradient and Hessian of the dual function of (D_n), needed in RSD to optimize (D_n), can be computed efficiently. The Hessian is useful for speeding up the convergence by applying a projected Newton method (Bertsekas, 1982).

References

Altun, Y., Tsochantaridis, I., & Hofmann, T. (2003). Hidden Markov support vector machines. *Proc. 20th ICML*.

Bertsekas, D. P. (1982). Projected Newton methods for optimization problems with simple constraints. *SIAM J. Control Optim.*, *20*, 221–246.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, *36*, 192–236.

Collins, M. (2002). Discriminative training methods for hidden Markov models. *Proc. Conf. Empirical Methods in Natural Language Processing*.

Hearn, D. W., Lawphongpanich, S., & Ventura, J. A. (1987). Restricted simplicial decomposition: Computation and extensions. *Math. Program. Stud.*, *31*, 99–118.

Rousu, J., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, *7*, 1601–1626.

Taskar, B., Guestrin, C., & Koller, D. (2004). Max-margin Markov networks. *Proc. NIPS 16*.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, *6*, 1453–1484.

Yu, H., & Rousu, J. (2007). An efficient method for large margin parameter optimization in structured prediction problems. Technical Report C-2007-87. Univ. Helsinki.