Selecting a sparse set of logical clauses for probabilistic binary classification

October 29, 2012

1 Problem

TODO(mwytock,wbishop): Merge with overall problem statement and introduction and rewrite in terms of logistic regression and ℓ_0/ℓ_1 penalty.

We consider the problem of learning a classifier for a label, $y \in \{0, 1\}$ with binary input features, $x \in \{0, 1\}^n$. In an optimization framework, we formulate this as choosing a function $f \in \mathcal{F}$ such that we minimize $\ell(f(x), y)$ where $\mathcal{F}$ is the class of functions under consideration and ℓ is our loss function. Typically, when n is large, we restrict $\mathcal{F}$ to linear functions with $O(n)$ parameters in order to make optimization computationally tractable.

Here, we consider the more expressive class of quadratic functions and demonstrate that when the dependence of y on x is sparse, we can recover the optimal f in time much less than $O(n^2)$.

We consider the class of log-linear models such that

$$p(y = 1|x; \theta) \propto \exp\left(\sum_i \theta_i x_i + \sum_{i,j} \theta_{ij} x_i x_j\right) \tag{1}$$

which we note has $n + n(n-1)/2 = O(n^2)$ parameters. However, we assume that our model is sparse and specifically we want to solve the optimization problem

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \;\; \ell(\theta) \\ &\text{subject to} \;\; \|\theta\|_0 \leq k \end{aligned} \tag{2}$$

It is easy to see that standard solutions to (2) require $O(n^2)$ operations since θ has $O(n^2)$ components.

2 Literature review

2.1 Ising model selection

Recent work [?] demonstrates that under certain conditions, we can consistently estimate the undirected graph associated with a discrete distribution under the Ising model. This model assumes that interactions between variables are at most pairwise, and thus the distribution can be parameterized as

$$p(z;\phi) \propto \exp\left(\sum_{i,j} \phi_{ij} z_i z_j\right) \tag{3}$$

In our case, this is insufficient for our since we are interested in estimating higher-order interactions, for example between (y, x_i, x_j) for some i and j. In fact, as is explained in the paper, if we consider an Ising model over $(y, x_i, \ldots, x_n)$, then the conditional distribution of y on $x_1, \ldots, x_n$ is directly equivalent to linear logistic regression

$$p(y|x_1, \ldots x_n) = \frac{\exp(2y \sum_{i=1}^n \theta_i x_i)}{\exp(2y \sum_{i=1}^n \theta_i x_i) + 1} \tag{4}$$

where we have abused notation slightly by formulating our problem using Ising model conventions with $y, x_1, \ldots, x_n \in \{-1, 1\}$ and θ_i representing the pairwise factor between y and x_i.

3 Theoretical analysis

In this section, we consider approaches that allow us to find optimal solutions to our optimization problem without requiring that we enumerate all possible feature combinations. First, we consider the scenario in which we are given a set of conditional independencies in the form of a Markov random field. Second, we propose a greedy iterative approach and demonstrate a simple case in which it recovers the true solution.

3.1 Markov random fields

A Markov random field represents a multivariate distribution with a graph in which the vertices represent variables and the edges represent conditional independencies. In our problem, we consider the multivariate distirbution over $(y, x_1, \ldots, x_n)$ and are interested in the neighborhood around y. Specifically, given a random field representing the true distribution we can conclude that

- y is conditionally independent of x_i if there is no edge between x_i and y.

- y may depend on $x_i x_j$ if and only if there exists a clique in the graph containing y, x_i and x_j

These observations allow us to immediately exclude many variables and combinations of variables from consideration when building a model. For example, if we are interested in learning all quadratic combinations of input features, we must consider at most $O(nd)$ where d is the maximal neighborhood in the graph which may be far smaller than n if the graph is sparse.

Although structure learning in graphical models is a hard problem in itself, recent work has demonstrated that under certain conditions we can learn the structure efficiently [?].

3.2 Greedy iterative approach

In this section, we consider a greedy iterative approach to solving (2) by first finding the best linear model and using that to find the best quadratic model.

Specifically, we follow these steps

1. Find $\theta^{(1)}$ which minimizes (2) subject to the additional constraint that θ be zero on any component corresponding to a quadratic term in x in (1)
2. Find $\theta^{(2)}$ which minimizes (2) subject to the additional constraints that for i such that $\theta^{(1)} = 0$ we have $\theta_i = 0$, and $\theta_{ij} = 0$ for all j

We note that the complexity of this approach is $O(\max(n, k^2))$, a significant improvement on $O(n^2)$ when $k << n$. Below, we explore sufficient conditions where this procedure will recover the true optimal solution such that $\theta^{(2)} = \theta^*$, the solution to (2).

3.2.1 Simple case

To start, we consider a toy example in where we have $k = 2$ and $n >> k$ with the squared error loss in the log probability space, defined by

$$\ell(\theta) = \sum_{x \in \{0,1\}^2} (\log p^* - \log p(y = 1|x; \theta))^2 \tag{5}$$

where $p^* = p(y = 1|x; \theta^*)$ is the probability under the true distribution. Furthermore, we assume that $p(y = 1|x; \theta^*) = p(y = 1|x_1, x_2; \theta^*)$ and thus the true distribution depends only on x_1 and x_2 and is thus parameterized by

$$\begin{aligned} \theta_1^* &= \log p(y = 1|x_1 = 1, x_2 = 0; \theta^*) \\ \theta_2^* &= \log p(y = 1|x_1 = 0, x_2 = 1; \theta^*) \\ \theta_{12}^* &= \log p(y = 1|x_1 = 1, x_2 = 1; \theta^*) \end{aligned} \tag{6}$$

By taking derivatives, its straightforward to see that an optimal solution to the first step of our algorithm which finds the best linear function in x is given by

$$\begin{aligned} \theta_1 &= \frac{\theta_1^* - \theta_2^* + \theta_{12}^*}{3} \\ \theta_2 &= \frac{\theta_2^* - \theta_1^* + \theta_{12}^*}{3} \end{aligned} \tag{7}$$

and $\theta_i = 0$ for $i > 2$. Furthermore, for any parameterization of this model such that $\theta_2^* \neq \theta_1^*$, this solution is unique which implies that the first step of our iterative algorithm will recover the correct support in $\theta^{(1)}$ and therefore in step 2 we will allow θ_1, θ_2 and θ_{12} to be nonzero which will gives us the optimal solution.

This analysis was done with respect to the true probabilities which in practice we would estimate from data. It is clear, that at least in this simple model we can expect our algorithm to recover the true solution in the limit of infinite data.