Cost-Sensitive Linear Regression with Costly Features

Robby Goetschalckx ROBBY.GOETSCHALCKX@CS.KULEUVEN.BE
Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan 200A 3001 Heverlee, Belgium

Scott Sanner SSANNER@NICTA.COM.AU
National ICT Australia, Tower A, 7 London Circuit, Canberra City ACT 2601, Australia

Kurt Driessens KURT.DRIESSENS@CS.KULEUVEN.BE
Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan 200A 3001 Heverlee, Belgium

1. Introduction

Linear regression models, some of the most widely used and well-studied models for regression, often assume that all features can be evaluated without difficulty, while in many real-world regression scenarios this assumption simply does not hold. For example, in a medical domain features used in linear regression may correspond to the results of various diagnostic tests which have a cost and incur discomfort to the patient, or in a financial setting a regression model often requires evaluating information-gathering features that incur time and monetary costs. These and related problems can be formalized more precisely with the notion of feature costs. Assuming that prediction error and feature costs can be measured in commensurable units, we can reformulate our linear regression problem w.r.t. a *parsimonious*[1] objective criterion: clearly there is no benefit in reducing error by using more features if the financial advantage gained by obtaining more accurate predictions does not outweigh the financial cost.

Finding the exact solution to this problem is a computationally hard problem to solve, therefore we examine efficient approximations to the optimization problem. Of particular interest are *least-angle regression* methods (Efron et al., 2004). We are able to provide a modified least angle regression approach for parsimonious linear regression called ParLiR that encourages sparsity in the feature weights according to the feature cost. As an empirical validation, we demonstrate that ParLiR provides an efficient and parsimonious solution in comparison to various other linear regression approaches. Theoretically, we were able to prove the important result that ParLiR ensures that every feature used in the regression reduces the error by at least its cost, thus proving parsimony.

[1]Def. *Parsimonious* (adjective): Exhibiting the quality of being careful with money or resources.

2. Related Work

Turney (Turney, 2000) provides an overview of ways that *cost-sensitivity* may be introduced into machine learning as well as a comprehensive bibliographical reference to historical work in this area. One of the categories considered is sensitivity to feature evaluation costs: this is the fundamental problem that we address. There has been work on the classification side of this problem, much of it specifically on decision-tree approaches, but such approaches do not easily extend to the regression of continuously varying functions on a continuous input space. As such, parsimonious extensions of linear regression to deal with feature costs as proposed in this paper comprise a novel contribution to machine learning. Finally, we note that while we do make use of least angle regression methods related to sparse linear regression methods such as *lasso* (Efron et al., 2004), these algorithms by themselves do not guarantee a parsimonious model w.r.t. feature costs.

3. Parsimonious Linear Regression

We begin with the problem formulation for parsimonious linear regression and then specify the ParLiR algorithm that efficiently approximates its solution.

Our problem is identical to the linear regression setting with the added modification that features are costly. Formally, we are given the following information: A set of input features (i.e., variables) $\mathcal{X} = \{f_1, \ldots, f_n\}$, where each feature variable $f_i \in \mathbb{R}$ (so $\mathcal{X} \in \mathbb{R}^n$), and all features have 0 mean and standard deviation 1, a positive finite cost c_i associated with each input feature f_i, a target response variable $y \in \mathbb{R}$ and a set of m data samples $\mathcal{D} = \{\langle \mathcal{X}, y \rangle\}$. Our objective is to find a linear regressor $\hat{y}(\mathcal{X}, \vec{w})$ w.r.t. data $\mathcal{D}$ and weight vector $\vec{w} = \langle w_0, \ldots w_n \rangle$ ($\vec{w} \in \mathbb{R}^{n+1}$) in the form $\hat{y}(\mathcal{X}, \vec{w}) = w_0 + \sum_{f_i \in \mathcal{X}} w_i f_i$. If we denote the set of features with non-zero weight by $\mathcal{F}_{\vec{w}}$, then we

can easily define the cost $C(\vec{w})$ of a particular selection of linear regression weights $\vec{w}$ as the following: $C(\vec{w}) = \sum_{f_i \in \mathcal{F}_{\vec{w}}} c_i$. We also define the usual average squared error function $E(\vec{w}, \mathcal{D})$ for weights $\vec{w}$ as the following $E(\vec{w}, \mathcal{D}) = \frac{1}{m} \sum_{\langle \mathcal{X}, y \rangle \in \mathcal{D}} \left(\hat{y}(\mathcal{X}, \vec{w}) - y\right)^2$.

Assuming that prediction error and feature costs can be measured in commensurable units, we now reformulate our linear regression problem of jointly minimizing both prediction error and feature cost: given input feature variables $\mathcal{X}$, target response variable y, and a target linear form $\hat{y}(\mathcal{X}, \vec{w})$ using weight variables $\vec{w}$, we define the parsimonious linear regression solution $\vec{w}$ to be a global optimum of the following unconstrained optimization problem: minimize $C(\vec{w}) + E(\vec{w}, \mathcal{D})$.

Unfortunately the above optimization objective is not convex in $\vec{w}$. Thus, we cannot apply unconstrained convex optimization techniques to directly solve this parsimonious linear regression problem. As such, good approximation algorithms are crucial to solving parsimonious linear regression problems as we discuss next.

Due to their sparsity properties, which are useful for performing implicit feature selection, we focus on a class of linear regression techniques collectively referred to as *least-angle regression* (LAR) methods, such as *lasso* and *forward stagewise regression* (Efron et al., 2004). One of the key ideas behind least angle regression is that one may perform regression by incremental line search in single feature dimensions, specifically ordering feature dimensions by the amount they correlate with the regression error of the current solution. Furthermore, doing so often yields sparse solutions when there is a restrictive constraint on the total sum of the weights. For lasso and related methods, this effectively also acts as a regularizer that combats overfitting by preventing too many features from being selected and the overall sum of weights from growing too large. Furthermore, least angle regression methods manage to closely approximate the optimal regression solutions to their respective problems formulated as quadratic programs.

For parsimonious linear regression, this single dimensional line search is an attractive approach because we can reprioritize the order in which features are selected for updating according to their correlation with the error *and* their associated feature cost. Even though the parsimonious linear regression optimization problem is quite different from the lasso objective, algorithmically, only minor modifications are required to approximate the solution to parsimonious linear regression. As such, we present a modified least angle regression algorithm called ParLiR to approximate the solution of parsimonious linear regression in Figure 1.

Parsimonious Linear Regression Approximation (ParLiR)

1. **Input:** a set $\mathcal{D}$ of m data samples represented as n m-length feature vectors $\vec{f}_1, \ldots, \vec{f}_n$, an m-length target vector $\vec{y}$, and step-size η
2. Initialize current selected feature set $\mathcal{F} = \emptyset$ and weight vector $\vec{w} = \langle w_0, w_1, \ldots, w_n \rangle$ with w_0 equal to the average value of $\vec{y}$ (this will give the residuals a mean of 0), and $w_i = 0; 1 \leq i \leq n$.
3. Define the residual vector $\vec{r}_{\vec{w}}$ for the current weight settings as $\vec{r}_{\vec{w}} = \vec{y} - \left[w_0 + \sum_{i=1}^{n} w_i \vec{f}_i\right]$
4. Repeat the following:
   (a) Calculate the cost-penalized correlation scores: $score_i = \frac{1}{m} \left| \vec{f}_i \cdot \vec{r}_{\vec{w}} \right| - \mathbb{I}[f_i \in \mathcal{F}] \sqrt{c_i}$
   (b) Find the feature f_i with the highest $score_i \geq \eta$; if no such feature found then halt and **Output:** $\vec{w}$.
   (c) **If** $f_i \notin \mathcal{F}$, let $\mathcal{F} = \mathcal{F} \cup \{f_i\}$ and let $w_i = w_i + \text{sgn}(\vec{f}_i \cdot \vec{r}_{\vec{w}}) \sqrt{c_i}$.
   (d) **Else** let $w_i = w_i + \text{sgn}(\vec{f}_i \cdot \vec{r}_{\vec{w}}) \eta$.

Figure 1. The ParLiR Algorithm

However, it is not immediately clear that modified least angle regression techniques will still produce a low-cost solution to the parsimonious linear regression problem. Therefore, we experimentally evaluate efficiency, approximation error and parsimony and prove formal theoretical guarantees on parsimony and a special case of optimality.

4. Experiments

We performed experiments on four different datasets. These datasets are available on Weka (Weka, 2004). We used the 'bodyFat', 'cholesterol', 'housing' and 'pwLinear' datasets. As there were no feature costs given for these examples, we used artificial costs, based on our own intuition: medical experiments are more costly than information such as *age* and *gender*, for example.

We multiplied the basic cost vectors by a varying cost-factor to examine how the algorithms behave for low and high costs.

All reported results are averages of 10 repetitions of 10-fold cross-validation with different partitionings in folds. The stepsize used was 0.01. We varied the cost-

factor from 0 to 10.

The algorithms used were forward Stagewise with maximal iteration number 100, forward stagewise until no feature had a correlation which was higher than the threshold (this corresponds to normal linear regression), and ParLiR.

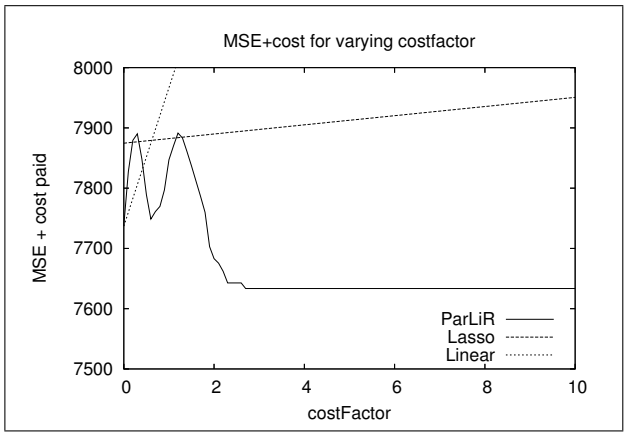

Figure 2. A comparison of Lasso, Linear Regression, and ParLiR on the *cholesterol* dataset.

We show the results on the 'cholesterol' dataset in 2. For the other datasets we observed similar behaviour. This figure shows the $C(\vec{w}) + E(\vec{w}, \mathcal{D})$ measure for varying cost-factors. Linear regression used all features irrespective of their costs. Lasso always uses the same subset of features. This implies that the costs paid by the algorithms increases linearly in terms of the cost-factor, while the mean sqaured error is constant. The comparison between Lasso and Linear regression is clear: lasso makes a higher approximation error, but the increase in cost spent on features increases less than linear. For most datasets, if the cost-factor is high, ParLiR clearly outperforms both linear and lasso methods since it is reluctant to increase the objective by using costly features. The sparsity induced by this also avoids overfitting. One can observe that for very low costs, ParLiR does a bit worse than the other algorithms. This can be explained by the fact that the algorithm only uses an approximation of the information-value of the feature, not the exact value. For a cost equal to 0, ParLiR produces the same result as linear regression showing that it does in fact produce the optimal solution in these cases even though it is using approximate least angle regression approaches.

5. Theoretical Results

We were able to prove the following two theoretical results (proofs omitted due to space considerations).

Theorem 1: Every feature which is introduced in step 4c of the ParLiR algorithm immediately reduces the mean squared error of the prediction by the value of its cost.

Theorem 2: In the simple case where feature vectors are mutually orthogonal, the ParLiR algorithm gives an approximation closer than $k\eta^2$ to the optimal solution to the parsimonious linear regression problem, with k the number of features which has a non-zero weight in the optimal solution.

6. Conclusions and Future Work

In this paper we introduced a paradigm for linear regression where features are only observable at a certain cost. We argued that finding the exact solution for this parsimonious linear regression problem is computationally hard. We introduced an adaptation of a least-angle regression algorithm, which efficiently finds an approximation to the solution. We have shown both empirically and theoretically that the solution found by this algorithm is cost-efficient.

There are various routes for future work. Linear regression is not well-suited for many domains, and more complex regression functions might be more appropriate. We have the feeling that many regression methods might be adapted to be cost-sensitive, such as artificial neural networks, support vector machines and kernel methods.

In the current setup, we use the same costly features for each example. It might be useful to employ the costless features to determine which costly features to use, thereby making the selection of features different for each sample. This possibly allows for a better resolution. Having a complex regression function which is adaptable according to the cost-free information would give a powerful, cost-efficient method for parsimonious linear regression.

References

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *ANNALS OF STATISTICS*, *32*, 407.

Turney, P. D. (2000). Types of cost in inductive concept learning. *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning (ICML-2000)*. Stanford, California.

Weka (2004). Weka 3: Data mining software in java. The University of Waikato, Dept. of Computer Science, Machine Learning lab. http://www.cs.waikato.ac.nz/ml/weka/.