Mutual Information Based Output Landmark Variable Selection

Shishir Pandey SHISHIR@TIFR.RES.IN
Rahul Vaze VAZE@TCS.TIFR.RES.IN
Tata Institute of Fundamental Research, Mumbai, 400005.

Abstract

Given a large dimensional input and output space, even simple regression is prohibitively costly. Typically, dimensionality reduction or feature selection is performed in the input space to simplify the problem. However, with modern paradigms, e.g. topic modelling, image classification, etc., even the output space is extremely large and further compounds the problem. Moreover, in contrast to input dimensionality reduction, dimensionality reduction in output is complicated. We propose a mutual information based output dimensionality reduction, that takes into account the relationship between the input and output which is essential for regression and classification problems.

1. Introduction

Machine learning deals with the finding a target function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space, after observing some finite set of samples $\{(x_i, y_i)\}$. In practice, the input space $\mathcal{X}$ can be very high dimensional. A lot of methods have been invented to deal with high dimensional input space. It is also envisionable that the output space is of high dimensions, for example topic modelling or image classification where the output might represent a number topics a document might belong to or various objects present in an image. More importantly the output variables are not completely random, they are generally correlated (for example, document tags of 'machine-learning' and 'statistics' might appear together more often than 'statistics' and 'NP hard'). Similarly, it is possible that outputs of multivariate regression problem might also be correlated. In (Balasubramanian & Lebanon, 2012) the authors build a two step model:

$$X \mapsto Y_L \tag{1}$$

$$Y_L \mapsto Y, \tag{2}$$

where $L \subseteq \{1, ..., K\}$, also known as the landmark variables (LMV), K is the dimension of the output space $\mathcal{Y}$ and $Y_L = \{Y_i : i \in L\}$. The assumption is that the non-landmark variables can be predicted from the LMV by expressing them as a sparse linear combination of LMV,

$$Y = AY_L + n. \tag{3}$$

If $|L| \ll K$ then the method can scale well for high dimension y.

In (Balasubramanian & Lebanon, 2012) the authors use the following group Lasso optimization problem to determine the set L and the matrix A:

$$\hat{A} = \underset{A \in \mathbb{R}^{k\times k}}{\arg\min} \parallel Y - YA \parallel_F^2 + \lambda_1 \parallel A \parallel_{1,2} + \lambda_2 \parallel A \parallel_1, \tag{4}$$

where $\parallel A \parallel_F \triangleq \sqrt{\text{tr} A^T A}$,
$\parallel A \parallel_{1,2} \triangleq \sum_{i=1}^{k} \sqrt{\sum_{j=1}^{k} A_{i,j}^2}$,
$\parallel A \parallel_1 \triangleq \sum_{i=1}^{k} \sum_{j=1}^{k} |A_{i,j}|$.

In (Balasubramanian & Lebanon, 2012), authors do not consider the dependence of L on the input variable X. This might not be optimal. In this work, we propose a novel approach for selecting LMV set L which is a function of the input.

2. Motivation and Formulation

To accomplish output dimensionality reduction while taking input into account, we propose a mutual information based LMV selection. Mutual information captures the amount of correlation between the input and the output. It has been used in machine learning for many purposes, e.g. feature selection, intrusion detection, etc.

In linear model the output is modelled as $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{n}$, where $\mathbf{n}$ is the noise. If $\mathbf{X}$ and $\mathbf{n}$ are Gaussian then

Appearing in Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013. Copyright 2013 by the author(s).

mutual information function

$$C(\mathcal{R}_\mathcal{L}) := \log\det\left(I + \boldsymbol{\beta}_{\mathcal{R}_\mathcal{L}}\boldsymbol{\beta}^{\dagger}_{\mathcal{R}_\mathcal{L}}\right), \qquad (5)$$

where $\boldsymbol{\beta}_{\mathcal{R}_\mathcal{L}}$ is $\boldsymbol{\beta}$ restricted to the rows in the set $\mathcal{R}_\mathcal{L}$. Selecting the components of the output $\mathbf{Y}$ is equivalent to selecting rows of the regression matrix $\boldsymbol{\beta}$. In general selecting the best subset is NP hard. However, as shown in (Vaze & Ganapathy, 2012) $C(\mathcal{R}_\mathcal{L})$ is submodular and hence a greedy algorithm can get $(1-1/e)$ of performance due to Theorem(2.1). Once $\mathcal{R}_\mathcal{L}$ is selected we use eqn.(6) to determine the expansion matrix.

$$\hat{A} = \underset{A\in\mathbb{R}^{k\times k}}{\arg\min} \parallel Y - Y_{\mathcal{R}_\mathcal{L}}A \parallel_F^2 + \lambda \parallel A \parallel_1, \qquad (6)$$

where $Y_{\mathcal{R}_\mathcal{L}}$ is restriction of rows of Y to those in set $\mathcal{R}_\mathcal{L}$.

Definition 2.1 *Let Ω be a set then a function $f : 2^\Omega \to \mathbb{R}$ is submodular if $\forall S, T \subset \Omega$ we have $f(S) + f(T) \geq f(S\cup T) + f(S\cap T)$.*

Theorem 2.1 *(Nemhauser et al., 1978) For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing greedily that provides the largest marginal increase in the function value. Let S^* be a set that maximizes the value of f over all k-element sets. Then $f(S) \geq (1-1/e)f(S^*)$.*

To recapitulate, we assume the following model $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{n}$. We learn $\boldsymbol{\beta}$ by ridge regression. Find L most useful rows greedily maximizing mutual information, $\boldsymbol{\beta}_{\mathcal{R}_\mathcal{L}}$. Then we get $Y_{\mathcal{R}_\mathcal{L}} = \boldsymbol{\beta}_{\mathcal{R}_\mathcal{L}}\mathbf{X}$. Learn the matrix A such that we minimize eqn.(6). When a new $\mathbf{X}$ arrives $\boldsymbol{\beta}_{\mathcal{R}_\mathcal{L}}\mathbf{X}A$ is the final answer.

3. Experiments

We have conducted experiments on synthetic data only. We generate the synthetic data as follows: We first generate the landmark output using random matrices, $\mathbf{X}$ is sampled from Gaussian distribution and $\boldsymbol{\beta}$ from uniform. We then use another random matrix to expand the output to a larger $\mathbf{Y}$ and Gaussian noise to it. Ridge regression without any cross-validation is done to find the regression matrix $\hat{\boldsymbol{\beta}}$ for the expanded output $\mathbf{Y}$. Greedily row subset $\mathcal{R}_\mathcal{L}$ is selected from rows of $\hat{\boldsymbol{\beta}}$, these are our LMV.

The results of simulation are shown in Fig.1. We see if the LMV assumed is less than the true LMV the RMSE between Y and the predicted $\hat{Y}$ is quite high. But, when the LMV is equal to or more than the number of LMV, then we have substantial reduction in RMSE.

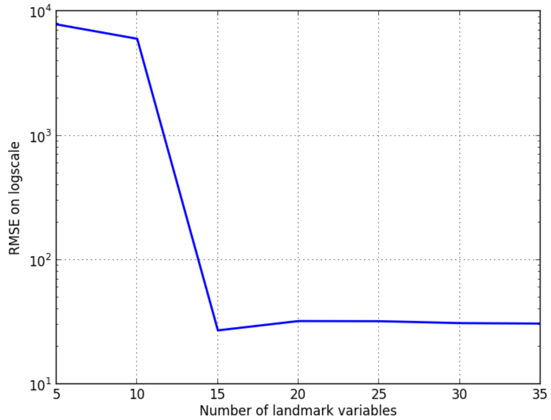

Figure 1. RMSE as the number of assumed number of LMV is changed. The true number of LMV is 15.

4. Conclusion

In multiple regression problem, if there is correlation between the output variables we can exploit this to speed up learning. Instead of making a model for each output variable we train only on the LMV and then go from LMV to the full output variable via the linear transformation. Selecting LMV is easy, since the greedy algorithms complexity is $O(LK)$ instead of $O(N^2)$ (for best subset) due to the submodularity of $C(\mathcal{R}_\mathcal{L})$. Hence, finding true number of landmark variables will help speeding up the regression process by two step model of section(1). We would also like to see how this method performs with respect to other methods for LMV selection.

References

Balasubramanian, Krishnakumar and Lebanon, Guy. The landmark selection method for multiple output prediction. In Langford, John and Pineau, Joelle (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pp. 983–990, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functionsI. *Mathematical Programming*, 14: 265–294, 1978. doi: 10.1007/BF01588971.

Vaze, R. and Ganapathy, H. Sub-modularity and antenna selection in mimo systems. *Communications Letters, IEEE*, 16(9):1446–1449, 2012. ISSN 1089-7798. doi: 10.1109/LCOMM.2012.070512.120912.