Semi-supervised learning via matrix completion

Charles Zheng and Trevor Hastie

May 20, 2015

Abstract

Matrix completion refers to the problem of inferring the missing entries of a given $m \times n$ matrix R. As such, the problem of *semi-supervised prediction* can be interpreted as a special case of matrix completion. The setting of semi-supervised learning involves an $m \times p$ matrix of features and a partially observed $m \times 1$ vector of labels; the goal is to derive a prediction rule for predicting the unobserved labels, as well as labels for newly observed features. Matrix completion can be applied in this setting to complete the combined matrix of features and labels, $R = (XY)$; here, the only missing entries are the unobserved labels. Hence, it is conceivable that existing approaches for matrix completion, such Soft-Impute (Mazumder et al 2010) could be applied to the problem of semi-supervised learning. However, generalizing to new examples requires the additional step of interpreting the result of the matrix completion as a *prediction rule*. Here we derive a prediction rule for Soft-Impute, and examine its qualities for the semi-supervised prediction problem.

1 Introduction

Suppose $R_{m\times n}$ is a partially observed matrix: we only observe R_{ij} for $(i,j) \in \Omega$. Let B denote a $m \times n$ boolean matrix where $B_{ij} = I(i,j) \in \Omega$; we will write $R \circ B$ to denote the matrix of observed entries.

Here is one way to predict the missing entries. Consider the following matrix completion problem (Mazumder et al 2010).

$$\text{minimize}_Z \frac{1}{2} \sum_{(i,j)\in\Omega} (R_{ij} - Z_{ij})^2 + \lambda ||Z||_\star$$

After solving this optimization problem, use Z_{ij} as a prediction of the missing entries R_{ij}.

Consider the solution $Z(\lambda)$ of this optimization problem. If $r = rank(Z(\lambda))$, then also $Z(\lambda) = U(\lambda)V(\lambda)^T$, where $U(\lambda)$, $V(\lambda)$ are the solutions to

$$\text{minimize}_{U_{m\times r}, V_{n\times r}} \frac{1}{2} \sum_{(i,j)\in\Omega} (R_{ij} - (UV^T)_{ij})^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2)$$

2 Semi-supervised learning

Can we interpret matrix completion as a form of semi-supervised learning? In semi-supervised learning, we have observed covariates $x^1, ..., x^m \in \mathbb{R}^p$ and *partially* observed responses $y^1, ..., y^{m_0}$ where $m_0 < m$. Let $X_{m\times p}$ denote the matrix of covariates and $Y = (y^1, ..., y^m)$ denote the full set of observed and unobserved responses.

A *supervised* approach would be to fit a model to the fully observed pairs,

$$y^i \approx x^i\beta + \beta_0$$

for $i = 1, ..., m_0$, and then predict the unobserved responses as $\hat{y}^i = x^i\beta + \beta_0$ for $i = m_0 + 1, ..., m$.

Now consider using matrix completion to solve the problem. Define the matrix

$$R_{m\times n} = [X|Y]$$

where $n = p + 1$. Here we have observed R_{ij} for all $j = 1, ..., n-1$ and for all $j = n$, $i = 1, ..., m_0$. The problem of predicting $y^{m_0+1}, ..., y^m$ is equivalent to predicting the missing elements $R^{m_0+1,n}, ..., R^{m,n}$. This approach is *semi-supervised* because it uses information from all the covariate vectors $x^1, \ldots, x^m$, not just the covariates with observed responses.

This suggests thinking of matrix completion as a method for learning a predictive model, which can be used to predict Y given X. However, it is not perfectly straightforward to interpret matrix completion as a predictive model like regression. In regression, the model gives a *prediction rule* for labelling a new observation X^*. In matrix completion, if we wanted to predict Y^* for a new observation X^*, we could do so by extending the matrix R by one row and re-running matrix completion. However, we argue that this process cannot be described as a “prediction rule” since it involves retraining the entrie model. In the following, we demonstrate that there *is* a way to interpret matrix completion in terms of a prediction rule: our proposed rule gives different results than re-running matrix completion on an extended matrix R.

3 Prediction rules for matrix completion

The problem of matrix completion can be phrased in terms of a population-based model as follows. Suppose we have a real-valued random vector $R \in \mathbb{R}^n$ with some unknown distribution $R \sim F$. Meanwhile, there also exists an n-dimensional boolean random vector B, such that R and B have a joint distribution G; alternatively, say that B has a distribution G_R conditional on R. Let $(R^1, B^1), ..., (R^m, B^m)$ be a sample of iid realizations from the joint distribution G. We then observe R_{ij} for each $(i, j) \in [m] \times [n]$ where $B_{ij} = 1$.

This yields our training set, from which we can derive a *prediction rule* for inferring missing entries of a new observation R^* for which we only observe the entries determined by B^*; more formally, a function f which maps a partially observed vector $(r_{*i_1}, ..., r_{*i_k})$ to a prediction of both observed and unboserved entries $(z_1, ..., z_n)$. Here we do not require the predictions to match on observed and unobserved entries. For notational purposes let $R^i \circ B^i$ denote the observed entries of row i, hence f maps $R^i \circ B^i$ to a prediction $Z^i \in \mathbb{R}^n$.

One recognizes that the goal of matrix completion is to find a prediction rule which minimizes the squared error of the missing entries within the sample:

$$\text{pre. error} = \sum_{i=1}^{m} \sum_{j:B_{ij}=0} (R_{ij} - f(R^i)_j)^2$$

However, it is not immediately obvious as to whether there always exists such a function f which describes the prediction made by matrix completion: i.e. f satisfying $f(R^i \circ B^i) = Z(\lambda)^i$, where $Z(\lambda)_{m\times n}$ is a minimizer of the objective

$$\text{minimize}_Z \frac{1}{2} \sum_{(i,j)\in\Omega} (R_{ij} - Z_{ij})^2 + \lambda ||Z||_\star$$

In fact, it is easy to see that such a function f always exists: just pick any function which maps $R^i \circ B^i$ to $Z(\lambda)^i$, and takes an arbitrary value anywhere else. The function is well-defined, because we can show that there exists $Z(\lambda)$ where for any $i, j \in [m]$ such that $R^i \circ B^i = R^j \circ B^j$, we have $Z(\lambda)^i = Z(\lambda)^j$.

However, now the problem is that the function f is not unique, and it also hardly resembles a proper *prediction rule* in the sense that rather than summarizing the data, it actually requires more information (order mn)

to describe than the original data. More importantly, such an arbitrarily constructed *prediction rule* can hardly be expected to generalized to new examples. To elaborate, suppose that we consider the goal of minimizing the generalization error on new examples, defined as:

$$\text{gen. error} = \mathrm{E}_{R^*,B^*} \sum_{j:B_{*j}=0} (R_{*j} - f(R^*)_j)^2$$

We resolve all three of these issues by presenting a prediction rule which is uniquely defined, which can be compactly described, and which (we will show) generalizes well under the given assumptions. To derive the prediction rule f, recall that $Z(\lambda)$ can be written as $Z(\lambda) = U(\lambda)V(\lambda)^T$, where $U(\lambda)$ and $V(\lambda)$ are the solution to

$$\text{minimize}_{U_{m\times r}, V_{n\times r}} \frac{1}{2} \sum_{(i,j)\in\Omega} (R_{ij} - (UV^T)_{ij})^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2)$$

for $r = Rank(Z(\lambda))$.

Now the key observation: supposing we only knew $V(\lambda)$, we could recover each row of $U(\lambda)$ using only information from the corresponding row of $R\circ B$. To see this, rewriting the objective function having fixed V (so we are only minimizing over U) and in terms of the individual rows, we get

$$U(\lambda) = \text{argmin}_{U_{m\times r}} \frac{1}{2} \sum_{i\in[m]} \sum_{j:(i,j)\in\Omega} (R_{ij} - (U^i V(\lambda)^T)_j)^2 + \frac{\lambda}{2}||U^i||^2$$

hence the objective function separates over rows of U, and

$$U(\lambda)^i = \text{argmin}_{\mathbb{R}^r} \sum_{j:(i,j)\in\Omega} (R_{ij} - (U^i V(\lambda)^T)_j)^2 + \lambda||U^i||^2$$

But this is simply a least squares problem. For fixed i, let $j_1, ..., j_{k_i}$ be the indices j such that $B_{ij} = 1$. Let $R^{[B^i]}$ denote a column vector consisting only of the observed entries of R^i, i.e. $R^{[B^i]} = (R_{ij_1}, ..., R_{ij_{k_i}})$. Meanwhile, let $V^{[B^i]}$ denote the $k_i \times r$ submatrix of $V(\lambda)$ with rows $V^{j_1}, ..., V^{j_{k_i}}$. Then we have

$$(U(\lambda)^i)^T = ((V^{[B^i]})^T V^{[B^i]} + \lambda I_r)^{-1} (V^{[B^i]})^T R^{[B^i]}$$

So far we have described a way of getting $U(\lambda)^i$ from $R^i \circ B^i$. In order to specify the prediction rule f, it remains to use the fact that $Z^i = U(\lambda)^i V(\lambda)^T$.

Hence our prediction rule is:

$$f(R^* \circ B^*) = [((V^{[B^*]})^T V^{[B^*]} + \lambda I_r)^{-1} (V^{[B^*]})^T R^{[B^*]}]^T V(\lambda)^T$$

This prediction rule is very general in the sense that it can take input with any missingness pattern. For the problem of semi-supervised learning, we can specialize to the missingness pattern where only the last entry of R^*, corresponding the to response, is missing. This yields a prediction rule for semi-supervised learning

$$y^* = \gamma x^*$$

where

$$\gamma = V(\lambda)^n ((V^{(-n)})^T V^{(-n)} + \lambda I_r)^{-1} (V^{(-n)})^T$$

Hence, the prediction rule for matrix completion is a special form of linear regression!