Summary of Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering

Jamel M. Thomas

April 2017

1 Summary:

In this paper, we consider a nonparametric regression algorithm to predict latent variable models. This is done by collaborative filtering, and was inspired by Netflix's movie rating system. The paper aims to give background into the algorithm, based on Taylor polynomial expansion around a point, and its connections to traditional collaborative filtering algorithms. Moreover, we give a theorem to upper bound the expected fraction of estimates with error greater than ϵ.

2 Introduction:

Blind regression is motivated by matrix completion with Netflix's movie rating system. Assume we have a partially observed matrix of m - users rows and n movies columns. We would like to complete the matrix, that is, predict the unobserved user-movie ratings given the data. Traditionally, these methods are implemented with some feature function of the data, $x_1(u)$ and $x_2(i)$, for the user u and movie i respectively. However, and issue arises in classical regression when $x = (x_1(u), x_2(i))$ is not observed. The solution proposed in this paper utilizes classical Taylor polynomial expansion estimation for unobserved functions of u and i. That is, this method only utilizes functional estimates of user u and movie i. Another issue arises as traditional collaborative filtering methods require the matrix to be low-rank, which may not be possible in real life situations. Finally, this proposed method sheds insight into the success of collaborative filtering via a nonparametric method. Some other applications of this method include social network analysis, image recovery, and predicting product demand.

2.1 Goals:

The goals of this paper include the following:

- Provide a nonparametric framework for regression over latent variables.
- Shed light into the success of collaborative filtering.
- Give the theorem of expected fraction of estimates with error greater than ϵ is γ^2/ϵ^2 plus a polynomially decaying term, for γ^2 additive entry-wise noise.

2.2 Assumptions:

Let user - u have a rating for a movie i, denoted $y(u,i)$, and let the system have a small fraction of this matrix.
Assume user - u and movie i be associated with features $x_1 \in X_1$ and $x_2 \in X_2$ with Borel σ-algebra probability measures.
Assume all entries are independently sampled with probability at least $max\{m^{\delta-1}, n^{\delta-\frac{1}{2}}\}, \delta > 0$
Assume there exists a function $f : X_1 \times X_2 \to \mathbb{R}$ such that the rating is given by $y(u,i) = f(x_1(u), x_2(i)) + \eta_{u,i}$, where the noise $\eta_{u,i}$ is independent, identically distributed, and bounded with mean 0.

3 Methods - Model Summary:

There are two parts to the model. The first involves estimating $f(x_1(u), x_2(i))$ for any unobserved pair (u,i). This uses first order local Taylor approximation expanded around points $(u,i'), (u',i),\ and\ , (u',i')$. This implies an estimate for y:

$$\hat{y}(u,i) \equiv y(u',i) + y(u,i') - y(u',i') \approx f(x_1(u), x_2(i)).$$

This model assumes $x_1(u)$ is close to $x_1(u')$ or $x_2(i')$, for some u' and i' within the matrix. However, since these variables are latent, we need another method. We would like to be able to upper bound the error terms. In the following, we upper bound the squared error of the above $\hat{y}(u,i)$ by the squared difference between commonly observed entry rows (u,v) or columns (i,j).

3.1 Set Up:

Let there be m users and n movies. The ratings for user $u \in \{1,\dots,m\}$ and $i \in \{1,\dots,n\}$ be given by $y(u,i) = f(x_1(u), x_2(i)) + \eta_{u,i}$. as before. Assume the following:

1. X_1 and X_2 are compact matrices with distance metrics d_{X_1} and d_{X_2} respectively: $d_{X_1}(x_1, x_1') \leq B_X, \forall x_1, x_1' \in X_1,\ and\ d_{X_2}(x_2, x_2') \leq B_X, \forall x_2, x_2' \in X_2$ (Bounded).
2. $f : X_1 \times X_2 \to \mathbb{R}$ is L-Lipschitz with respect to ∞ - product metric (Uniform Continuity).
3. Latent features of $x_1(u)$ and $x_2(i)$ are sampled according to Borel probability measures and are independent.
4. The noise is independent and identically distributed, bounded, with mean zero and variance γ^2.
5. The rating of entry (u,i) is revealed with probability p.

3.2 Notation:

Let $M_{u,i} = 1$ if the rating is revealed on the entry (u,i), and 0 otherwise. Let $N_1(u)$ be the set of column entries in row u. That is, the set of movie entries for user u. Moreover, let $N_2(i)$ be the set of row entries for column i. That is, the set of user entries for movie i. These are:

$$N_1(u) = \{i : M(u,i) = 1\}\ and\ N_2(i) = \{u : M(u,i) = 1\}.$$

Let $N_1(u,v) = N_1(u) \cap N_1(v), u \neq v$. That is, the commonly observed entries for row (u,v). Moreover, let $N_2(i,j) = N_2(i) \cap N_2(j), i \neq j$. That is, the commonly observed entries for columns (i,j). Call this the *overlap* between rows and columns.

4 Algorithm Preface:

In order to predict the ratings, consider the classical Taylor polynomial expansion around a point $(x_1(v), x_2(j))$ for some $u \neq v$ and $i \neq j$. That is:

$$f(x_1(u), x_2(i)) \approx f(x_1(v), x_2(j)) + (x_1(u) - x_1(v)) \frac{\partial f(x_1(v), x_2(j))}{x_1} + (x_2(i) - x_2(j)) \frac{\partial f(x_1(v), x_2(j))}{x_2}$$

However, we do not know the latent features $x_1(u)$ and $x_2(i), f$ or the partial derivative. However, if we consider the same classical Taylor expansion for $f(x_1(v), x_2(i))$ and $f(x_1(u), x_2(j))$, around $(x_1(v), x_2(j))$, we can solve the system of equations to uncover

$$f(x_1(u), x_2(i)) \approx f(x_1(v), x_2(i)) + f(x_1(u), x_2(j)) - f(x_1(v), x_2(j)).$$

This holds if the first order Taylor approximation is accurate. As stated above, we do not know the latent variables, but we can now substitute $y(u, i) = f(x_1(u), x_2(i)) + \eta_{u,i}$. If $\eta_{u,i}$ is small, we uncover $\hat{y}(u, i) = y(u, j) + y(v, i) - y(v, j)$.

5 Results 1 - Algorithm:

The following algorithm predicts the unknown entry for position (u, i) using available data. Given a parameter, $\beta \geq 2$, define the β - *overlapping* neighbors of u and i as

$$S_u^\beta(i) = \{v : v \in N_2(i), v \neq u, |N_1(u, v)| \geq \beta\}$$

$$S_i^\beta(u) = \{j : j \in N_1(u), i \neq j, |N_2(i, j)| \geq \beta\}$$

This is, in essence, the set of similar rows with overlapping agreement over the i^{th} movie, and the set of similar columns with overlapping agreement over the u^{th} row. For each $v \in S_u^\beta(i)$ compute the observed row variance between u and v, and column variance between i and j respectively.

$$s_{uv}^2 = \frac{1}{2|N_1(u, v)|(|N_1(u, v)| - 1)} \sum_{i,j \in N_1(u,v)} ((y(u, i) - y(v, i)) - (y(u, j) - y(v, j)))^2$$

$$s_{ij}^2 = \frac{1}{2|N_2(i, j)|(|N_2(i, j)| - 1)} \sum_{u,v \in N_1(i,j)} ((y(u, i) - y(v, i)) - (y(u, j) - y(v, j)))^2.$$

Let $B^\beta(u, i)$ be the set of positions (v, j) that all $y(v, j), y(u, j)$ and $y(v, i)$ are observed, with commonly observed ratings between observations (u, i) and (v, j) to be at least β.

$$B^\beta(u, i) = \{(v, j) \in S_u^\beta(i) \times S_i^\beta(u) : M(v, j) = 1\}.$$

Finally, compute the estimate:

$$\hat{y}(u, i) = \frac{\sum_{(u,i) \in B^\beta} w_{u,i}(v, j)(y(u, j) + y(v, i) - y(v, j))}{\sum_{(v,j) \in B^\beta(u,i)} w_{ui}(v, j)}$$

where the weights $w_{ui}(v, j)$ are a function of the row and column variances. The choice of this weight function will result in a different algorithm.

5.1 User-User or Item-Item Nearest Neighbor Weights:

If we evenly distribute weights among entries within a nearest neighbor window by row via empirical variance, we uncover an estimate that is asymptotically equivalent to the classical user-user nearest neighbor collaborative filtering algorithm. Similarly, if we distribute within a nearest neighbor window by column via empirical variance, we uncover the same asymptotically equivalent form of the classical item-item collaborative filtering algorithm.

5.2 User-Item Gaussian Kernel Weights:

As a variant of this algorithm, we compute the weights according to a Gaussian kernel function with bandwidth λ. $w_{vj} = exp(-\lambda min\{s_{u,v}^2, s_{i,j}^2\})$. This variant improves upon the classical user-user and item-item collaborative filtering.

6 Results 2 - Theorem

For a fixed $\epsilon > 0$, as long as $p \geq max\{m^{-1+\delta}, n^{-1/2+\delta}\}(where \delta > 0)$, for any $\rho = \omega(n^{-2\delta/3})$, the user-user nearest-neighbor variant of our method with $\beta = np^2/2$ achieves the following:

$$\mathbb{E}(Risk_\epsilon) \leq \frac{3\rho + \gamma^2}{\epsilon^2}(1 + \frac{3 * 2^{1/3}}{\epsilon} n^{-\frac{2}{3}\delta}) + O(exp(-\frac{1}{4}Cm^\delta) + m^\delta exp(-\frac{1}{5B^2} n^{\frac{2}{3}\delta}))$$

where $B = 2(LB_X + B_\eta),\ and\ C = h(\sqrt{\frac{\rho}{L^2}})^{1/6}$. This theorem indicates that the choice of β to grow with np^2 ensures that the overlap between rows goes to infinity as n goes to infinity. The essence of the indicates that the expected risk at ϵ, that is the fraction of estimates whose error is greater than ϵ, is a function of γ^2/ϵ^2 plus a monotonically decreasing polynomial function. The ρ is introduced for the purpose of analysis and is not used in the algorithm. The full proof, along with theoretical lemmas, can be found in the paper.

7 Experiments and Conclusions:

Utilizing the Gaussian weights for the final estimator and $\beta = 2$, the proposed method was competitive with user-user, item-item, and softImpute methods. Testing was preformed on two data sets, one from MovieLens and the other from Netflix. The proposed algorithm outshines both user-user and item-item methods in both tests, but shyly comes up short in the Netflix data set to softImpute. However, softImpute performed the worst on the MovieLens set.

In this paper, we introduced the concept of blind regression over latent variable models. The running example, movie rating prediction, served as a proxy for other applications. We provided the framework for regression over these latent variables, and provided some intuition into the algorithms. Moreover, we shed light into the success of traditional collaborative filtering methods, as they can be though of as a specific application of this model. Finally, we give a theorem to bound the estimates with error greater than the threshold ϵ. This work is important as the general assumptions of a matrix being "low-rank" may not be met in practice. Future work may include extending to multivariate functions f, tensor completion, and possibly capturing general noise models.