

Learning of Highly-Filtered Data Manifold Using Spectral Methods

Oleg Roderick¹ and Ilya Safro¹

Argonne National Laboratory, Argonne, IL, USA, (roderick, safro)@mcs.anl.gov*

Abstract. We propose a scheme for improving existing singular value decomposition-based tools for recovering and predicting decisions. Our main contribution is an investigation of advantages of using a functional, rather than popular linear approximation of the response of an unknown, complex model. A significant attractive feature of the method is the demonstrated ability to make predictions based on a highly filtered data set.

1 Introduction

The problem of prediction of model-constrained decisions is important in many theoretical and applied fields, such as data mining, factor analysis and uncertainty quantification. This class of problems is often formulated as an optimization problem of finding the best fit between artificial and real-life data manifolds. Since an explicit structure behind the real-life data is not available, high-quality manifold learning techniques [2] play an important role in optimization.

We propose a method that improves on the standard approaches to the problem of finding the best fit to observed data that are based on singular value decomposition (SVD) and on the extraction of latent factors [3]. The novelty of the method is twofold: its generalization of a widely used *linear representation* of the relationship between the training set and outputs, and its construction of an adaptive *high-order polynomial interpolation scheme* to estimate the relative probability of each possible outcome. Empirically, the precision of prediction grows with the increase in the order of the polynomial basis, indicating that the proposed approach is beneficial for learning and prediction problems with nonlinear relationships between model response and inputs. A distinctive feature of our method is that it treats decision outcomes as *events*, rather than *real-valued outputs* of some functional. The proposed scheme is demonstrated on a database of movie ratings, provided in [1].

2 Motivation and general scheme

The general question is how to predict the response of a model dependent on many inputs, i.e. to discover the unknown relationship $\mathcal{F}(q) : T \cup Q \rightarrow \mathbb{R}$, where T and Q are the training set and the query set, respectively, and q is a particular point. The

* This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

effects of some of the inputs on the model can be inaccessible, or very complicated. The response is then estimated based on the available observations of the model behavior for a set of inputs. A mathematical response model $F \approx \mathcal{F}$ is constructed so that the known decisions are reproduced almost perfectly, and the deviation from truth for the unknown decisions is minimal. Commonly, \mathcal{F} is described as a linear function. In modelling of nonlinear effects, however, a polynomial interpolation provides higher quality.

Latent factor methods [4] estimate the response of an unknown function based on a small number of quantities derived from the training set using essentially data compression techniques. Suppose that the most important factors $S = (s_1, s_2, \dots, s_k)$ are used to approximate the response to query q :

$$F(q, T) \sim F'(q, S),$$

where $S = S(T)$. We introduce a constructive approach to design a class of decision functions θ based on subsets of latent factors. The members of this class will simulate a work of different F' and S . The function θ will be constructed as an explicit polynomial expression on a set of factors, that is, an expansion in the polynomial basis with the coefficients obtained by regression on the training set. The steps used in our approach are: (a) application-oriented preprocessing; (b) Randomized filtering of the training set; (c) Latent factors identification; (d) Estimation of decision probabilities by polynomial approximation on latent factors; (e) Application-oriented postprocessing.

We shall formulate the proposed method using notation convenient for our applied example: a group of decision-making agents (*users*) is given access to a set of similar media or products (*content*). The users evaluate some of the units of content by assigning one of the suggested tags, or *ratings*, to each unit. Given samples of rated content, we seek to recover, or predict, the ratings that are not available.

3 Proposed method

Let $M = \{m_i\}_{i=1}^N$, $U = \{u_i\}_{i=1}^n$ and $R = \{r_i\}_{i=1}^K$ be the sets of content, users and available ratings, respectively. Denote by \mathcal{R} the binary approval of the real rating assignment (1 for "approves some r_i for the content", 0 for "does not approve some r_i content"). To meet our goal, the decision-making function θ , needs to be able to estimate a likelihood of each rating:

$$\theta : U \times M \times R \rightarrow (-\epsilon, 1 + \epsilon) \text{ (or } \theta(u_i, m_j, r_k) = \xi_r), \quad (1)$$

where $\xi_r \approx 1$ corresponds to a highly likely rating r_k and $\xi_r \approx 0$ corresponds to an unlikely rating r_k . A sufficiently small distortion ϵ is allowed to account for the interpolation error and other numerical effects. We suggest a representation of a decision of u_i on m_j is a function of other users decisions on m_j and the main optimization problem is formulated as follows.

Problem: Given $u_i, r_k, M_i \subseteq M$, construct θ such that

$$\sum_{m \in M_i} \|\theta(u_i, m, r_k) - \mathcal{R}(u_i, m, r_k)\|_2 \text{ is minimized.} \quad (2)$$

Let $\Omega_k \in \{0, 1\}^{N_i \times n}$, where $N_i = |M_i|$, be a matrix extracted from a training set $\Omega_k(p, q) = \mathcal{R}(u_p, m_q, r_k)$. As a *preprocessing* step, one has to prepare Ω_k per r_k that will participate in the decision. If the value $\mathcal{R}(u_p, m_q, r_k)$ is not available, then $\Omega_k(p, q) = 0$, with the *exception* of the position corresponding to the current query. We denote it by entry 1, to avoid the implication that every rating r_k is extremely unlikely.

The division of data with different numerical ratings into several binary fractions is explained by the relationship between the numerical values and missing entries. In frequency matrices of usual information retrieval problems, zero entries are meaningful, that is, they are comparable with very small matrix entries. This situation enables the use of the popular SVD-based methods. In the rating decision problems, however, zero has a significantly different meaning: it reflects an *event* of a missing data. Thus, once we define *event recovery* as a main goal, a better representation of data can be a set of events (that one can achieve with the binary structures Ω_k) rather than unified in one matrix numerical values.

We introduced additional uncertainty to the system by filtering out most of the known data from the originally defined Ω_k . A nonzero value related to the pair (m_p, u_q) was removed from the corresponding Ω_k with probability 0.85 if the total number of values in the respective row $\Omega_k(p, :)$ was greater than some sufficiently small threshold.

A few principal factors influencing the likelihood of r_k are extracted by SVD of Ω_k :

$$\Omega_k = \Phi \cdot \Sigma \cdot \Upsilon^T = \sum_i \phi_i \sigma_i v_i^T \quad (3)$$

where σ_i are the singular values listed in ascending order. It is well known that the optimal lower-rank approximation of Ω_k is a truncated version of the decomposition:

$$\hat{\Omega}_k = \hat{\Phi} \cdot \hat{\Sigma} \cdot \hat{\Upsilon}^T = \sum_{i=1}^{\eta} \phi_i \sigma_i v_i^T, \quad (4)$$

where η is a desired number of principal factors. The projection of query q onto a lower-dimensional subspace with the basis determined by SVD is defined as

$$\hat{q} = \hat{\Phi} \hat{\Sigma} q \in \mathbb{R}^{\eta}. \quad (5)$$

The subspace is spanned by the left singular vectors ϕ_i , defined by

$$\Omega_k \Omega_k^T \phi_i = \sqrt{\sigma_i} \phi_i. \quad (6)$$

Let $(\phi_1, \phi_2, \dots, \phi_{\eta}) = \hat{\Phi}$ be the first η columns of Φ , that is, the eigenvectors corresponding to η largest eigenvalues. The factors $S = (s_1, s_2, \dots, s_{\eta})$ determining the likelihood of rating r_k to unit of content m_j are defined as vectors

$$S = \hat{\Phi} \cdot \text{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_{\eta}}) \cdot \Omega_k(:, j), \quad (7)$$

and the desired polynomial (instead of the linear) decision-making function θ is constructed as an expansion with polynomials $\Psi = \{\psi_j(S)\}$ and coefficients $x_j \in \mathbb{R}$:

$$\theta = \sum_j x_j \psi_j(S) = \sum_j x_j \psi_j(s_1, s_2, \dots, s_{\eta}). \quad (8)$$

This basis $\Psi = \{\psi_j\}$, $\psi_j(s_1, s_2, \dots, s_\eta) = \prod_{l=1}^\eta p^{(\tau_l)}(s_l)$, consists of products of all possible combinations of single variable polynomials $p^{(\tau_l)}$ of order τ_l . A trivial basis $p^{(i)}(\alpha) = \alpha^i$ led us to sufficiently good numerical results; other choices such as Chebyshev and Hermite polynomials also worked well. The expansion coefficients x_i are obtained by solving the linear regression equations $\sum_l x_l \psi_l(S) = \mathcal{R}(u_i, m_j, r_k)$. This system of linear equations has as many right-side entries as there are known decisions of the current user. At the same time, there should be at least one equation per polynomial in the basis. Final recovering of $\mathcal{R}(u_i, m_j, r_k)$ is done by

$$r_{predicted} = \left(\sum_{k=1}^K r_k \theta(u_i, m_j, r_k) \right) / \left(\sum_{k=1}^K \theta(u_i, m_j, r_k) \right). \quad (9)$$

4 Numerical example and discussion

The numerical experiments were performed on randomly chosen data points (u_i, m_j, r_k) . During each experiment, we predicted $2 \cdot 10^5$ ratings for the points extracted from [1]. We measured the root mean-square error, $RMSE = \left(\frac{1}{n} \sum (r_{true} - r_{predicted})^2 \right)^{1/2}$.

An error obtained using simple averaging of ratings given by all users provides an RMSE upper bound of 1.05. A *linear* method (the interpolating basis consisted of only linear functions) produced an improvement to 0.90. In the experiments with higher interpolation orders (2 and 3), we observed an improvement to 0.95 and 0.90, respectively. Using even higher orders severely restricted the number of factors.

In our work, we generalized a widely used method for working with the latent factors of an information model. The generalization consists of a high-order polynomial interpolation scheme rather than linear combination. The presented algorithmic approach is highly adaptive and can be reinforced by iterative parameter learning methods. For an applied example, we introduced an event matrix model as a baseline for a latent factor methods, which can describe better a fact of missing data and successively interact with the high-order polynomial scheme. The experiments on data reinforced by introduction of additional aggressive uncertainty exhibited significant improvement in comparison to the linear method and an improvement produced by an increase in interpolation order from 2 to 3.

Overall, the method appears to be competitive in its class, requires a moderate implementation and computational cost, and can be combined with sophisticated post-processing techniques. We recommend considering the high-order interpolation scheme for data recovery and prediction algorithms that are based on latent factors.

References

1. Netflix prize problem. <http://www.netflixprize.com/>.
2. S. Kumar A. Talwalkar and H. Rowley. Large-scale manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
3. Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 707–720, London, UK, 2002. Springer-Verlag.
4. Richard L. Gorsuch. *Factor Analysis*. Hillsdale, NJ: Erlbaum, 1983.