Semi-supervised Tensor factorization

Zhuoyan Xu

Feb 21 discussion

1 Some Views of Tensor multiway clustering

1.1 Gaussian case

Consider loss function in matrix form:

$$\min_{U,A,B,C} ||X - U \times_1 A \times_2 B \times_3 C||_F^2$$

Where $X \in R^{d_1 \times d_2 \times d_3}$, $U \in R^{r_1 \times r_2 \times r_3}$.

Degree of Freedom

The degree of freedom of this problem is :

$$r_1 r_2 r_3 + r_1^{d_1} + r_2^{d_2} + r_3^{d_3} - r_1! - r_2! - r_3!$$

In this case:

- $r_1 r_2 r_3$ comes from the μs in core matrix U.
- For mode-1, each observations need to be assigned to one of the r_1 clusters, and there are total d_1 observations in mode 1, which makes $r_1^{d_1}$..
- Since for each membership matrix, when shuffling two groups and the corresponding estimated μ in U makes no change(for e.g. $UA = (UP^{-1})(PA)$), which makes $-r1!$.

Alleviate Constrain

Consider matrix form, rank(A) = r_1, rank(B) = r_2, rank(C) = r_3. For membership matrix A(similar as B,C) There are 3 constrains:

- $A \in R^{d_1 \times r_1}$.
- A = $[a_{ij}]$, $a_{ij} \in \{0, 1\}$.
- $A^T A = I$

Under full constrains, A is membership matrix. If we exclude second constrain(0,1 constrain), we got sparse matrix. If we exclude sparse constrain, we got SVD orthogonal matrix.

1.2 Binary case

loss function/criterion

Why not use

$$\min_{U,A,B,C} ||X - U \times_1 A \times_2 B \times_3 C||_F^2$$

1. Since the y is binary, cross entropy is a better loss function than MSE.
2. logit transformation make sure the output probability(after sigmoid or softmax) is in [0,1].
3. MLE estimate is asymptotically most efficient and robust. In Gaussian case, MLE estimate is also the optimal solution of MSE.

2 Extension to include covariate

Consider the following scenario:

Suppose $Y \in R^{d_1 \times d_2 \times d_3}$ is a 0,1 tensor contains the information about scholars went conferences and their keyword. The mode-1 is individuals, mode-2 is conference, mode-3 is keyword. Y = $[y_{ijk}]$, and:

$$y_{ijk} = \begin{cases} 1 & individual\ i\ publish\ at\ conference\ j\ about\ keyword\ k \\ 0 & otherwise \end{cases}$$

If elements in A,B,C are all binary(A,B,C is membership matrix), we call it **hard clustering**. Otherwise we call it **soft clustering**.

Suppose $X \in R^{d_1 \times p}$ is the additional information about individuals(mode-1 of Y), such as the partition or university or age or gender of these individuals. As for the information about university or gender, the X can be binary use Onehot encoder to denote it, as for the information age or other, the X can be continuous. This case we call it **supervised clustering**

2.1 Gaussian case

Consider the question might be hard to solve, we begin with Gaussian case, which makes elements in Y is continuous and Gaussian distribution. We still use the MSE loss, there are two ideas to solve it.

Regression

Under this condition, we just replace A with X:

$$\min_{U,A,B,C} ||Y - U \times_1 X \times_2 B \times_3 C||_F^2$$

Where $U \in R^{P \times r_2 \times r_3}$, $X \in R^{d_1 \times P}$. We just use iteration method to get result.

Penalty

Under this condition, we add a penalty:

$$\min_{U,A,B,C} ||Y - U \times_1 A \times_2 B \times_3 C||_F^2 + \alpha||A - X||_F^2 + \lambda||A||_{2,1}$$

Where $\alpha,\ \lambda$ is penalty parameter, the last penalty is regularization requires $A \in R^{d_1 \times p}$ to be sparse.

More complexity

If we still keep the dimension of A is $R^{d_1 \times r_1}$ If we suppose $W \in R^{d_1 \times C}$(divide individuals into C classes), there is a matrix coefficients $W \in R^{r_1 \times p}$ connect A and X such as:

$$X = AW$$

we have:

$$\min_{U,A,B,C} ||Y - U \times_1 A \times_2 B \times_3 C||_F^2 + \alpha||AW - X||_F^2 + \lambda||W||_{2,1}$$

This is concretely discussed in [`Bokai Cao, Chun-Ta Lu`,*Semi-supervised Tensor Factorization for Brain Network Analysis*]. This paper use ADMM to compute the result.

2.2 Binary case

Classification

Under this condition, we just replace A with X:

$$logit(\mathbb{E}[Y]) = log(\frac{\mathbb{E}[Y]}{1 - \mathbb{E}[Y]}) = U \times_1 X \times_2 B \times_3 C$$

the log-likelihood function would be:

$$l(\mu) = \sum_{i=1}^{d_1} \sum_{k=1}^{d_2} \sum_{k=1}^{d_3} \{Y_{ijk}([UXBC]_{ijk}) - log(1 + \exp{[UXBC]_{ijk}})\}$$

Where $U \in R^{P \times r_2 \times r_3}$, $X \in R^{d_1 \times P}$. We just use iteration method to get result.

Penalty

Under this condition, we add a penalty:

$$\max_{U,A,B,C} \sum_{i=1}^{d_1} \sum_{k=1}^{d_2} \sum_{k=1}^{d_3} \{Y_{ijk}([UABC]_{ijk}) - log(1 + \exp{[UABC]_{ijk}})\}$$
$$subject\ to\ A = X$$

Where $A \in R^{d_1 \times p}$ to be sparse.

More complexity

If we still keep the dimension of A is $R^{d_1 \times r_1}$ If we suppose $W \in R^{d_1 \times C}$(divide individuals into C classes), there is a matrix coefficients $W \in R^{r_1 \times p}$ connect A and X such as:

$$X = AW$$

we have:

$$\max_{U,A,B,C} \sum_{i=1}^{d_1} \sum_{k=1}^{d_2} \sum_{k=1}^{d_3} \{Y_{ijk}([UABC]_{ijk}) - log(1 + \exp [UABC]_{ijk})\}$$
$$subject\ to\ AW = X$$

,