Notes on Statistical and Machine Learning

Nonnegative Matrix Factorization

Chapter: *31* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 14, Unsupervised Learning* in Hastie, Tibshirani, and Friedman (2009).

I. Nonnegative Matrix Factorization

1. **Problem Statement:** Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a data matrix with all entries being nonnegative. We want to find matrices $\mathbf{W} \in \mathbb{R}^{n\times r}$ and $\mathbf{H} \in \mathbb{R}^{r\times p}$ such that

$$\mathbf{X} \approx \mathbf{W}\mathbf{H},$$

where we require $r \leq \max\{n, p\}$. In addition, we assume that $w_{i,k} \geq 0$ and $h_{k,j} \geq 0$ for all $i = 1, 2, \cdots, n$, $k = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, p$.

2. **Objective Function:** To obtain the desired $\mathbf{W}$ and $\mathbf{H}$, we consider the following criterion

$$L(\mathbf{W}, \mathbf{H}) := \sum_{i=1}^{n}\sum_{j=1}^{p}\big(x_{i,j}\log[\mathbf{W}\mathbf{H}]_{i,j} - [\mathbf{W}\mathbf{H}]_{i,j}\big), \tag{1}$$

where $[\mathbf{A}]_{i,j}$ denotes the (i, j)-th entry of the matrix $\mathbf{A}$. Equivalently, L above can also be expressed as

$$L(\mathbf{W}, \mathbf{H}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{i,j}\log\left(\sum_{k=1}^{r} w_{i,k}h_{k,j}\right) - \left(\sum_{k=1}^{r} w_{i,k}h_{k,j}\right)\right).$$

Notice that (1) is the log-likelihood function from a model in which $x_{i,j}$ has a Poisson distribution with the mean $[\mathbf{W}\mathbf{H}]_{i,j}$.

3. **Algorithm to Maximize (1):** Note that L is convex in $\mathbf{W}$ and $\mathbf{H}$ separately, but is *not* convex jointly in $\mathbf{W}$ and $\mathbf{H}$. A minorize-maximization algorithm is proposed.

   (a) *Minorization Function and Its Consequence:* A function $g(x, y)$ is said to minorize a function $f(x)$ if

$$g(x, y) \leq f(x), \qquad \text{and} \qquad g(x, x) = f(x),$$

   for all x, y in the domain. This is useful for maximizing f since f is nondecreasing under the update

$$x^{(s+1)} = \arg\max_{x} g(x, x^{(s)}).$$

(b) *Minorization Function for L:* For our objective function L in (1), it can be shown that a minorization function for L is

$$g(\mathbf{W}, \mathbf{H}; \mathbf{W}^{(s)}, \mathbf{H}^{(s)}) := \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{r} x_{i,j} \frac{a_{i,k,j}^{(s)}}{b_{i,j}^{(s)}} \big(\log w_{i,k} + \log h_{k,j}\big) - \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{r} w_{i,k} h_{k,j}, \tag{2}$$

where

$$a_{i,k,j}^{(s)} := w_{i,k}^{(s)} h_{k,j}^{(s)}, \qquad \text{and} \qquad b_{i,j}^{(s)} := \sum_{\ell=1}^{r} w_{i,\ell}^{(s)} h_{\ell,j}^{(s)}.$$

The key point to show $g(\mathbf{W}, \mathbf{H}; \mathbf{W}^{(s)}, \mathbf{H}^{(s)}) \le L(\mathbf{W}, \mathbf{H})$ is to use the following result: for any set of r positive values $\{y_1, y_2, \cdots, y_r\}$, any set of r positive values $\{c_1, c_2, \cdots, c_r\}$ satisfying $\sum_{k=1}^{r} c_k = 1$, we must have

$$\log\left(\sum_{k=1}^{r} y_k\right) \ge \sum_{k=1}^{r} c_k \log\left(\frac{y_k}{c_k}\right),$$

which is a consequence of Jensen's inequality. In addition, observe that

$$\sum_{k=1}^{r} \frac{a_{i,k,j}^{(s)}}{b_{i,j}^{(s)}} = 1.$$

(c) *Algorithm:* Start from $\{w_{i,k}^{(0)}\}_{i=1,\cdots,n;k=1,\cdots,r}$, $\{h_{k,j}^{(0)}\}_{j=1,\cdots,p;k=1,\cdots,r}$, update them by

$$w_{i,k}^{(s+1)} \longleftarrow w_{i,k}^{(s)} \frac{\sum_{j=1}^{p} h_{k,j}^{(s)} x_{i,j} / [\mathbf{W}^{(s)}\mathbf{H}^{(s)}]_{i,j}}{\sum_{j=1}^{p} h_{k,j}^{(s)}}, \tag{3}$$

$$h_{k,j}^{(s+1)} \longleftarrow h_{k,j}^{(s)} \frac{\sum_{i=1}^{n} w_{i,k}^{(s)} x_{i,j} / [\mathbf{W}^{(s)}\mathbf{H}^{(s)}]_{i,j}}{\sum_{i=1}^{n} w_{i,k}^{(s)}}. \tag{4}$$

Eventually, the algorithm converges to a local maximum of L.

Remark. Update equations (3) and (4) can be obtained by setting the partial derivatives of g in (2) with respect to $w_{i,k}$ and $h_{k,j}$ to 0, respectively.

II. Archetypal Analysis

1. **Main Idea:** *Archetypal analysis* approximates data points by prototypes that are themselves linear combinations of data points.

   Rather than approximating each data point by a *single* nearby prototype (like K-means clustering), archetypal analysis approximates each data point by a *convex combination* of a collection of prototypes.

   Remark. The use of a convex combination forces the prototypes to lie on the convex hull of the data cloud.

2. **Problem Statement:** Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a data matrix. We want to find matrices $\mathbf{W} \in \mathbb{R}^{n\times r}$ and $\mathbf{H} \in \mathbb{R}^{r\times p}$ such that

$$\mathbf{X} \approx \mathbf{WH},$$

where we require $r \leq n$.

We make the following assumptions:

(a) $w_{i,k} \geq 0$ and $\sum_{k=1}^{r} w_{i,k} = 1$ for all $i = 1, 2, \cdots, n$;

(b) $\mathbf{H} = \mathbf{BX}$, where $\mathbf{B} \in \mathbb{R}^{r\times n}$ satisfies $b_{k,i} \geq 0$ and $\sum_{i=1}^{n} b_{k,i} = 1$ for all $k = 1, 2, \cdots, r$.

Remark 1. By Assumption (a), the n data points (rows of $\mathbf{X}$) in p-dimensional space are represented by convex combinations of the r archetypes (rows of $\mathbf{H}$).

Remark 2. By the restrictions on $\mathbf{B}$ in Assumption (b), the archetypes themselves are convex combinations of the data points.

3. **Objective Function:** We minimize the following criterion

$$J(\mathbf{W}, \mathbf{B}) := \|\mathbf{X} - \mathbf{WBX}\|_F^2, \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The criterion J is convex in $\mathbf{W}$ and $\mathbf{B}$ separately, but *not* jointly. We can minimize J in an alternating fashion, with each separate minimization involving a convex optimization. The algorithm converges to a local minimum of J.

4. **Comparison between Nonnegative Matrix Factorization and Archetypal Analysis:**

(a) *Goals are different:*

i. Nonnegative matrix factorization aims to approximate the columns of $\mathbf{X}$, and the main output of interest are the columns of $\mathbf{W}$ representing the primary nonnegative components in the data;

ii. Archetypal analysis focuses on the approximation of the rows of $\mathbf{X}$ using the rows of $\mathbf{H}$, which represent the archetypal data points.

(b) *Assumptions on r:*

i. Nonnegative matrix factorization assumes that $r \leq p$. With $r = p$, we can get an exact reconstruction simply choosing $\mathbf{W}$ to be the data $\mathbf{X}$ with columns scaled so that they sum to 1;

ii. Archetypal analysis requires $r \leq n$, but allows $r > p$.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.