CPSC 8420 Advanced Machine Learning
Week 6: Unsupervised Learning

Dr. Kai Liu

September 22, 2020

Learning Outcomes

Our goal for today's lecture is to understand:

- PCA and Projection
- K-means and its variations
- Non-negative Matrix Factorization (NMF) with solutions through Multiplicative Updating Algorithm (MUA)
- NMF with solutions via Alternating Minimization

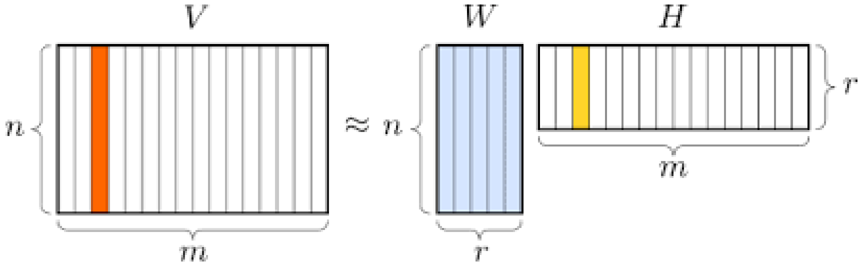
Non-Negative Matrix Factorization

Disadvantage of PCA

Consider Error Construction formulation of PCA: $\|x_i - U\lambda_i\|^2, s.t. U^T U = I$, that each data point is approximately represented by a linear combination of U_i with coefficients $\lambda_i := U^T x_i$, apparently it can be negative. However, in real-life, some operations are only additive, thus we may add non-negativeness constraint on the factor.

Another example is image processing, that each pixel should be within $[0, 255]$, negative pixel is meaningless. To enhance the interpretability, we introduce Non-negative Matrix Factorization.

Non-Negative Matrix Factorization



Non-Negative Matrix Factorization

$$\begin{bmatrix} 4.2 & 3.5 & 1 & 1.5 \\ 4 & 3.8 & 1.2 & 1.4 \end{bmatrix} \approx \begin{bmatrix} 4 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} \color{red}{0.9} & \color{red}{0.8} & 0.1 & 0.2 \\ 0.1 & 0.2 & \color{green}{0.9} & \color{green}{0.8} \end{bmatrix}$$

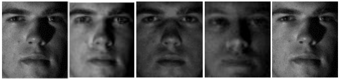
$$\approx \begin{bmatrix} 1 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 0.1 & 0.2 & \color{red}{0.9} & \color{red}{0.8} \\ \color{green}{0.9} & \color{green}{0.8} & 0.1 & 0.2 \end{bmatrix}$$

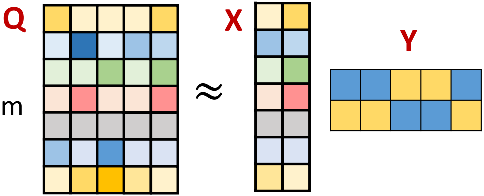
MUA for NMF

Non-negative Matrix Factorization problem:

$$\min_{X,Y \geq 0} h(X,Y) = \frac{1}{2}\|Q - XY\|_F^2$$

5 images of 2 people

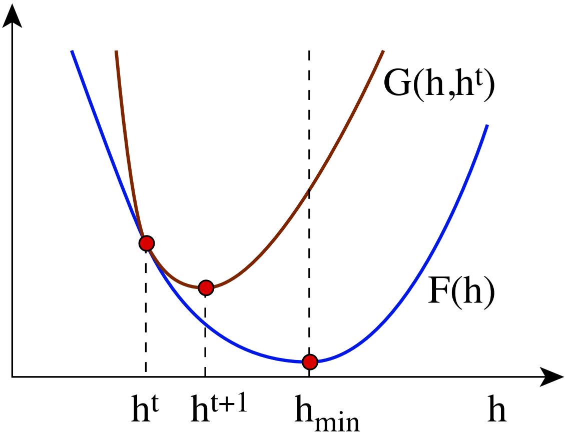




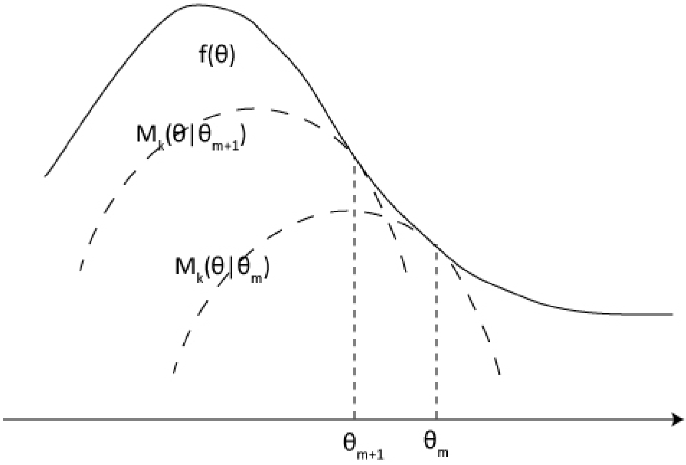
Multiplicative Updating Algorithm (MUA) (Lee & Seung, 2001):

$$Y_{ij} \leftarrow Y_{ij}\frac{(X^TQ)_{ij}}{(X^TXY)_{ij}}, \quad X_{ij} \leftarrow X_{ij}\frac{(QY^T)_{ij}}{(XYY^T)_{ij}}$$

Convergence and Majorize-Minimization



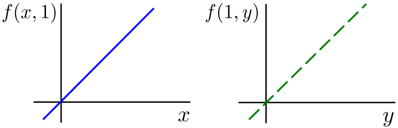
Minorize-Maximization



Alternating-Minimization

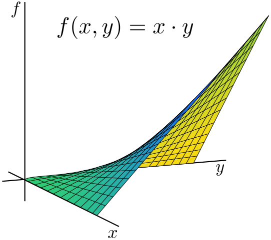
$$\min_{X,Y\geq 0} h(X,Y) = \frac{1}{2}\|Q - XY\|_F^2$$

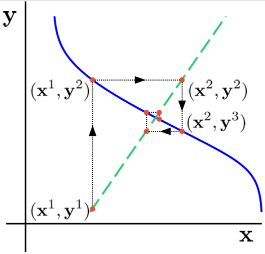
1. The NMF objective is **Nonconvex**.
2. **Convex** w.r.t. each component (X, Y)



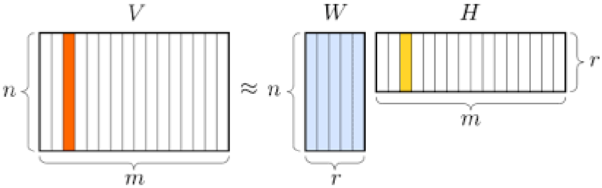
$$f(x,1) : \mathbb{R} \to \mathbb{R}$$
$$f(1,y) : \mathbb{R} \to \mathbb{R}$$

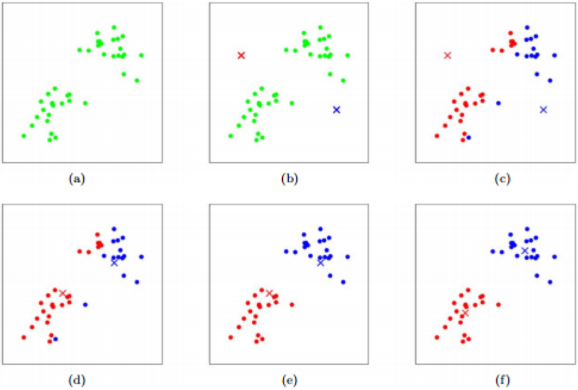
MARGINALLY CONVEX FUNCTION





K-means v.s. NMF





K-means v.s. NMF

NMF	a	b	c	d
C1	0.9	0.15	0.8	0.25
C2	0.2	0.8	0.1	0.8

K-means	a	b	c	d
C1	1	0	1	0
C2	0	1	0	1

Convergence of Gradient Descent

Strongly Convex Strongly Smooth Function

We say a continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is α-strongly convex (SC) and β-strongly smooth (SS) if for every $x, x^+ \in \mathbb{R}^p$, we have:

$$\frac{\alpha}{2}\|x^+ - x\|_2^2 \le f(x^+) - f(x) - \langle \nabla f(x), x^+ - x \rangle \le \frac{\beta}{2}\|x^+ - x\|_2^2, \quad (1)$$

based on which we will have:

$$\begin{aligned} f(x^+) &\le f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ f(x^+) &\ge f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2. \end{aligned} \quad (2)$$

Replacing x^+ with x^*, then we will have:

$$\frac{1}{2\beta}\|\nabla f(x)\|^2 \le f(x) - f(x^*) \le \frac{1}{2\alpha}\|\nabla f(x)\|^2 \quad (3)$$

Linear Convergence Rate

$$\begin{aligned} f(x^+) - f(x^*) &\leq f(x) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ &\leq f(x) - f(x^*) - \frac{\alpha}{\beta}(f(x) - f(x^*)) \\ &= (1 - \frac{\alpha}{\beta})(f(x) - f(x^*)) \end{aligned} \tag{4}$$

which implies $\frac{f(x^+)-f(x^*)}{f(x)-f(x^*)} = 1 - \frac{\alpha}{\beta}$, is the definition of linear convergence. Then to obtain ϵ-suboptimal result, we need $\mathcal{O}(log\frac{1}{\epsilon})$ iterations, which is way faster than sub-linear rate $\mathcal{O}(\frac{1}{\epsilon})$.

Non Strongly Convex

$$\begin{aligned} f(x^+) - f(x^*) &\leq f(x) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ &\leq \langle \nabla f(x), x - x^* \rangle - \frac{1}{2\beta}\|\nabla f(x)\|^2 \end{aligned} \tag{5}$$

on the other hand we have:

$$\begin{aligned} \|x^+ - x^*\|^2 &= \|x - \eta \nabla f(x) - x^*\|^2 \\ &= \|x - x^*\|^2 - 2\eta\langle \nabla f(x), x - x^* \rangle + \eta^2\|\nabla f(x)\|^2 \\ &= \|x - x^*\|^2 - 2\eta(\langle \nabla f(x), x - x^* \rangle - \frac{\eta}{2}\|\nabla f(x)\|^2), \end{aligned} \tag{6}$$

then we have
$\langle \nabla f(x), x - x^* \rangle - \frac{\eta}{2}\|\nabla f(x)\|^2 = \frac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)$, and
therefore $f(x^+) - f(x^*) \leq \frac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)$

Non Strongly Convex

Summation the equation above from $k = 0$ to $k = T - 1$, we have: $\sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 - \|x_T - x^*\|^2}{2\eta} \leq \frac{\|x_0 - x^*\|^2}{2\eta}$, then $f(x_T) - f(x^*) \leq \frac{1}{T} \sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2T\eta}$, which is sub-linear convergence rate. Then to obtain ϵ-suboptimal result, we need $T = \mathcal{O}(\frac{1}{\epsilon})$ iterations.