CPSC 8420 Advanced Machine Learning
Week 5: Unsupervised Learning

Dr. Kai Liu

September 17, 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

PCA and Projection

Computing SVD

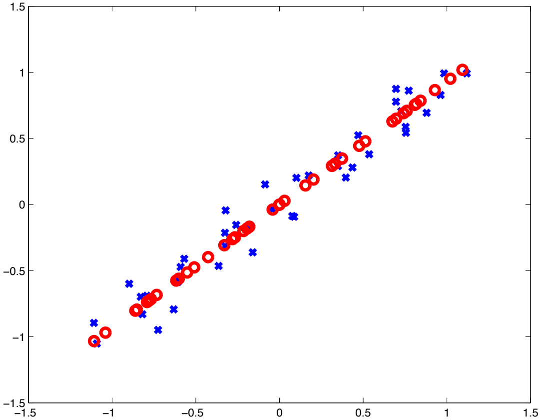
Assum we have $X \in \mathbb{R}^{n \times p}$, $[U, S, V] = svd(X^T X)$, then the complexity is $\mathcal{O}(p^3)$. And to construct $X^T X$, it will take $\mathcal{O}(p^2 n)$, if $p \gg n$, then it is likely computationally demanding.

A More Efficient Way

Now consider XX^T, suppose u is its eigenvector: $XX^T\mathrm{u} = \sigma\mathrm{u}$, then $X^TXX^T\mathrm{u} = \sigma X^T\mathrm{u}$, then immediately we know $\frac{X^T\mathrm{u}}{\|X^T\mathrm{u}\|}$ is an eigenvector of X^TX with eigenvalue of σ. We can therefore calculate the PCA solution by calculating the eigenvalues of XX^T instead of X^TX. The complexity is $\mathcal{O}(n^3)$ plus $\mathcal{O}(n^2p)$, which is significantly more efficient.

A Perspective From Reconstruction Error

In 2D space, assume each example is of the form $(x, x + y)$ where x is uniformly generated from $[-1, 1]$ and y is sampled from Gaussian distribution with mean 0 and standard deviation of 0.1. Blue x's and red circles representing before and after reconstruction.



Principal Component Regression

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Projection

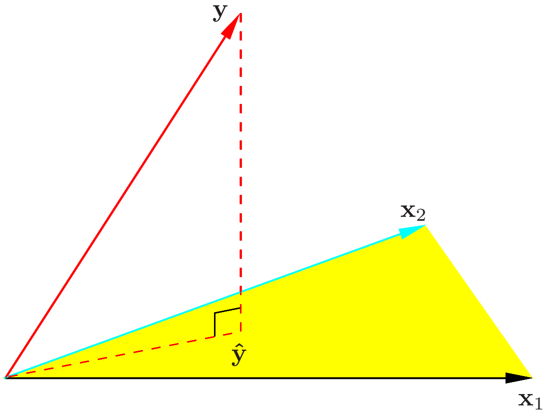
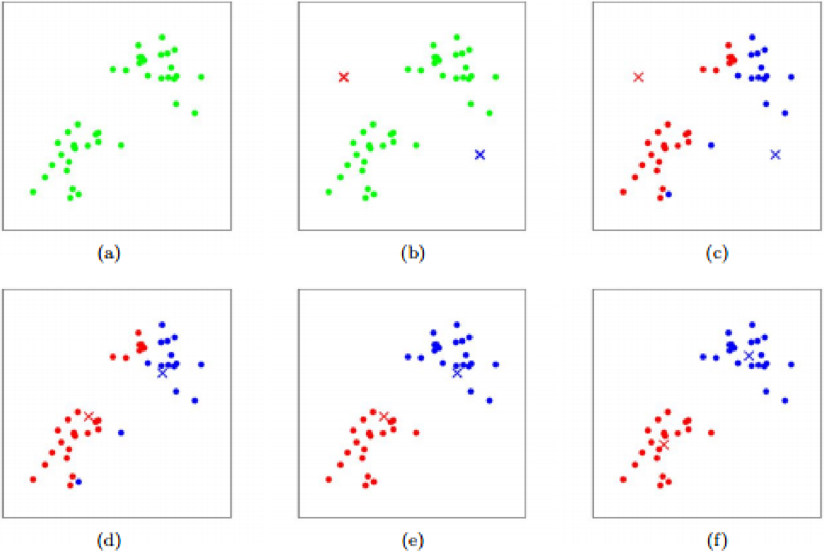


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

K-means and its Variations

Clutering with K-means


(a) (b) (c)

(d) (e) (f)

K-means Algorithm

input: $\mathcal{X} \subset \mathbb{R}^n$; Number of clusters k
initialize: Randomly choose initial centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
repeat until convergence
$\forall i \in [k]$ set $C_i = \{\mathbf{x} \in \mathcal{X} : i = \operatorname{argmin}_j \|\mathbf{x} - \boldsymbol{\mu}_j\|\}$
(break ties in some arbitrary manner)
$\forall i \in [k]$ update $\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$

K-means Objective

$$G(C_1, \ldots, C_k) = \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^n} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2.$$

Think over the question that why the objective is monotonically non-increasing in each step within the loop?

Schrödinger's cat

Though the objective is monotonically non-increasing, there is no guarantee on the number of iterations the K-means algorithm needs in order to reach convergence. In fact, K-means might converge to a point which is not even a local minimum.

Given a data set $\{1, 2, 3, 4\}$ with initial centers $\{2, 4\}$ and $K = 2$. We break ties in the definition of C_i by assigning i to be the smallest value in $argmin_j \|x - u_j\|$.

K-means Variations

- The K-medoids objective function is similar to the K-means objective, except that it requires the cluster centroids to be members of the input set.
- The K-median objective function is quite similar to the K-medoids objective, except that the 'distortion' between a data point and the centroid of its cluster is measured by distance, rather than by the square of the distance.

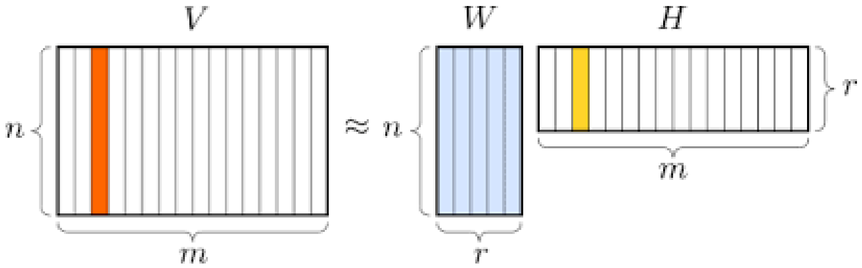
Non-Negative Matrix Factorization

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

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

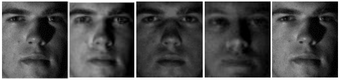
$$\approx \begin{bmatrix} 1 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 0.1 & 0.2 & \color{red}{0.9} & \color{red}{0.8} \\ \color{green}{0.9} & \color{green}{0.8} & 0.1 & 0.2 \end{bmatrix}$$

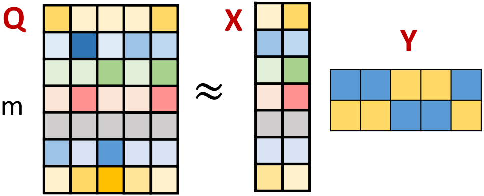
MUA for NMF

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

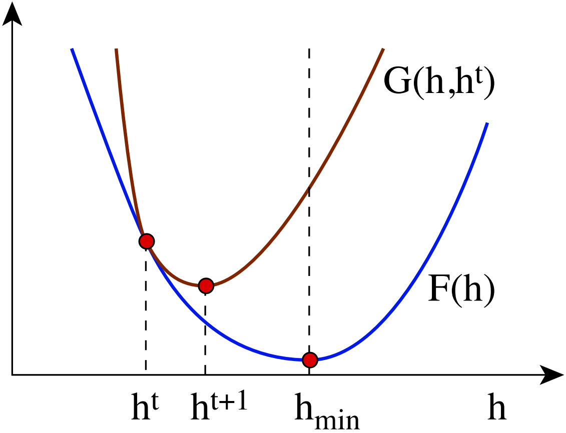




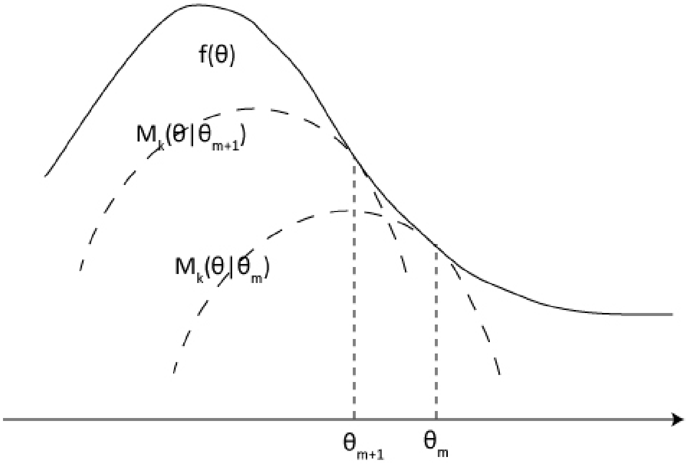
Multiplicative Updating Algorithm (MUA) (Lee & Seung, 2001):

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

Convergence and Majorize-Minimization



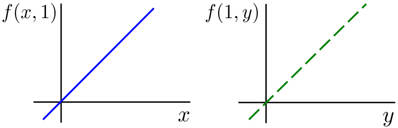
Minorize-Maximization



Alternating-Minimization

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

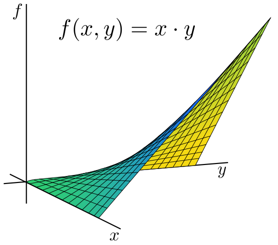
1. The NMF objective is **Nonconvex**.
2. **Convex** w.r.t. each component (X, Y)

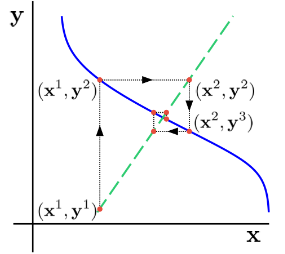


— $f(x,1) : \mathbb{R} \to \mathbb{R}$

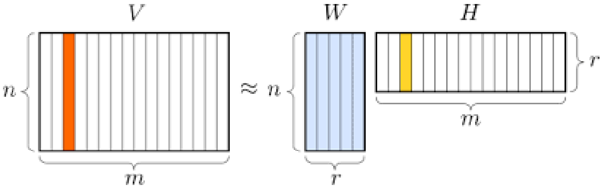
-- $f(1,y) : \mathbb{R} \to \mathbb{R}$

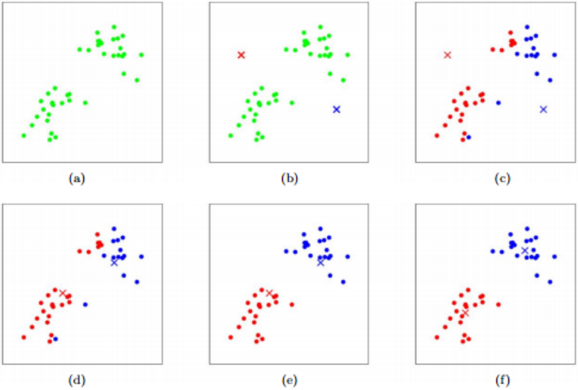
MARGINALLY CONVEX FUNCTION





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