THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Dictionary Learning and Non-negative Matrix Factorisation

Tongliang Liu

Announcements

THE UNIVERSITY OF SYDNEY

- Assignment 1 is online now
  - Assignment 1 due on 6/10/2022, 11:59pm
  - Group-based (2-3 students per group). Find you teammates by yourselves.

Assignment 1

THE UNIVERSITY OF SYDNEY

Summary

The objective of this assignment is to implement Non-negative Matrix Factorisation (NMF) algorithms and analyse the robustness of NMF algorithms when the dataset is contaminated by large magnitude noise or corruption. You should implement at least two NMF algorithms and compare their robustness.

Assignment 1

THE UNIVERSITY OF SYDNEY

Data

ORL dataset: it contains 400 images of 40 distinct subjects. All images are cropped and resized to 92×112 pixels.

Extended YaleB dataset: it contains 2414 images of 38 subjects. All images are manually aligned, cropped, and then resized to 168×192 pixels.

THE UNIVERSITY OF SYDNEY

Review

PAC learning framework

Definition:

training sample size is measured by polynomial function

efficiently probably approximately

it holds for all data (from any sample distribution)

upper bounds holds true with high probability

A hypothesis class H is said to be PAC (probably approximately correct)-learnable if there exists an learning algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution D on $X \times Y$, the following holds for any sample of size $n > poly(1/\delta, 1/\epsilon)$ and the hypothesis h_S learned by $\mathcal{A}$:

$$p\left\{R(h_S) - \min_{h \in H} R(h) \leq \epsilon\right\} \geq 1 - \delta.$$

PAC learning framework

If a hypothesis class is not learnable, it means that we cannot learn the best classifier (or classifiers close to the best) by designing a learning algorithm using a training sample with size larger the a polynomial function, e.g., $n > poly(1/\delta, 1/\epsilon)$

We need a exponentially large training sample size to learn the best classifier (or classifiers close to the best). It could be impossible to find a such a large training sample or a classical computer can handle it.

PAC-learnable checking: Empirical risk minimisation

ERM algorithm: empirical risk

$$h_S = \arg\min_{h \in H} R_S(h) = \arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h).$$

Check if the following hold when $n > poly(1/\delta, 1/\epsilon)$

$$p\left\{ R(h_S) - \min_{h \in H} R(h) \leq \epsilon \right\} \geq 1 - \delta.$$

Generalisation error

We have $R_S(h_S) \leq R_S(h^*)$, $h^* = \arg\min_{h \in H} R(h)$ then

$$
\begin{aligned}
R(h_S) - \min_{h \in H} R(h) &= R(h_S) - R(h^*) \\
&= R(h_S) - R_S(h_S) + R_S(h_S) - R_S(h^*) + R_S(h^*) - R(h^*) \\
&\leq R(h_S) - R_S(h_S) + R_S(h^*) - R(h^*) \\
&\leq |R(h_S) - R_S(h_S)| + |R(h^*) - R_S(h^*)| \\
&\leq \sup_{h \in H} |R(h) - R_S(h)| + \sup_{h \in H} |R(h) - R_S(h)| \\
&= 2 \sup_{h \in H} |R(h) - R_S(h)|.
\end{aligned}
$$

Hypothesis complexity

event A: there exist an hypothesis, sup unl $\geq \epsilon$

event B: the union of event $\geq \epsilon$ (at least 1 events)

A happens, B always happen

$p\{\sup_{h\in H} |R(h) - R_S(h)| \leq \epsilon\}$

If A replies B, then $p\{A\} \leq p\{B\}$

Union bound

$$p\{\cup_{i=1}^{n} A_i\} \leq \sum_{i=1}^{n} p\{A_i\}.$$

$$p\left\{\sup_{h\in H} |R(h) - R_S(h)| \geq \epsilon\right\}$$

(sup: event A)

$$\leq p\{\cup_{h\in H} |R(h) - R_S(h)| \geq \epsilon\}$$

(union: event B)

$$\leq \sum_{h\in H} p\{|R(h) - R_S(h)| \geq \epsilon\}$$

$$p\{|R(h) - R_S(h)| \geq \epsilon\} \leq 2\exp\left(\frac{-2n\epsilon^2}{M^2}\right)$$

$$\leq 2|H|\exp\left(\frac{-2n\epsilon^2}{M^2}\right).$$

PAC learning checking

If the hypothesis class is of finite hypotheses, it is PAC learnable. Because

$$p\left\{R(h_S) - \min_{h\in H} R(h) \leq 2\sup_{h\in H}|R(h) - R_S(h)| \leq 2M\sqrt{\frac{\log|H| + \log(2/\delta))}{2n}}\right\} \geq 1-\delta.$$

Since $\delta = 2|H|\exp\left(\frac{-2n\epsilon^2}{M^2}\right)$. We have

$$n = \frac{M^2}{\epsilon^2}\log\left(\frac{2|H|}{\delta}\right).$$

$$n > poly(1/\delta, 1/\epsilon)$$

PAC learning framework

Definition:

A hypothesis class H is said to be PAC (probably approximately correct)-learnable if there exists an learning algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution D on $X \times Y$, the following holds for any sample of size $n > poly(1/\delta, 1/\epsilon)$ and the hypothesis h_S learned by $\mathcal{A}$:

$$p\left\{R(h_S) - \min_{h \in H} R(h) \leq \epsilon\right\} \geq 1 - \delta.$$

VC dimension

$$
\begin{aligned}
& p\left\{\sup_{h\in H}|R_S(h)-R(h)|\geq\epsilon\right\} \\
& \leq 2p\left\{\sup_{h\in H}|R_S(h)-R_{S'}(h)|\geq\epsilon/2\right\} \\
& \leq 4p\left\{\sup_{h\in H}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\ell(X_i,Y_i,h)\right|\geq\epsilon/4\right\} \\
& \leq 4p\left\{\cup_{h\in H'}\frac{1}{n}\left|\sum_{I=1}^{n}\sigma_i\ell(X_i,Y_i,h)\right|\geq\epsilon/4\right\} \\
& \leq 4\Pi_H(n)p\left\{\frac{1}{n}\left|\sum_{I=1}^{n}\sigma_i\ell(X_i,Y_i,h)\right|\geq\epsilon/4\right\} \\
& \leq 8\Pi_H(n)\exp(-n\epsilon^2/32M^2).
\end{aligned}
$$

hypothesis ⇓ just consider finite prediction output ⇓ find representative

VC dimension

Let H be a hypothesis set with $\mathrm{VC\ dimension}(H) = d$ then for all $n \geq d$

$$\Pi_H(n) \leq \left(\frac{en}{d}\right)^d.$$

The proof is in Chapter 3 of the book "Foundations of ML"

VC dimension

Definition:
VC dimension
The VC dimension of a hypothesis class H is the size of the largest set that can be fully shattered by H:

$$\text{VC dimension}(H) = \max_n \{n : \Pi_H(n) = 2^n\}.$$

Note that VC dimension is designed for binary classification problem. Why?

VC dimension

Definition:

Shattering

The data points $\{X_1, \ldots, X_n\}$ is said to be shattered by a hypothesis class H when H realises all possible binary predictions. That is $\Pi_H(n) = 2^n$.

PAC learning checking

If the hypothesis class is of finite VC dimension, it is PAC learnable. Because

finite: $\log|H|$ replaced by $d\log(en/d)$

$$p\left\{R(h_S) - \min_{h\in H} R(h) \leq 2\sup_{h\in H}|R(h) - R_S(h)| \leq 2M\sqrt{\frac{32(d\log(en/d) + \log(8/\delta))}{n}}\right\} \geq 1-\delta.$$

Since $\delta = 8\left(\frac{en}{d}\right)^d \exp(-n\epsilon^2/32M^2)$, we have

$$n = \frac{32M^2}{\epsilon^2}(d\log(en/d) + \log(8/\delta)).$$

$$n > poly(1/\delta, 1/\epsilon)$$

Generalisation error

We have proven that

$$R(h_S) \leq R(h^*) + 2 \sup_{h \in H} |R(h) - R_S(h)|.$$

estimation error

$R(h_S) - R(h^*) \leq 2 \sup_{h \in H} |R(h) - R_S(h)|$

generalisation error

generalisation error upper bound

Some discussion on the differences between training error and estimation error.

Generalisation bound

Why the bound is useful?

To prove PAC learnable.
To analyse the test error.
To improve learning algorithm, e.g., if we can prove

$$R(h_S) \leq R(h^*) + 2 \sup_{h \in H} |R(h) - R_S(h)|$$

$$\triangleq UB(h_S).$$

We could propose a new algorithm,

$$h_S = \arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h) + UB(h),$$

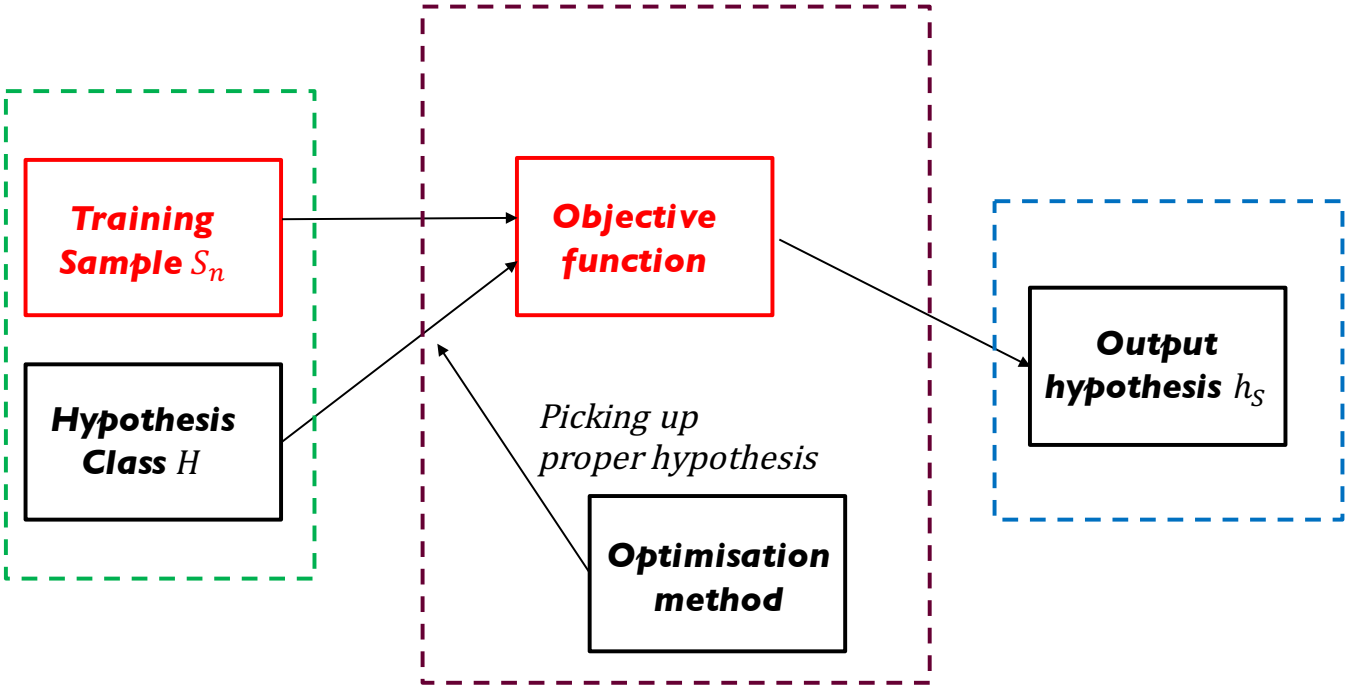
which is call the regularised ERM algorithm.

PAC learning framework

For a given learning algorithm, the PAC learning framework can be used to calculate how many examples we will need such that for a given confidence level, e.g., with probability at least $1 - \delta$, and a given approximation error ϵ, the learned hypothesis h_S satisfies $R(h_S) - R(h^*) \leq \epsilon$.

Note that the claim holds for all distributions of the training data, including the worst case. This means the bound is loose for some applications. We may need a smaller training sample size.

Machine Learning Algorithms



THE UNIVERSITY OF SYDNEY

Dictionary learning

unsupervised.

Dictionary learning

What is a dictionary in machine learning?

A dictionary is a collection of words in one specific languages.

Can we find some common "words" (elements) to express data?

Dictionary learning

What is a dictionary in machine learning?

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&2&2&4\\1&2&1&2\\1&2&2&4\\1&2&1&2\end{bmatrix}=\begin{bmatrix}1&2\\1&1\\1&2\\1&1\end{bmatrix}\begin{bmatrix}1&2&0&0\\0&0&1&2\end{bmatrix}$$

$R^{4\times4}$ $R^{4\times2}$ $R^{2\times4}$

1 nature basis

dictionary

representation matrix

$$\begin{bmatrix}1&2\\1&1\\1&2\\1&1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix}=a\times\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+b\times\begin{bmatrix}2\\1\\2\\1\end{bmatrix}$$

new basis, can represent the original $R^{4\times4}$ matrix

dictionary

Dictionary learning

What is a dictionary in machine learning?

Let $x \in \mathbb{R}^d$, $D \in \mathbb{R}^{d \times k}$

d × k dictionary

k × 1 new representation

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$

Note that $\|x\|_2 = \sqrt{x^\top x}$ is the ell 2 norm.

Given $x_1, \ldots, x_n \in \mathbb{R}^d$

we want the original data point and the new representation as close as possible

$$\{D^*, \alpha_1^*, \ldots, \alpha_n^*\} = \arg\min_{D \in \mathbb{R}^{d \times k}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \|x_i - D\alpha_i\|^2.$$

Dictionary learning

eg. PCA, SVD

Note that

Estimated.
$\hat{X} = DR - R$

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i - D\alpha_i\|^2 = \frac{1}{n}\|X - DR\|_F^2,$$

where $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$,

k dimensions
$R = [\alpha_1, \alpha_2, \ldots, \alpha_n] \in \mathbb{R}^{k\times n}$,

$$\|X\|_F = \sqrt{\mathrm{trace}(X^\top X)} = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n} X_{i,j}^2}$$ is the Frobenius norm of X.

for matrix ↑

l2-norm for vector

Dictionary learning

Note that

$$\arg \min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2,$$

where $\mathcal{D}$ and $\mathcal{R}$ are some specific domains for D and R.

Optimisation

Objective:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

The objective is convex with respect to either R or D but not to both.

iteratively update

for one, optimise the other

Fix R, solve for D

$$\min_{D \in \mathcal{D}} \|X - DR\|_F^2$$

Fix D, solve for R

$$\min_{R \in \mathcal{R}} \|X - DR\|_F^2$$

Engan, Kjersti, Sven Ole Aase, and J. Hakon Husoy. "Method of optimal directions for frame design." Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. Vol. 5. IEEE, 1999.

Optimisation

Objective:

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

The objective is convex with respect to either D or R but not to both.

Suppose D^* and R^* are the local minimisers for the objective, we have

$$X \approxeq D^* R^* = (D^* A)(A^{-1} R^*).$$

D' — new dictionary

$AA^{-1} = I$

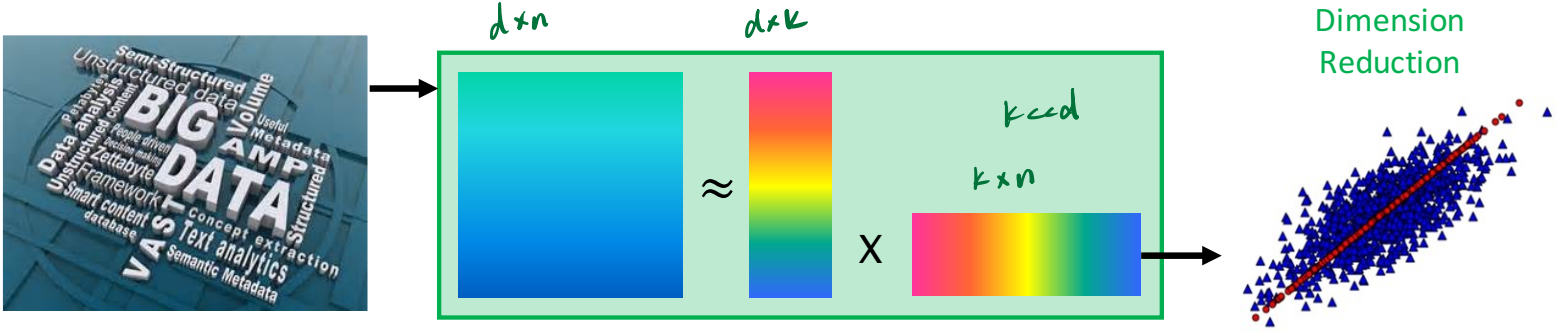
Normalisation (optional):

$$D_{:,i} \leftarrow D_{:,i} / \|D_{:,i}\|$$

if we make sure each column in dictionary is normalised, then A can only be I

norm for each column = 1

Dictionary learning application

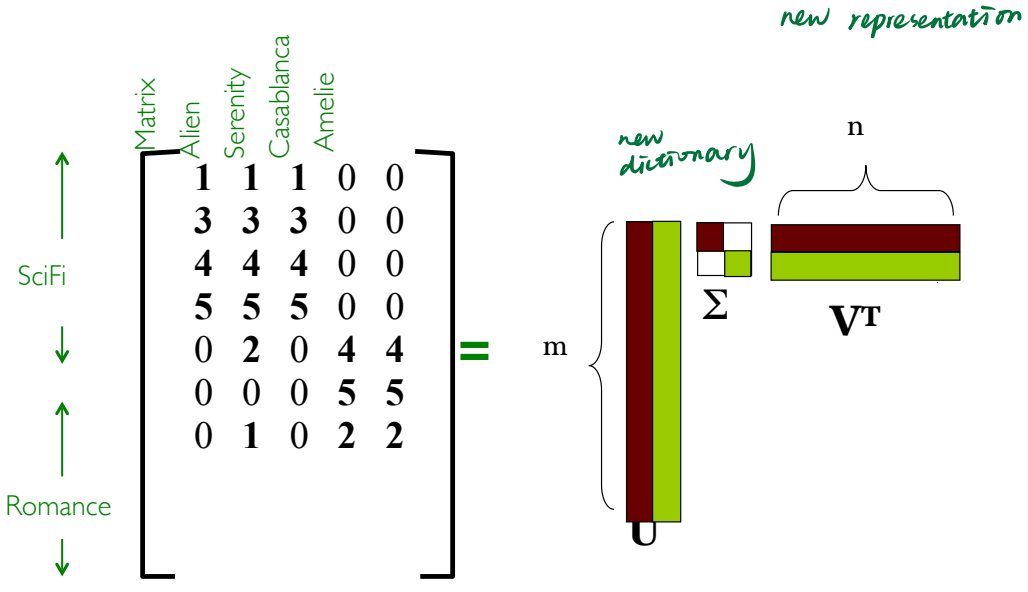


Dictionary learning

Problem 2

special requirement ?

SVD: $X = U\Sigma V^T$



special case of SVD

Dictionary learning

PCA: $A = U\Lambda U^T$

dictionary eigenvector

eigenvalue

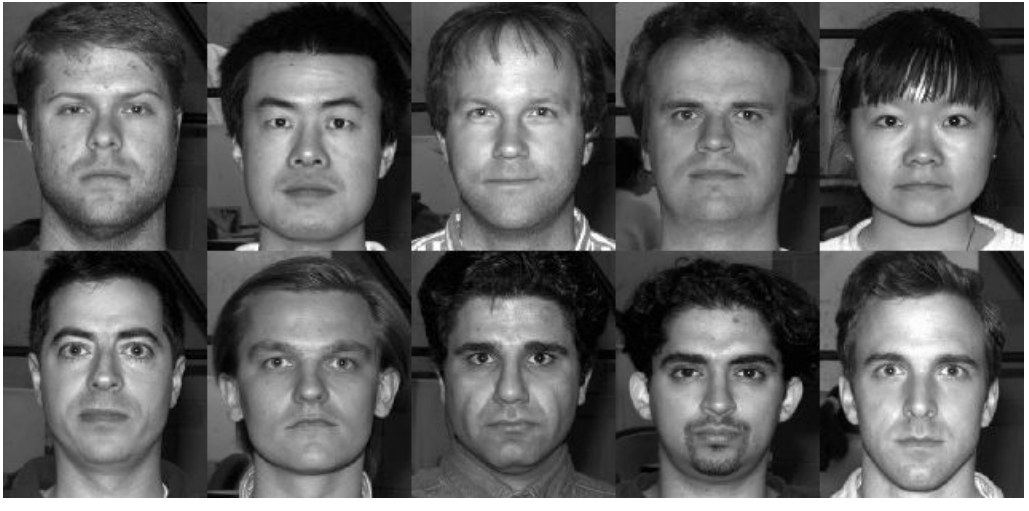
representation

$$A = \underbrace{\begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \\ | & | & & | \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} — & \mathbf{u}_1^T & — \\ — & \mathbf{u}_2^T & — \\ & \vdots & \\ — & \mathbf{u}_n^T & — \end{bmatrix}}_{R}$$

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Special requirement: columns of D are orthonormal to each other.

Dictionary learning application



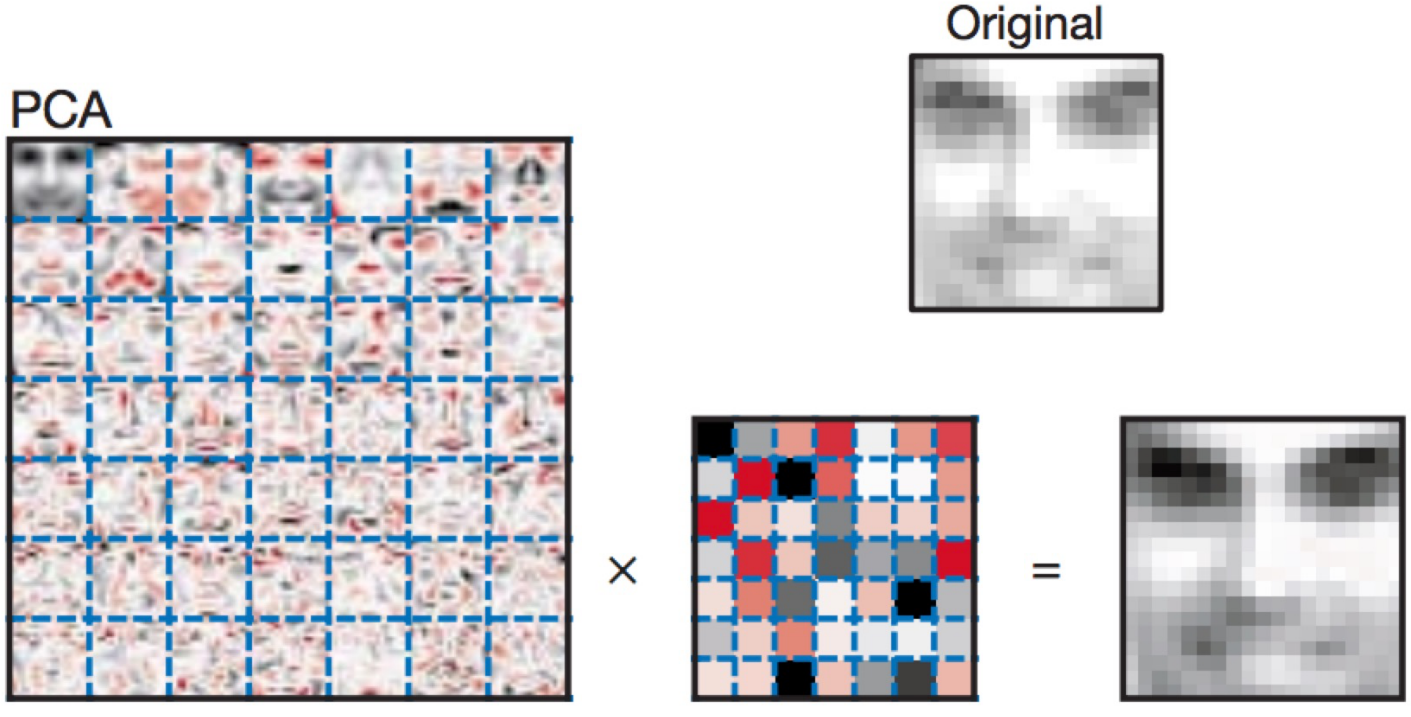
Question: what are the common bases D of human face?

The Yale Face Database B

Dictionary learning

PCA: $A = U\Lambda U^T$

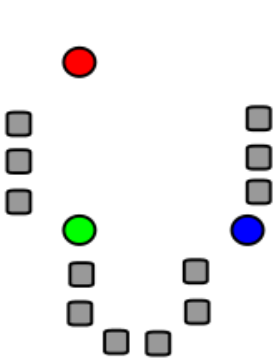
$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$

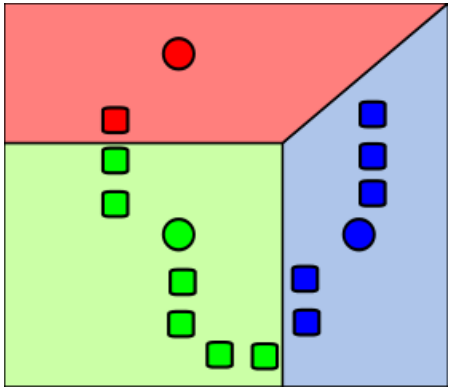


Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

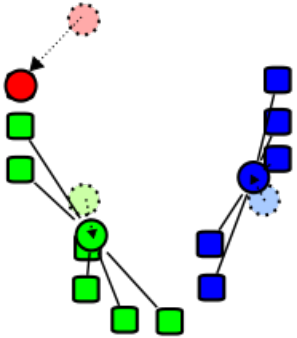
Dictionary learning

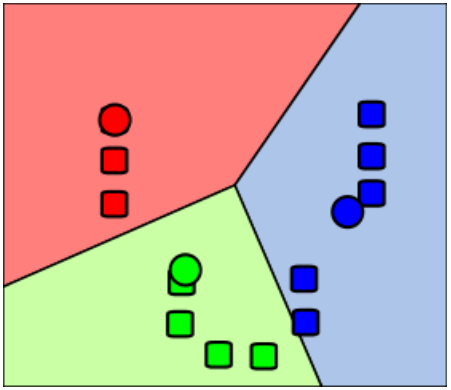
K-means clustering:





use k centroid to represent





An updating rule to find the k centroids.

Dictionary learning

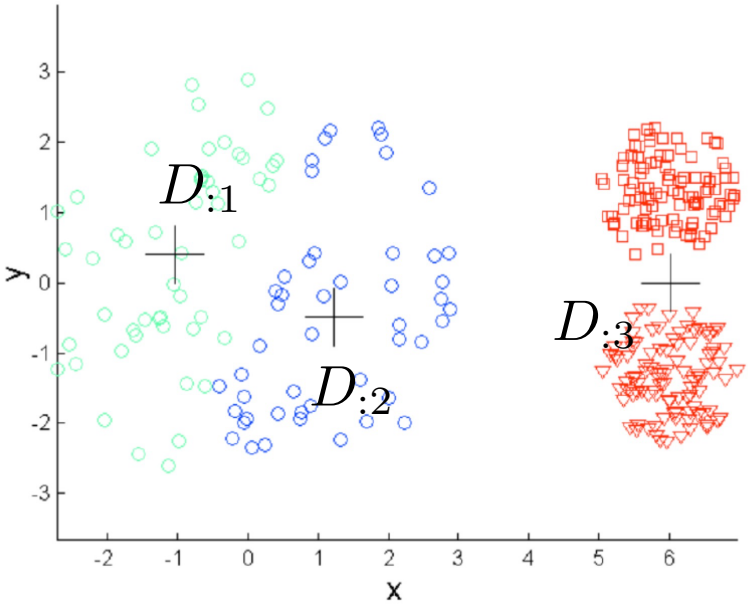
K-means clustering:

have reconstruction error

$$\min_{D\in\mathcal{D},R\in\mathcal{R}} \|X - DR\|_F^2$$

column one hot vector

larger k, small error

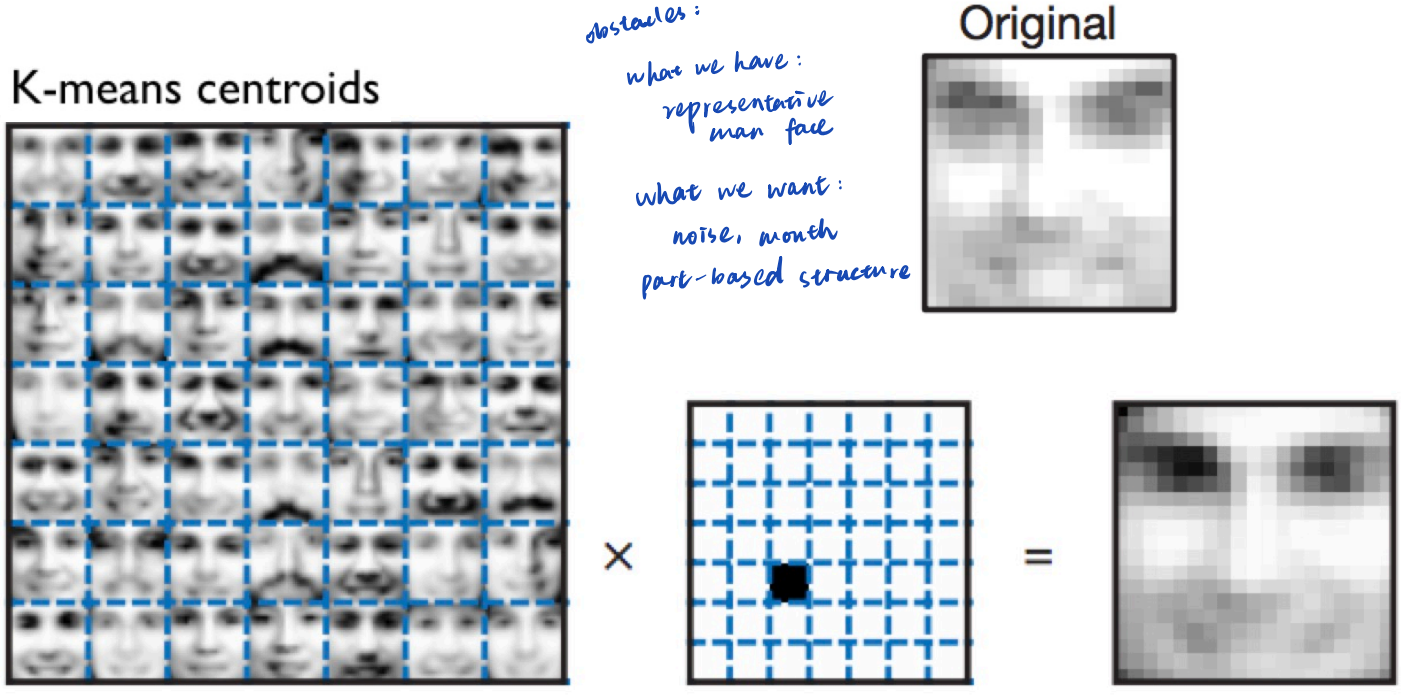


Special requirement: each column of R only one have entry equals to one, the other entries are all zeros.

Dictionary learning

K-means clustering:

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

Revisit human faces

Question: is that possible to extract parts-based structures of human faces? E.g., eye, nose, mouth, ear, forehead, chin…

The Yale Face Database B

THE UNIVERSITY OF SYDNEY

Non-negative matrix factorisation

Non-negative matrix factorisation

- **Why non-negativity of data?**

Data is often nonnegative by nature
Image intensities
Movie ratings
Document-term counts
Microarray data
Stock market values

Non-negative matrix factorisation

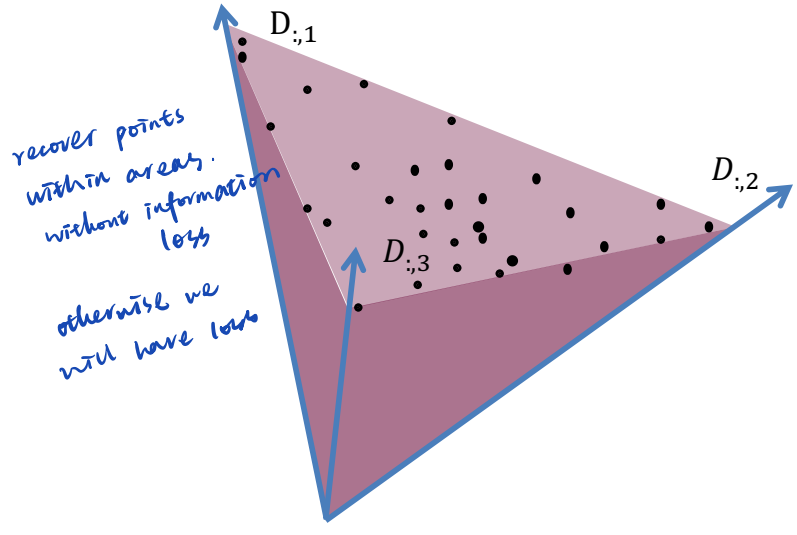
THE UNIVERSITY OF SYDNEY

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Special requirement: $\mathcal{D} = \mathbb{R}_+^{d \times k}, \quad \mathcal{R} = \mathbb{R}_+^{k \times n}$.

Geometry interpretation of non-negativity constraints

$\theta x + (1 - \theta)y$



$\theta \in [0, 1]$

$$\mathcal{D} = \mathbb{R}_+^{d \times k}, \quad \mathcal{R} = \mathbb{R}_+^{k \times n}.$$

$$\begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = a \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + b \times \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

Geometry interpretation of non-negativity constraints

Question: what will the bases (columns of D, e.g., $D_{:1}$) be look like?

Geometry interpretation of non-negativity constraints

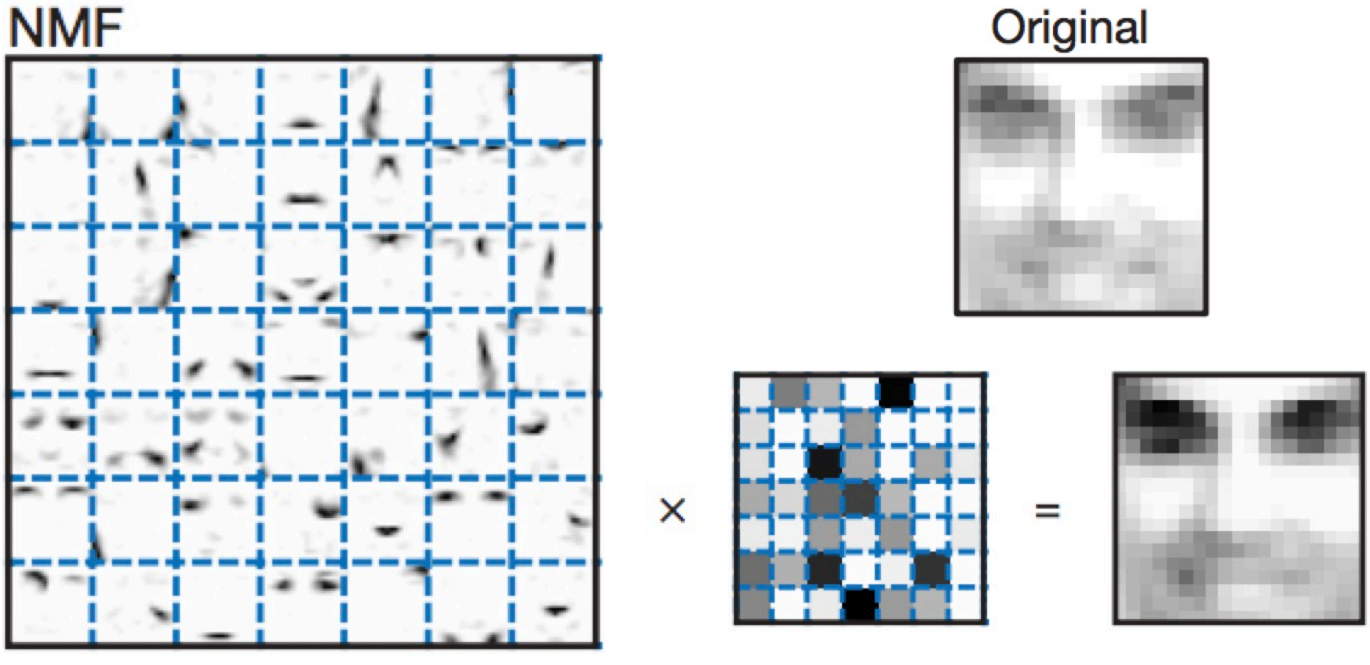
THE UNIVERSITY OF SYDNEY

These constraints lead to a **parts-based representation** because they allow **only additive**, not subtractive, **combinations**.

Non-negative matrix factorisation

parts representation

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \|x - D\alpha\|^2.$$



Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788.

NMF optimisation

MUR (Multiplicative Update Rules):

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Problem 3. How to derive this derivation.

Fix D, solve for R

$$\frac{\partial \|X - DR\|_F^2}{\partial R} = -2D^\top X + 2D^\top DR$$

The Matrix Cookbook: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Online helping tool: http://www.matrixcalculus.org/

$$\|X - DR\|_F^2 = \text{trace}((X - DR)^\top (X - DR))$$

NMF optimisation

MUR (Multiplicative Update Rules):

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Fix D^k, solve for R^{k+1}

$$\frac{\partial \|X - DR\|_F^2}{\partial R} = -2D^\top X + 2D^\top DR$$

$$R_{i,j}^{k+1} = R_{i,j}^k + \frac{\eta_{i,j}}{2}(2D^{k\top} X - 2D^{k\top} D^k R^k)_{i,j}$$

$$\eta_{i,j} = \frac{R_{i,j}^k}{(D^{k\top} D^k R^k)_{i,j}}$$

↑ set learning rate

NMF optimisation

MUR (Multiplicative Update Rules):

$$\min_{D \in \mathcal{D}, R \in \mathcal{R}} \|X - DR\|_F^2$$

Fix D^k, solve for R^{k+1}

$$R_{i,j}^{k+1} = R_{i,j}^k \frac{({D^k}^\top X)_{i,j}}{({D^k}^\top D^k R^k)_{i,j}}$$

Fix R^{k+1} , solve for D^{k+1}

$$D_{i,j}^{k+1} = D_{i,j}^k \frac{(X {R^{k+1}}^\top)_{i,j}}{(D^k R^{k+1} {R^{k+1}}^\top)_{i,j}}$$

NMF optimisation

MUR (Multiplicative Update Rules):

$$R_{i,j}^{k+1} = R_{i,j}^{k} \frac{(D^{k\top} X)_{i,j}}{(D^{k\top} D^{k} R^{k})_{i,j}} \qquad D_{i,j}^{k+1} = D_{i,j}^{k} \frac{(X R^{k+1\top})_{i,j}}{(D^{k} R^{k+1} R^{k+1\top})_{i,j}}$$

update R, D simultaneously.

Theorem non-convex, may have local optimal D / R.

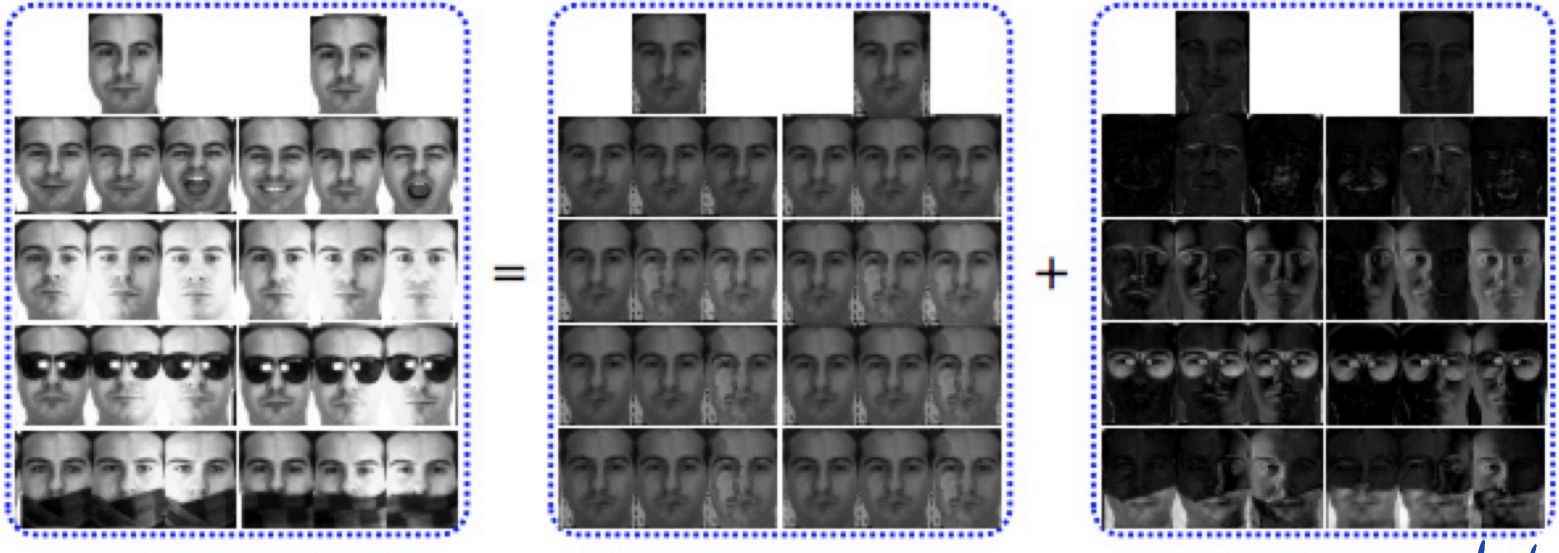
The above update rule will make the objective $\|X - DR\|_F^2$ non-increasing. The objective is invariant if and only if D^{k+1} and R^{k+1} are at a stationary point of the distance.

local optimal

Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances in neural information processing systems. 2001.

NMF applications

Face processing



Guan, Naiyang, et al. "Truncated Cauchy Non-negative Matrix Factorization for Robust Subspace Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

Pros and cons

- **Pros**:

Great interpretability

- **Cons**:

Factorisation is not unique

Feature redundancy among the bases

since. non-convex.

non-negative constraint : error larger, more information loss

NMF variants

ℓ_2-norm, NMF is not robust.

- Model developments with **various loss functions** in addition to squared loss
- to compensate for feature redundancy in the columns ⇒ eg. require columns to be sparse to remove redundancy
- to improve generalisation ability

If we can derive the upper bound,
→ include the upper bound in the object function
reduce reconstruction error