1 Unsupervised Learning

2020年5月24日 星期日　　下午4:06

- Clustering

一、K-means:

1. clustering $X=\{x^1,\cdots x^n,\cdots x^N\}$ into K cluster
2. initialize cluster center c^i, $i=1,2,\cdots k$ (k random x^n from X)

Repeat: for all x^n in X: $b_i^n \begin{cases}1 & x^n \text{ is most "close" to } c^i\\ 0 & \text{otherwise}\end{cases}$

updating all c^i: $c^i=\sum_{x^n} b_i^n x^n \Big/ \sum_{x^n} b_i^n$

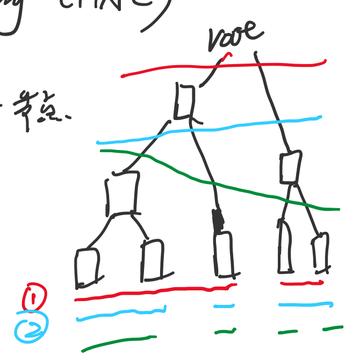
二、Hierarchical Agglomerative Clustering (HAC)

1. build a Tree

算两两相似度，相似的结为子节点。

2. pick a threshold

不同颜色切割。



— Dimension Reduction 降维

- PCA: principle component analysis $z=Wx$

$z=Wx$

Reduce to 1D: $z_1=w^1\cdot x$ (z_1 variance as large as possible)

$Var(z_1)=\sum_{z_1}(z_1-\bar{z}_1)^2$ $\|w^1\|_2=1$ $z_2=w^2\cdot x$

$Var(z_2)=\sum_{z_2}(z_2-\bar{z}_2)^2$ $\|w^2\|_2=1$ $w^1\cdot w^2=0$ (正交)

$W=\begin{bmatrix}(w^1)^T\\(w^2)^T\\\vdots\end{bmatrix}$ 求解: $z_1=w^1\cdot x$

$\bar{z}_1=\sum z_1=\sum w^1\cdot x=w^1\cdot\sum x=w^1\cdot\bar{x}$

$Var(z_1)=\sum_{z_1}(z_1-\bar{z}_1)^2=\sum\left(w^1(x-\bar{x})\right)^2$

$=(w^1)^T\sum(x-\bar{x})(x-\bar{x})^T w^1=(w^1)^T\cdot cov(x)\cdot w^1$

$S=Cov(x)$ $(w^1)^T\cdot S\,w^1$ (Maximizing, $\|w^1\|_2=1$)

$Sw^1=\alpha\cdot w^1$ (特征向量)

$(a\cdot b)^2=(a^Tb)^2=a^Tba^Tb$

— PCA another view

$x\approx c_1w^1+c_2w^2+\cdots+c_kw^k+\bar{x}$ $\begin{bmatrix}c_1\\c_2\\\vdots\\c_k\end{bmatrix}$ represent a digit image

component

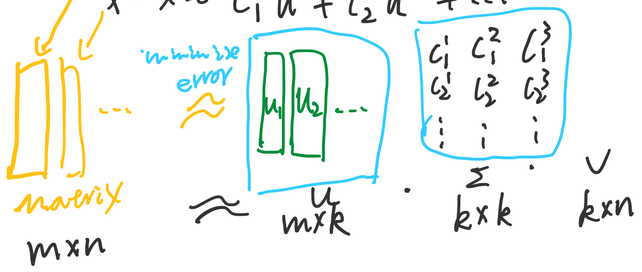
$x-\bar{x}\approx c_1w^1+c_2w^2+\cdots+c_kw^k=\hat{x}$

error: $L=\min_{\{w^1,\cdots w^k\}}\sum\left\|(x-\bar{x})-\left(\sum_{k=1}^{k}c_kw^k\right)\right\|_2$

PCA: $z=Wx$ $\begin{bmatrix}z_1\\z_2\\\vdots\\z_k\end{bmatrix}=\begin{bmatrix}w_1^T\\w_2^T\\\vdots\\w_k^T\end{bmatrix}x$

$x^1-\bar{x}\approx c_1^1w^1+c_2^1w^2+\cdots$

$x^2-\bar{x}\approx c_1^2w^1+c_2^2w^2+\cdots$



- Unsupervised
- Linear

Non-linear

PCA

LDA

Word Embedding: 词嵌入

- Generate Word Vector is unsupervised
- A word can be understood by its context

1. Count Based

• w_i, w_j 经常同时出现，$V(w_j)$ 和 $V(w_i)$ 接近

— Glove Vector

2. Prediction Based

① CBOW - continous bag of word

w_{i-1}, w_{i+1} → NN → w_i

② skip-gram: w_i → NN → w_{i-1}, w_{i+1}

3, Document Embedding