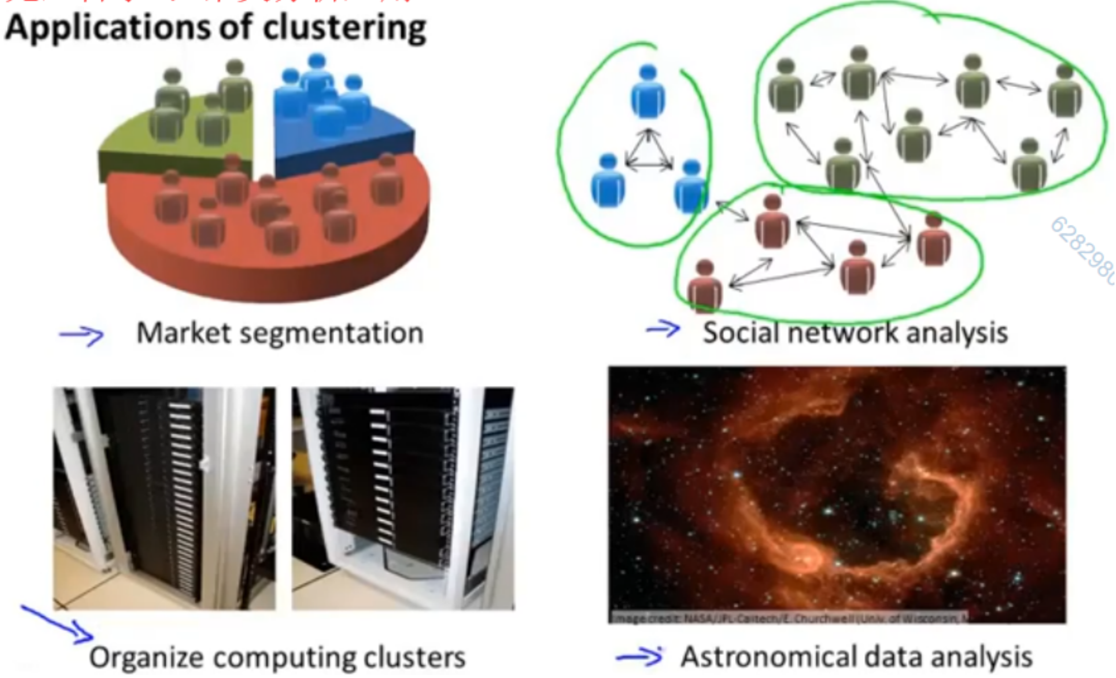
11.吴恩达-机器学习+无监督学习

笔记本： 日常
创建时间： 2019/11/18 9:13　**更新时间：** 2019/11/18 10:28
作者： 296645429@qq.com

无监督学习-聚类分析应用



Applications of clustering

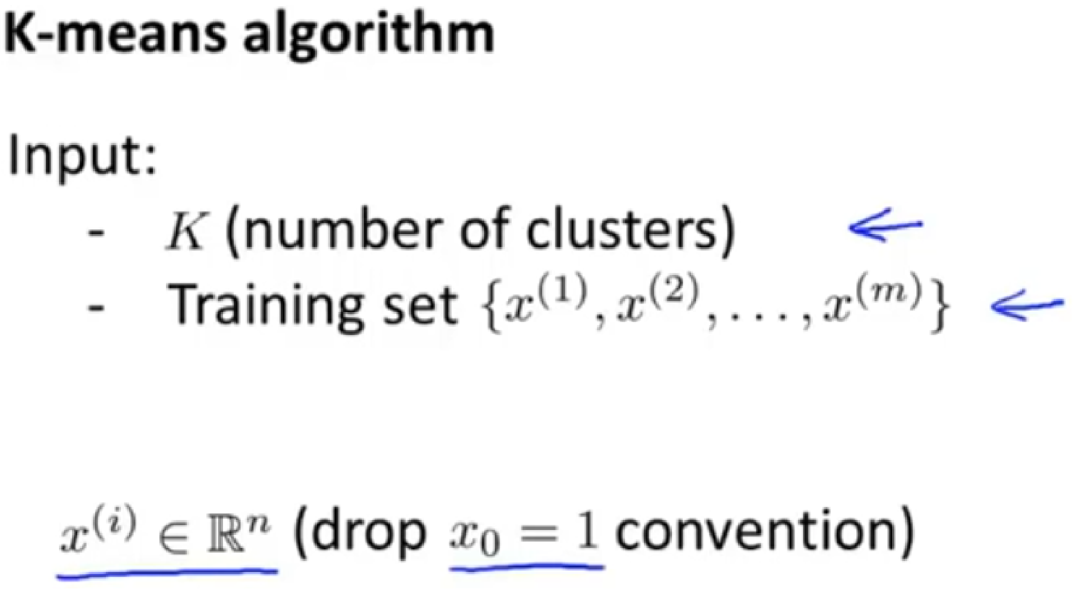
Market segmentation

Social network analysis

Organize computing clusters

Astronomical data analysis

Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, M



K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$\underline{x^{(i)} \in \mathbb{R}^n}$ (drop $\underline{x_0 = 1}$ convention)

K-means algorithm

μ_1 × μ_2 ×

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$
Repeat {
    for $i$ = 1 to $m$
        $c^{(i)}$ := index (from 1 to $K$) of cluster centroid
                closest to $x^{(i)}$
    for $k$ = 1 to $K$
        $\mu_k$ := average (mean) of points assigned to cluster $k$
}
```

聚类划分步骤 Cluster assignment step

$\min_k \|x^{(i)} - \mu_k\|^2 \rightarrow c^{(i)}$

移动聚类中心步骤 Move centroid

$x^{(1)}, x^{(5)}, x^{(6)}, x^{(10)} \rightarrow c^{(1)}=2, c^{(5)}=2, c^{(6)}=2, c^{(10)}=2$

$\mu_2 = \frac{1}{4}\left[x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}\right] \in \mathbb{R}^n$

（失真）代价函数

K-means optimization objective

→ $c^{(i)}$ = index of cluster (1,2,…,K) to which example $x^{(i)}$ is currently assigned

→ μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$) K $k \in \{1,2,\ldots,K\}$

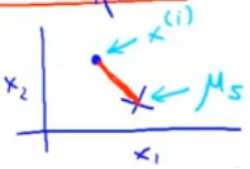
$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned $x^{(i)} \rightarrow 5$ $c^{(i)} = 5$ $\mu_{c^{(i)}} = \mu_5$

Optimization objective:

$$\rightarrow J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m}\sum_{i=1}^{m} \boxed{\|x^{(i)} - \mu_{c^{(i)}}\|^2} \leftarrow$$

$$\min_{\substack{c^{(1)}, \ldots, c^{(m)}, \\ \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

Distortion



第一步即簇分配步骤即是最小化损失函数的步骤

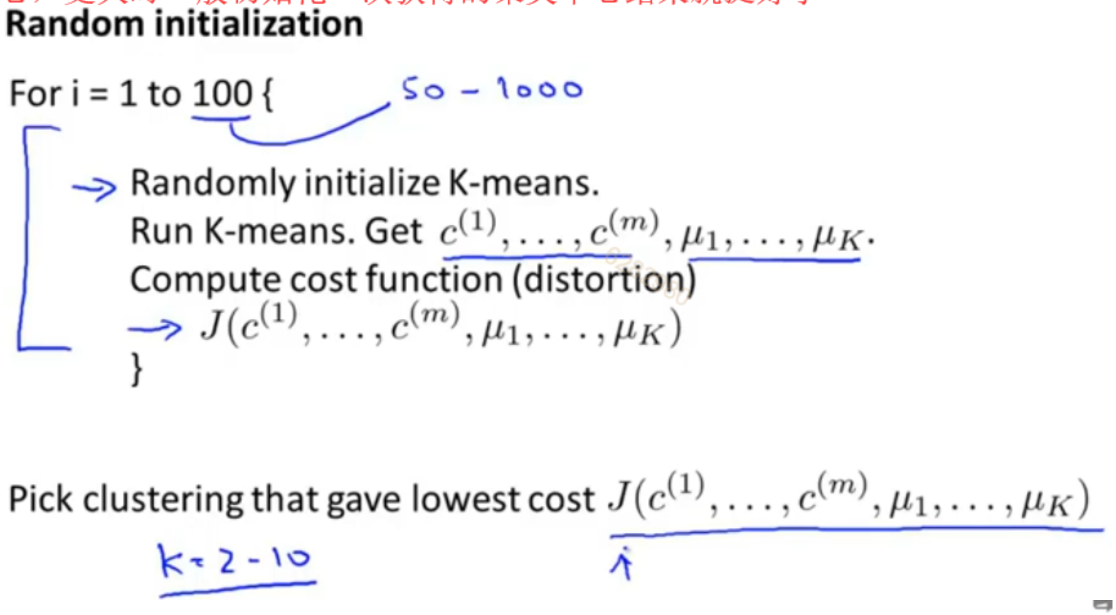
K-means algorithm

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$
Repeat {
    for $i$ = 1 to $m$
        $c^{(i)}$ := index (from 1 to $K$) of cluster centroid
                closest to $x^{(i)}$
    for $k$ = 1 to $K$
        $\mu_k$ := average (mean) of points assigned to cluster $k$
}
```

Cluster assignment step
Minimize $J(\ldots)$ wrt $\boxed{c^{(1)}, c^{(2)}, \ldots, c^{(m)}}$ ←
(holding $\mu_1, \ldots, \mu_K$ fixed)

move centroid
minimize $J(\ldots)$ wrt $\boxed{\mu_1, \ldots, \mu_K}$

聚类中心随机初始化，一般K在2-10时循环初始化多次取最小代价值的聚类中心，更大时一般初始化一次获得的聚类中心结果就挺好了



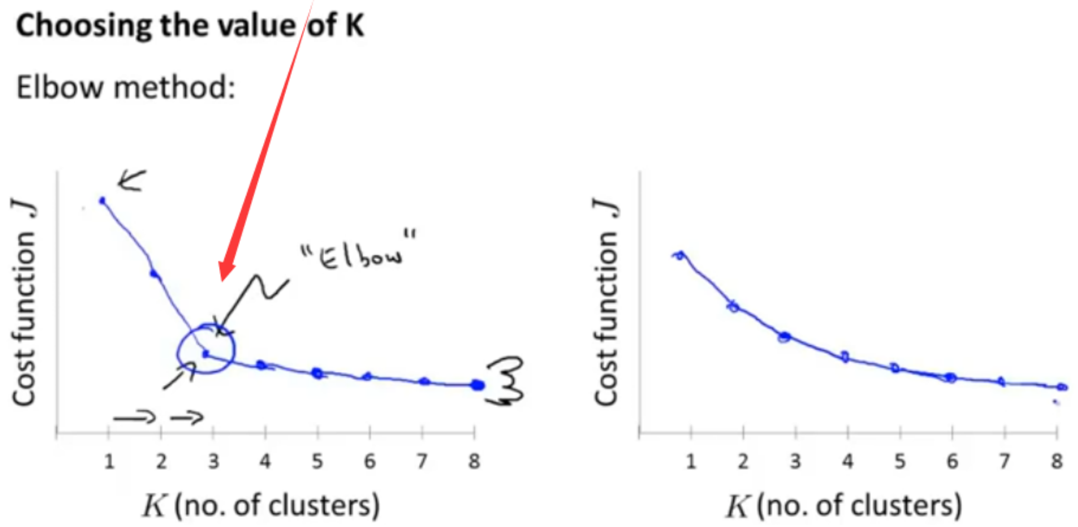
Random initialization

```
For i = 1 to 100 {          50 - 1000
    Randomly initialize K-means.
    Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.
    Compute cost function (distortion)
        $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$
}
```

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

K = 2 - 10

聚类数量选择方法：肘部方法



Choosing the value of K

Elbow method: