K-Means聚类算法

简介：是unsupervised Learning Algo. 将所有近似的example聚集到一个cluster中。

算法步骤：(1) 随机初始化K个centroids

往往是在原dataset中随机选K个example，而不是人为选定。

(2) 对每个 $x^{(i)}$，将其匹配到离其最近的那么centroid，使用idx(i)来保存centroid的编号

(3) 更新centroid的值。即用该centroid内所有 $x^{(i)}$ 的平均值（mean）来更新该centroid。

代价函数：$J(c^{(1)}, c^{(2)} \dots c^{(m)}, \mu^{(1)}, \mu^{(2)} \dots \mu^{(K)}) = \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - \mu_c^{(i)} \|$

就是保证每个 $x^{(i)}$ 与其对应的centroid的距离总和最小

如何选择K：一般使用 elbow method：逐渐增大K，在 J(θ) 变化不大的那个K即取为最终的K。

K-Means作图像compression：原本图像每个pixel 24bit.
(RGB表示的，每个8bit，表0-255之间integer)。回应一个picture
有hundreds of thousands colors. 现在，将~~所有~~piture用
K个color表示。即将所有3D中的颜色点进行聚类。最
后用centroid的color值替换每个 $x^{(i)}$ 的color值。只不过现
在每个pixel只需要存储centroid的索引即可。