模式识别 hw11　　学号　　2016310503

Problem 1

1.1 ~~若某step没有converge 即 E-step中 $r_{nk}^{(t)} = r_{nk}^{(t-1)}$~~

对应一组 r_{nk} $(n=1,\cdots N;\ k=1,\cdots K)$ $\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$ $(k=1,\cdots K)$

为右侧定的

因于一共只有 K^N 种 $r_{nk}(n=1,\cdots N;\ k=1,\cdots K)$ 的配置.

[illegible] 最多这么多次 iteration. 一定收敛.

因为每次更新 E-step 后 $J_{after\,E}^{(t)} \leq J_{after\,M}^{(t-1)}$

同时 $J_{after\,M}^{(t)} \leq J_{after\,E}^{(t)}$ ∴不会 re-enter 一个 r_{nk} 配置两次

1.2 $\gamma(z_{nk}) = p(z_n=k|x_n)$

$p(z_n=k, x_n) = \pi_k \frac{1}{(2\pi\varepsilon)^{\frac{D}{2}}} \exp\{-\frac{1}{2\varepsilon}\|x_n-\mu_k\|^2\}$

$$p(z_n=k|x_n) = \frac{p(z_n=k, x_n)}{\sum_k p(z_n=k, x_n)} = \frac{\pi_k \exp(-\frac{1}{2\varepsilon}\|x_n-\mu_k\|^2)}{\sum_k \pi_k \exp(-\frac{1}{2\varepsilon}\|x_n-\mu_k\|^2)}$$

$$\lim_{\varepsilon\to 0} p(z_n=k|x_n) = \lim_{\varepsilon\to 0} \frac{\pi_k \exp(-\frac{1}{2\varepsilon}(\|x_n-\mu_k\|^2 - \|x_n-\mu_{k'}\|^2))}{1+\sum_{k\neq k'} \pi_k \exp(-\frac{1}{2\varepsilon}(\|x_n-\mu_k\|^2 - \|x_n-\mu_{k'}\|^2))}$$

其中 $k' = \arg\min_k \|x_n-\mu_k\|^2$

当 $\varepsilon\to 0$ 时 $\|x_n-\mu_k\|^2 - \|x_n-\mu_{k'}\|^2 \geq 0$

只在 $k=k'$ 时取 "="

此时 $\lim_{\varepsilon\to 0} p(z_n=k'|x_n) = 1$

$$\lim_{\substack{\varepsilon\to 0 \\ k\neq k'}} p(z_n=k|x_n) = \frac{\pi_k \cdot 0}{1+0} = 0$$

∴ $\gamma(z_{nk}) \xrightarrow{\varepsilon\to 0} r_{nk}$

Problem1.3 实在没有证明出来… 感觉 epsilon 消不掉啊…

Problem 2

2.1. 复杂度分析

k-means:

假设迭代 ITER 次，一共数据量为 N. 分 K 个 cluster，每个数据维度为 D

每次迭代的计算新 cluster 中心步骤需要 O(ND)，新的 assignments 步骤需要计算每个数据点到每个中心的距离并找到最小值 O(NKD)。一共需要$O(NKD \times ITER)$的时间。

Spectral clustering:

计算两两数据点之间的相似度，然后构建 W，从而构建 L 需要$O(N^2)$，求 L 的最小的 M 个特征值(对应 M 个近似连通分量)对应的特征向量需要$O(N^3)$. 当然在 L 对称的情况下(全连接图或者对称 k 近邻图时)，做特征值分解有数值加速算法，实际计算并不需要$O(N^3)$时间。然后再调用一次 k-means 算法需要$O(NKM \times ITER)$时间.

2.2. 实验结果

k-means 算法中用不同的 seed 运行，初始化的 random partition 不同，结果会不同。

解决方法: k-means，以及其他的 EM 类算法都可能 stuck at local minimum，可以通过多次的 restart 试验，根据每个 sample 到中心的平均距离或者(在概率模型情况下)训练集的 likelihood 选择最好的一次运行。

seed	NMI	Average square dist
12345	0.4639	3115
1234	0.4525	3108
123	0.5209	3091
12	0.4610	3115

以下实验基本上都在 ITER=30 次左右收敛, N=100，固定 random seed 为 12345，即 initialization 相同是算法结果为:

方法	M	近邻图	Normalize	NMI
k-means	-	-	-	0.4639
Spectral clustering	10	全连接	否	0.1057
	5	k 近邻 k=5	否	0.5817
	8		否	0.6173
	10		否	0.5783
	20		否	0.4615
	10	k 近邻 k=5	是	0.5396
	10	k 近邻 k=10	否	0.5766
	10	k 近邻 k=20	否	0.5369

在我的实验中，谱聚类的结果大多数情况下比 k-means 好，如果用全连接效果十分差，原因可能是在全连接的情况下要找出 M 个近似联通分量的确十分不精准。

可能是由于数值精度的原因, normalize 之后的谱聚类结果没有 normalize 之前好。在我的谱聚类实现中两两节点之间的相似度计算如下:

```
W = exp(-pdist(X)/280);
W = squareform(W);
```

其中 280 是一个我考虑 784 维向量平均每维度的像素值差距为 10 时的欧式距离, 设置为一个超参数. 并没有时间调整实验了。

2.3. 怎么选择 cluster 个数

可以根据类间方差的平均值和类内方差做对比，如果一个类的类内方差太大，应该需要继续拆分。

可以使用非参贝叶斯的方法，用一个 Dirichlet process 作为 sample partition 情况的先验，liklihood function 可以用 gaussian likelihood。实现可以用其中国餐馆过程形式做 gibbs sampling, 为了跟 k-means 更类似，每个桌子上的方差可以设置的大一点... 在一定 sample 次数后更新每个桌子对应的 cluster 中心。没有时间写诶...

2.4. 比较

spectral clustering... 相对于 k-means, spectral clustering 虽然需要计算一次特征值分解，但是由于数值加速算法的存在，这个计算通常也比较快。由于对原始 feature 维度从 D 降维到了 M 维，在 k-means 过程中也更加快，同时对 feature 中的噪声也更加不敏感。