CS182 HW4 王鹏豪 2021533138

1. As for the 2 type of PCA

① PCA based on projected variance maximization

let $v_1, v_2, \dots, v_d$ denote the d principal components,

$v_i \cdot v_j = 0,\ i \neq j$ and $v_i \cdot v_i = 1,\ i = j$.

~~Anu~~ Assume that data is centered.

Let $X = [X_1, X_2, \dots, X_n]$, we have $\frac{1}{n}\sum_{i=1}^{n}(v^T x_i)^2 = v^{+}XX^Tv$

we need to find vector that maximizes sample variance of projected data

which is $\max_{v} (v^T X X^T v)$ s.t. $v^T v = 1$

$$= Tr(VV^T X^T X VV^T) = Tr(X^{+} X VV^T)$$

② PCA based on projected error minimization.

$\frac{1}{n}\sum_{i=1}^{n} \| x_i - (v^T x_i) v \|^2$. we have that

$$\min_{v} \| X - XVV^T \|^2 = Tr((X - XVV^T)^T (X - XVV^T))$$
$$= Tr((I - VV^T)^T X^T X (I - VV^T))$$
$$= Tr(X^T X (I - vv^T))$$
$$= Tr(X^T X - X^T X VV^T)$$
$$= Tr(X^T X) - Tr(X^T X VV^T)$$

as the $Tr(X^{+}X)$ is a constant, the

$\max Tr(X^{+} X VV^T)$ is same as $\min -Tr(X^T X vv^T)$

so we get that PCA based on projected variance maximization is equivalent to PCA based on projected error minization.

2.

(a) Cost function:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} ||X_i - \mu_j||_2^2 \; y_i^j$$, where $y_i^j = 1$ when $y_i = j$, and 0 for otherwise

(b) as defined in (a) , we need to minimize the cost function , after each calculation, the mean will be updated, we denote the new mean is μ'

$$f = \sum_{i=1}^{n}\sum_{j=1}^{k} ||X_i - \mu_j'||_2^2 \; y_i^j$$, where $y_i^j = 1$, when $y_i = j$, 0 for others

$$\frac{\partial f}{\partial \mu_j'} = -2 \sum_{i \in C_j} (X_i - \mu_j') = 0$$, then we get that

$|C_j| \mu_j' = \sum_{i \in C_j} X_i$, where $|C_j|$ is the number of points in the corresponding jth cluster, so we get that $\mu_j' = \frac{\sum_{i \in C_j} X_i}{|C_j|}$, as the definition of centroid , we find that the each μ_j need to be the mean of the sample points assigned to the cluster i.

(c) The It do not follow the advice , there will be mainly 2 problems:

① When we met the case that a sample point has serveral clusters means that are equally close, when we move to the next iteration, it not stay at the last cluster, we may fall into a case that we find a lacal minimum solution, which result the wrong solution.

② If we do not stay at the last cluster, the sample point may continously move among the serval cluster , which will need cost more iterations, so will make the time longer or even not converge.

(d) CD