باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

مقدمه ای بر یادگیری ماشین - گروه دکتر امینی
پاییز ۱۴۰۱

تمرین تئوری چهارم

۱. مهلت تحویل این تمرین مطابق تاریخ اعلام شده در سامانه CW می باشد.

۲. ۱۴ روز تاخیر مجاز برای تحویل تمارین در اختیار شما خواهد بود.

۳. سقف تاخیر برای تحویل هر تمرین ۷ روز خواهد بود و پس از آن پاسخنامه تمرین منتشر خواهد شد.

۴. ابهامات و مشکلات خود از دو سوال اول را می توانید با آقای توانا و باقی سوالات را با آقای بقایی مطرح کنید.

@mtv_tavana @erfunba

1 Second principal component

فرض کنید:

$$J(v_2, z_2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - z_{i1}v_1 - z_{i2}v_2)^T(x_i - z_{i1} - z_{i2}v_2)8$$

نشان دهید:

$$if : \frac{\partial j}{\partial z_2} = 0 \to z_{i2} = v_2^T x_i$$

2 PCA residual error

نشان دهید:

$$||x_i - \sum_{j=1}^{K} z_{ij}v_j||^2 = x_i^T x_i - \sum_{j=1}^{K} v_j^T x_i x_i^T v_j$$

3

Recall that we discussed three clustering: Scale Invariance, Richness, and Consistency. Consider the Single Linkage clustering algorithm. 1. Find which of the three properties is satisfied by Single Linkage with the Fixed Number of Clusters (any fixed nonzero number) stopping rule. 2. Find which of the three properties is satisfied by Single Linkage with the Distance Upper Bound (any fixed nonzero upper bound) stopping rule. 3. Show that for any pair of these properties there exists a stopping criterion for Single Linkage clustering, under which these two axioms are satisfied.

4

Given a metric space (χ, d), where $|\chi| < \infty$, and $k \in \mathbb{N}$, we would like to find a partition of χ into $C_1, ..., C_k$ which minimizes the expression

$$G_{k-diam}((\chi, d), (C_1, ..., C_k)) = \underset{j\in[d]}{max}\, diam(C_j),$$

where diam $(C_j) = max_{x,x' \in C_j} d(x, x')$ (we use the convention $diam(C_j) = 0$ if $|C_j| < 2$).
Similarly to the k-means objective, it is NP-hard to minimize the k-diam objective. Fortunately, we have a very simple approximation algorithm: Initially, we pick some $x \in \chi$ and set $\mu_1 = x$.
Then, the algorithm iteratively sets

$$\forall j \in \{2, ..., k\}\, ,\ \mu_j = \underset{x\in\chi}{argmax} \underset{i\in[j-1]}{min} d(x, \mu_i).$$

Finally, we set

$$\forall i \in [k],\ C_i = \left\{ x \in X : i = \underset{j\in[k]}{argmin}\ d(x, \mu_j) \right\}.$$

Prove that the algorithm described is a 2-approximation algorithm. That is, if we denote its output by $\hat{C}_1, ..., \hat{C}_k$, and denote the optimal solution by $C_1^*, ..., C_k^*$, then,

$$G_{k-diam}((\chi, d), (\hat{C}_1, ..., \hat{C}_k)) \leq 2 \cdot G_{k-diam}((\chi, d), (C_1^*, ..., C_k^*)).$$

Hint: Consider the point μ_{k+1} (in other words, the next center we would have chosen, if we wanted k+1 clusters). Let $r = \underset{j\in[k]}{min}\, d(\mu_j, \mu_{k+1})$ Prove the following inequalities

$$G_{k-diam}((\chi, d), (\hat{C}_1, ..., \hat{C}_k)) \leq 2r$$

$$G_{k-diam}((\chi, d), (C_1^*, ..., C_k^*)) \geq r.$$

5

Recall that a clustering function, F, is called Center-Based Clustering if, for some monotonic function $f : \mathbb{R}_+ \to \mathbb{R}_+$, on every given input (χ, d), $F(\chi, d)$ is a clustering that minimizes the objective

$$G_f((\chi, d), (C_1, ..., C_k)) = \underset{\mu_1,...,\mu_k\in\chi'}{min} \sum_{i=1}^{k} \sum_{x\in C_i} f(d(x, \mu_i)),$$

where χ' is either χ or some superset of χ. Prove that for every k > 1 the k-diam clustering function defined in the previous exercise is not a center-based clustering function.
(Hint: Given a clustering input (χ, d), with $|\chi| > 2$, consider the effect of adding many close-by points to some (but not all) of the members of χ, on either the k-diam clustering or any given center-based clustering.)