题目 1 *判断题 16 分*

(a) From the "Occam's razor" (that is derived from the PAC learning framework), we can deduce that given two rules that are consistent with the (empirical) data, the complicated rule is necessarily worse (than the simple one).

由基于 PAC 学习理论的 "Occam 剪刀" 原理，我们可推导出：对于任意两个与（经验）数据一致的规则，复杂的规则一定比简单的规则更坏。

(b) Using the SVM (support vector machine), we can classify the data according to the inner products between vectors corresponding to the data points and the labels of training data.

利用支持向量机的方法对数据进行分类时，其结果只取决于数据向量之间的内积及训练数据的标签 (labels)。

(c) Since there exists a (k, ε)-coreset of size $\text{poly}(k, \frac{1}{\varepsilon})$ for the Euclidean k-means problem for any $\varepsilon > 0$, we know that there exists an algorithm that runs in polynomial (in $n, k, \frac{1}{\varepsilon}$) time and a solution C such that $D(A, C^*) \leq D(A, C) \leq (1+\varepsilon)D(A, C^*)$, where C^* is the optimum solution, and $D(A, C')$ denotes the (k-means) cost of the input set A with respect to the center set C'.

对于欧式 k-means 问题，由于存在针对该问题的大小为 $\text{poly}(k, \frac{1}{\varepsilon})$ 的 (k, ε)-核心集 (这里的参数 $\varepsilon > 0$)，那么一定存在一个运行时间为 (关于 $n, k, \frac{1}{\varepsilon}$ 的) 多项式的算法，该算法可以找到一个解 C 满足 $D(A, C^*) \leq D(A, C) \leq (1+\varepsilon)D(A, C^*)$，这里的 C^* 为最优的解，且 $D(A, C')$ 表示输入集 A 相对于中心集 C' 的（k-means）代价。

(d) Let $A \subset \{0,1\}^d$ be a set of n data points, each being a d-dimensional 0, 1-vector. For every two points $a, a' \in \{0,1\}^d$, define its distance to be $D(a, a') = \sum_{i=1}^{d} |a_i - a'_i|$. Then based on this distance function, it is equivalent to use the k-means and k-median methods to perform clustering on A.

假设 $A \subset \{0,1\}^d$ 是一个包含 n 个点的数据集，其每个数据是一个 d 维的 0, 1 向量。对于任意两个点 $a, a' \in \{0,1\}^d$，定义其距离为 $D(a, a') = \sum_{i=1}^{d} |a_i - a'_i|$。那么使用基于该距离函数的 k-means 和 k-median 算法对 A 聚类，效果是等价的。

上述陈述中对的有： _____

错的有： _____

题目 2 单选题 4 分

Which of the following is correct about the VC-dimension of set systems?

(a) All set systems have finite VC-dimension.

(b) The VC-dimension of the set system defined by $X = \mathcal{R}^d$ and the set $\mathcal{H}$ of all the spheres in $\mathbb{R}^d$ is d.

(c) The VC-dimension of the set system defined by $X = \mathcal{R}$ and the set $\mathcal{H}$ of all the intervals in $\mathbb{R}$ is 1.

(d) The VC-dimension of the set system defined by $X = \mathcal{R}^d$ and the set $\mathcal{H}$ of all the halfspaces in $\mathbb{R}^d$ is $d + 1$.

关于集合系统的 VC-维，下面哪个说法是正确的：

(a) 所有集合系统的 VC-维都是有限的。

(b) 由 $X = \mathcal{R}^d$ 及 $\mathcal{R}^d$ 中的所有的球所定义的的集合系统的 VC-维是 d。

(c) 由 $X = \mathcal{R}$ 及 $\mathcal{R}$ 中所有的区间所定义的集合系统的 VC-维是 1。

(d) 由 $X = \mathcal{R}^d$ 及 $\mathcal{R}^d$ 中所有的半平面所定义的的集合系统的 VC-维是 $d + 1$。

你的选择： _____

题目 3 *10 分*

For any two points $a, b \in \mathbb{R}^2$, define their distance to be $D(a,b) = \|a-b\|_2$. Answer the following questions:

(a) For any two subsets $A, B \subseteq \mathbb{R}^2$ of finite numbers, what is the definition of the single linkage cost between A, B?

(b) Apply the single linkage based hierarchical clustering algorithm to the following 4 points in $\mathbb{R}^2$: $x_1 = (1,3), x_2 = (2,2), x_3 = (-1,1), x_4 = (-2,2)$. Draw the dendrogram according to the hierarchical clustering. It suffices to draw the corresponding tree and mark all the leaves with point labels.

对于 $\mathbb{R}^2$ 中的任意两个点 a, b，定义其距离为 $D(a,b) = \|a-b\|_2$。回答下面的问题：

(a) 对于 $\mathbb{R}^2$ 中的任意两个有限子集 A, B，它们之间的 single linkage 的代价是如何定义的?

(b) 使用基于 single-linkage 的层次聚类方法对 $\mathbb{R}^2$ 中的 4 个点 $x_1 = (1,3), x_2 = (2,2), x_3 = (-1,1), x_4 = (-2,2)$ 进行聚类。画出相应的层次聚类所定义的树状图，只需画出树状结构及叶子节点编号即可。

题目 4 *10 分*

Consider an instance space X consisting of integers 1 to T for some sufficiently large even number T, and a target concept c^* where $c^*(i) = 1$ for $\frac{T}{2}+1 \leq i \leq T$, $c^*(i) = 0$ otherwise. If your hypothesis class $\mathcal{H}$ is $\{h_j, 1 \leq j \leq T \mid h_j(i) = 1 \text{ for } i \geq j \text{ and } h_j(i) = 0 \text{ for } i < j\}$, how large a training set S do you need to ensure that with probability at least 99%, any consistent hypothesis in $\mathcal{H}$ will have true error less than 10%? Briefly explain how you derive your answer.

考虑实例空间 $X = \{1, 2, \ldots, T\}$，这里的 T 为足够大的偶数。考虑目标概念（concept）c^*，其满足当 $\frac{T}{2} + 1 \leq i \leq T$ 时，$c^*(i) = 1$；对于其余的 i，$c^*(i) = 0$。如果你的 hypothesis 类 $\mathcal{H}$ 为 $\{h_j, 1 \leq j \leq T \mid h_j(i) = 1$当 $i \geq j$；$h_j(i) = 0$ 当 $i < j\}$。我们需要选择多大的训练集合 S 才能保证以至少 99% 的概率，$\mathcal{H}$ 中的任意一个一致的 hypothesis 具有的真实误差不超过 10%? 简要解释你得出结论的依据。

题目 5 *20 分*

Let $(X, \mathcal{H})$ be a set system and D be a probability distribution over X. Let S_1, S_2 be two sets, each consisting of n points drawn independently from D. Let A be the event that there exists a set $h \in \mathcal{H}$ of probability mass (with respect to D) at least $\frac{\varepsilon}{2}$ that is disjoint from S_1. Let B be the event that there exists $h \in \mathcal{H}$ that is disjoint from S_1 but contains at least $\frac{\varepsilon n}{4}$ points in S_2. Prove that if $n \geq \frac{16}{\varepsilon}$, then $\Pr[B \mid A] \geq \frac{1}{2}$.

令 $(X, \mathcal{H})$ 为一个集合系统。令 D 是 X 上的一个概率分布。令 S_1, S_2 为 X 的两个子集，其中每个集合包含 n 个从 X 中按照分布 D（独立）抽样出的点。令 A 表示以下事件：$\mathcal{H}$ 中存在一个集合 h，其 (相对于分布 D 的) 概率质量 (probability mass) 不低于 $\frac{\varepsilon}{2}$ 且 h 与 S_1 不相交。令 B 表示以下事件：$\mathcal{H}$ 中存在一个集合 h，满足 h 与 S_1 不相交但是 h 包含了 S_2 中至少 $\frac{\varepsilon n}{4}$ 个点。证明当 $n \geq \frac{16}{\varepsilon}$ 时，$\Pr[B \mid A] \geq \frac{1}{2}$。

题目 6 *20 分*

Consider the Euclidean k-means problem. For an input set A and a center set C, define $D(A,C) = \sum_{a\in A}\min_{c\in C}\|a-c\|^2$. Let P,Q be two multisets of n points in $\mathbb{R}^d$, and let $\varepsilon \in (0,1)$, $k \geq 1$. Let $\Lambda \geq 0$. Let $\pi : P \to Q$ be a bijective mapping such that $\sum_{x\in P}\|x-\pi(x)\|^2 \leq \frac{\varepsilon^2}{16}\cdot\Lambda$. Prove that for any set C of k centers, it holds that

$$|D(P,C) - D(Q,C)| \leq \varepsilon \cdot \max\{\Lambda, D(P,C)\}.$$

Hint: You may consider to use the inequality $\sqrt{\Lambda \cdot D(P,C)} \leq \max\{\Lambda, D(P,C)\}$.

考虑欧氏空间的 k-means 问题。对于输入集 A 及中心集 C，定义 $D(A,C) = \sum_{a\in A}\min_{c\in C}\|a-c\|^2$。令 P,Q 是 $\mathbb{R}^d$ 中的两个包含 n 个点的集合。令 $\varepsilon \in (0,1)$，$k \geq 1$。令 $\Lambda \geq 0$。假设 $\pi : P \to Q$ 是一个双射其满足：$\sum_{x\in P}\|x-\pi(x)\|^2 \leq \frac{\varepsilon^2}{16}\cdot\Lambda$。证明对于任意一个具有 k 个中心的集合 C，下面的不等式成立：

$$|D(P,C) - D(Q,C)| \leq \varepsilon \cdot \max\{\Lambda, D(P,C)\}$$

提示：可以考虑使用不等式 $\sqrt{\Lambda \cdot D(P,C)} \leq \max\{\Lambda, D(P,C)\}$。

题目 7 *20 分*

Consider the Euclidean k-means problem as in the previous question. For an arbitrary set C', define $c(C') = \frac{1}{|C'|}\sum_{x\in C'} x$ to be its centroid.

For the input set A, let C^* be the set of k centers corresponding to an optimal solution. Let $\mathcal{P} = \{P_1, \ldots, P_k\}$ be the corresponding clustering of A, i.e., each cluster P_i is induced by some center in C^* (i.e., the centroid of P_i is a center in C^*). For any set C of centers, we say *the cluster P_i is bad with respect to C* if $D(P_i, C) > 10 \cdot D(P_i, c(P_i))$.

Consider the subroutine D^2-sampling(A, k) of the k-means++ algorithm. Let $i \geq 2$. In the beginning of the i-th iteration, we define $\mathcal{B}_i$ to be the set of all clusters in $\mathcal{P}$ that are bad with respect to C^{i-1} (i.e., the set of the first $i-1$ centers chosen by the subroutine). Prove that in the i-th iteration of D^2-sampling, if $D(A, C^{i-1}) > 20 \cdot D(A, C^*)$, then with probability at least $1/2$, the i-th center c_i sampled by the subroutine is from $\mathcal{B}_i$.

与前一题一样，考虑欧氏空间的 k-means 问题。对于任意集合 C'，定义 $c(C') = \frac{1}{|C'|}\sum_{x\in C'} x$ 为其 centroid。

对于输入集 A，令 C^* 是一个具有 k 个中心点的最优解。令 $\mathcal{P} = \{P_1, \ldots, P_k\}$ 为与之相应的一个聚类，即每个 cluster P_i 的 centroid 都对应于 C^* 中的某个中心点。对于任意一个中心集 $C \subseteq \mathbb{R}^d$，我们称P_i 相对于 C 来说是坏的，如果 $D(P_i, C) > 10 \cdot D(P_i, c(P_i))$。

考虑 k-means++ 算法中的子程序 D^2-sampling(A, k)。令 $i \geq 2$。在第 i 次迭代之前，我们定义 $\mathcal{B}_i$ 为 $\mathcal{P}$ 中所有的相对于 C^{i-1} 来说是坏的的 clusters 所构成的集合，这里 C^{i-1} 为子程序找到的前 $i-1$ 个中心构成的集合。证明在子程序 D^2-sampling 的第 i 次迭代中，如果 $D(A, C^{i-1}) > 20 \cdot D(A, C^*)$，那么以至少 $1/2$ 的概率，子程序抽样出来的第 i 个中心 c_i 来自于 $\mathcal{B}_i$。