Institut für Informatik
Prof. Dr. Heiko Röglin
Michael Etscheid

universitätbonn

Probabilistic Analysis
of Algorithms
Summer 2013

Problem Set 9

Problem 1

For $X \subseteq \mathbb{R}^d$, $|X| = n$ and $k \leq n$, let $C_1, \ldots, C_k$ be a partition of X. Furthermore, let $c_1, \ldots, c_k \in \mathbb{R}^d$. Define the potential $\phi = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2$. Show the following two claims.

a) If a center c_i is exchanged by the center of mass $c_{i'} = \frac{1}{|C_i|} \cdot \sum_{x \in C_i} x$, the potential ϕ drops by $|C_i| \cdot \|c_{i'} - c_i\|_2^2$.

b) If a point $x \in C_i$ switches to a cluster $C_{i'}$, $i' \neq i$, and the distance between x and the bisector of c_i and $c_{i'}$ is ε, the potential ϕ drops by $2\varepsilon \|c_{i'} - c_i\|_2$.

Problem 2

We say that a point set $X \subseteq \mathbb{R}^d$ is ε-separated if for any hyperplane $\mathcal{H}$, there are at most $2d$ points in X with distance ε of $\mathcal{H}$.

Suppose k-means is run on an ε-separated point set $X \subseteq \mathbb{R}^d$. Show that if one cluster gains or loses a total of at least $2kd$ points within a single iteration, then the potential drops by at least $4\varepsilon^2/n$.

Problem 3

Let $Y_1, \ldots, Y_d$ be independent normally distributed random variables with variance σ^2 and mean μ_i for each variable Y_i. Then

$$f_i(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right)$$

is the density function for each variable Y_i. The distribution of $Y = (Y_1, \ldots, Y_d)$ is called *d-dimensional normal distribution* with variance σ^2.

a) Let $f\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be the density function of Y. Show that for any $y_1, \ldots, y_d \in \mathbb{R}$, $f(y_1, \ldots, y_d) = f_1(y_1) \ldots f_d(y_d)$. Derive a formula for f.

b) Deduce that if $x \in \mathbb{R}^d$ is chosen according to a d-dimensional normal distribution with variance σ^2, the probability that x is in a fixed ball of radius ε is at most $(\varepsilon/\sigma)^d$.

Problem 4

The following claim can be used without a proof: Let P be a set of at least d points in $\mathbb{R}^d$, and let $\mathcal{H}$ be an arbitrary hyperplane. Then there exists a hyperplane $\mathcal{H}'$ passing through d points of P such that $\max_{p \in P}(\mathrm{dist}(p, \mathcal{H}')) \leq 2d \cdot \max_{p \in P}(\mathrm{dist}(p, \mathcal{H}))$.

Show that if the n points in X are chosen according to independent d-dimensional normal distributions with variance σ^2, then X is ε-separated with probability at least $1 - n^{2d}(4d\varepsilon/\sigma)^d$ for every $\varepsilon > 0$.