Prof. Dr. Johannes Blömer 14.01.2016
submission due: 21.01.2016, F1.110, 11:00

Clustering Algorithms

WS 2015/2016

Handout 9

Exercise 1:
Let $A = (r_{ij}) \in \mathbb{R}^{k \times d}$, where each r_{ij} is chosen according to $\mathcal{N}(0,1)$, $u \in \mathbb{R}^d$, $\|u\|_2 = 1$. Define the random variables

$$X_i = \sum_{j=1}^{d} r_{ij} u_j \quad \text{and} \quad Y = \|A \cdot u\|_2^2 = \sum_{i=1}^{k} X_i^2.$$

As we know from the lecture,

$$E[X_i] = 0,\ \operatorname{Var}(X_i) = 1 \text{ and } E[Y] = k.$$

Now use Chebyshev's inequality and determine for which k

$$\Pr\big((1-\epsilon)k \le \|A \cdot u\|_2^2 \le (1+\epsilon)k\big) \ge 1 - \frac{1}{3n^2}.$$

Hint: $E[X_i^4] = 3$.

Exercise 2:
If the random variable X is distributed according to the Gaussian distribution with mean μ and variance σ^2 ($X \sim \mathcal{N}(\mu, \sigma^2)$), then

$$\Pr(X \le a) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{a} \exp(-(x-\mu)^2/(2\sigma^2))\, dx.$$

If $X \sim \mathcal{N}(0,1)$, then $E[X] = 0$ and $\operatorname{Var}(X) = 1$.

Prove that

(a) If $X \sim \mathcal{N}(0,1)$, then $\sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

(b) If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $E[X] = \mu$ and $\operatorname{Var}(X) = \sigma^2$.

Exercise 3:
We are given a point set $P \subset \mathbb{R}^d$, $|P| = n$, a γ-approximation algorithm for the k-means problem, and an embedding $\pi : P \to \mathbb{R}^{c\log(n)/\epsilon^2}$ as given by Johnson-Lindenstrauss.

First, we apply the embedding and obtain a new point set $\pi(P) = \{\pi(p) \mid p \in P\} \subset \mathbb{R}^{c\log(n)/\epsilon^2}$ such that

$$(1-\epsilon) \cdot D_{l_2}(p,q) \le D_{l_2}(\pi(p), \pi(q)) \le (1+\epsilon) \cdot D_{l_2}(p,q)$$

for all $p, q \in P$.
Second, we use the γ-approximation algorithm of the k-means problem wrt. $\pi(P)$ and obtain a partition $\{C_1^\pi, \dots, C_k^\pi\}$ of $\pi(P)$.
Third, we obtain our solution wrt. P by defining a partition $\{C_1, \dots, C_k\}$ of P such that

$$\pi(C_i) = C_i^\pi.$$

Now show that

(a) for all $A \subset \mathbb{R}^d$

$$\frac{1}{2|A|} \sum_{p,q \in A} D_{l_2^2}(p, q) = D_{l_2^2}(A, \mu(A))$$

(b) we can bound the costs of our final solution by

$$D_{l_2^2}(P, \{\mu(C_1), \dots, \mu(C_k)\} \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 \gamma \cdot \sum_i D_{l_2^2}(O_i, \mu(O_i))$$

where $\{O_1, \dots, O_k\}$ denotes the optimal partition for the k-means problem wrt. P.