University of Augsburg, Institute of Computer Science — SS 2019
Prof. Dr. P. Fischer — 1. July 2019
J. Kastner, L. Rudenko — Solution 8

Analyzing Massive Data Sets

Exercise 1: BFR - variance (homework)

For calculation of variance vector of a cluster $\overrightarrow{\sigma^2}$ in each dimension we have used the following formula: $\overrightarrow{\sigma^2} = \frac{\overrightarrow{SUMSQ}}{N} - (\frac{\overrightarrow{SUM}}{N})^2$. Note, that all dimensions in the data points must be independent, because it is the condition for elliptical distribution of points along the axis. Thus, σ_i^2, SUM_i and $SUMSQ_i$ can be also calculated separately for each dimension i. So to prove that this formula is correct, we can prove it for one particular dimension i.

We do not know the true parameters for normal distribution of cluster points in this dimension, but we can estimate them by calculating sample mean and sample variance of points already assigned to a cluster.

We therefore consider only one dimension for calculation of σ^2 and the values of points in this dimension can be regarded as **random variable** X, then **expected value** $\mathbb{E}$ is:

$\mu := \mathbb{E}(X) = \sum_{i \in I}(x_i * p_i)$, where x_i is the value of random variable and p_i the probability for the value x_i.

All sample points in cluster have the same probability of $\frac{1}{N}$, where N is the number of elements in the cluster C.

$\Rightarrow \mu_i = \frac{\sum_{x \in C}(x_i)}{N}$, sample mean of a cluster for dimension i, cp. Sheet 06 (centroid),
and $\sum_{x \in C}(x_i) \equiv SUM_i$.

$Var(X) := \mathbb{E}(X - \mu)^2 = \sigma^2$, sample variance of a cluster for dimension i is also:
$\sigma_i^2 = \frac{\sum_{x \in C}(x_i - \mu_i)^2}{N}$

$$\sigma_i^2 = \frac{\sum_{x \in C}(x_i - \mu_i)^2}{N} = \frac{\sum_{x \in C}(x_i - \frac{SUM_i}{N})^2}{N} = \frac{\sum_{x \in C}(x_i^2 - 2x_i\frac{SUM_i}{N} + (\frac{SUM_i}{N})^2)}{N} =$$

$$= \frac{\sum_{x \in C}(x_i^2) - 2 * \frac{SUM_i}{N} * \sum_{x \in C}(x_i) + \sum_{x \in C}(\frac{SUM_i}{N})^2}{N} = \frac{SUMSQ_i}{N} - \frac{\frac{2 * SUM_i^2}{N}}{N} + \frac{N * (\frac{SUM_i}{N})^2}{N} =$$

$$= \frac{SUMSQ_i}{N} - 2 * (\frac{SUM_i}{N})^2 + (\frac{SUM_i}{N})^2 = \frac{SUMSQ_i}{N} - (\frac{SUM_i}{N})^2$$

Alternative solution would be to calculate variance with Steiner translation theorem $Var(X) := \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

Exercise 2: Boolean Retrieval (homework)

DNF: $(bread\ AND\ NOT\ game)\ OR\ (bread\ AND\ NOT\ work)$

Result:

- $(bread\ AND\ NOT\ game)$: D_3
- $(bread\ AND\ NOT\ work)$: D_1, D_3
- $(bread\ AND\ NOT\ game)\ OR\ (bread\ AND\ NOT\ work)$: D_1, D_3

Exercise 3: Boolean Retrieval (live)

The solution was discussed in the exercise.

Exercise 4: Fuzzy IR-model (live)

The solution was discussed in the exercise.