Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Janina Sontheim, Maximilian Hünemörder

Knowledge Discovery and Data Mining I

WS 2019/20

Exercise 2: Central Tendencies, Aggregation, Histograms

Exercise 2-1 Central Tendencies

You wake up one morning suddenly struck by an idea. A new and exciting purpose for your life! Triggered by a thought ascending from the deep dark parts of your dreams, you wonder how many ChocolateBuddies™ are in a bag of ChocolateBuddies™. The solution is easy, you rip open one of the hundreds of bags - you have somehow lying around in your bedroom - and start counting. After you determine that this bag contains an arbitrary amount n_1 of CBs™, you are not satisfied. You open up another m bags and realise that they all contain differing amounts of chocolately goodness. You are devastated! When somebody asks you about the amount of ChoclateBuddies™ in a bag, you want to give them a satisfying answer. You therefore want an estimated amount n_{opt} that is always the optimal answer to your problem.

(a) Show how to find the central tendency n_{opt} that minimizes the sum of squared distances of the samples to your estimate n_{est}.

$$\begin{aligned} \mathcal{L}(n_{est}) &= {d_1}^2 + {d_2}^2 + {d_3}^2 + ... + {d_m}^2 \\ n_{opt} &= \operatorname*{argmin}_{n_{est}} \mathcal{L}(n_{est}) \end{aligned}$$

(b) Show how to find the central tendency n_{opt} that minimizes the sum of absolute distances of the samples to your estimate n_{est}.

$$\begin{aligned} \mathcal{L}(n_{est}) &= |d_1| + |d_2| + |d_3| + ... + |d_m| \\ n_{opt} &= \operatorname*{argmin}_{n_{est}} \mathcal{L}(n_{est}) \end{aligned}$$

(c) Suddenly you have a moment of pure clarity and you begin to notice that ChocolateBuddies™ actually come in several different colors. Try finding the color that minimizes the sum of distances to the color of each single CB™ using the trivial metric.

$$d(o,p) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad o_i = p_i \\ 0 & \text{iff} \quad o_i \neq p_i \end{cases}$$

Exercise 2-2 Incremental Aggregation

Given a Data Warehouse with e.g. 10 million entries, additional 1000 entries arrive each day. Rather than recomputing the desired aggregates, an incremental adaptation to the new data should be supported. In order to accelerate the (re-)computation, precomputed intermediate results shall be stored and intermediate results for the new entries shall be computed. What (and how many) values suffice when considering the following aggregates? For each measure note whether it is an algebraic, holistic or distributive measure.

(a) Product.

(b) Mean.

(c) Variance.

(d) Median.

Exercise 2-3 Privacy

Given the following table

Key	**Quasi-Identifier**			**Sensitive**
Name	*Sex*	*Age*	*Zip*	*Disease*
Alice	F	24	10000	Heart Disease
Bob	M	22	10000	Lung Cancer
Charlotte	F	24	10000	Breast Cancer
Dave	M	22	10000	Lung Cancer
Emma	F	20	10000	Heart Disease
Francis	M	20	10000	Heart Disease
Garry	M	22	10000	Lung Cancer
Harry	M	20	10000	Heart Disease
Iris	F	21	10001	Flu
John	F	21	10001	Flu
Kendra	F	20	10000	Heart Disease
Lisa	F	20	10000	Lung Cancer

(a) k-Anonymity:

(i) Determine the largest k such that the table is k-anonym. Explain which rows contradict the $(k+1)$-anonymity.

(ii) You may now use suppression on the columns. Assume that by removing one digit from *Age* or *Zip*, or suppressing the *Sex* attribute, you lose one "value". What is the minimal value loss required to achieve 5-anonymity?

(b) Distinct l-Diversity

(i) What is one shortcoming of k-anonymity compared to l-diversity? Which attack exploits this weakness?

(ii) Given that a dataset is k-anonymous, but not $(k+1)$-anonymous. What implications does this have on the distinct l-diversity of the dataset? Give a lower and upper bound for l.

(iii) Knowing only the distribution of the sensitive attribute values; What bounds can you derive for l in distinct l-diversity?

(iv) What is the largest l such that the above mentioned dataset is distinct l-diverse?

(v) Assume suppressing the last digit of the *Zip* column and generalising *Age* to $\{(-\infty, 22], (22, +\infty)\}$. For what value of l can distinct l-diversity now be guaranteed.