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

Knowledge Discovery in Databases

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}.

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

First let us start with a few small transformations:

$$\begin{aligned} \mathcal{L}(n_{est}) &= {d_1}^2 + {d_2}^2 + {d_3}^2 + ... + {d_m}^2 \\ &= (n_1 - n_{est})^2 + (n_2 - n_{est})^2 + (n_3 - n_{est})^2 + ... + (n_m - n_{est})^2 \\ &= \sum_i (n_i - n_{est})^2 \end{aligned}$$

Now in order to find the n_{opt} we need to find $\operatorname{argmin}_{n_{est}} \mathcal{L}(n_{est})$, i.e. the value for which the Loss Function is minimal. Instead of using a time consuming optimization algorithm, we can instead zeroise the deriviation of the Loss Function:

$$\begin{aligned}
\frac{\partial}{\partial n_{est}}\mathcal{L}(n_{est}) &= 0 \\
&= \frac{\partial}{\partial n_{est}}\sum_i (n_i - n_{est})^2 \\
&= \sum_i \frac{\partial}{\partial n_{est}}(n_i - n_{est})^2 \\
&= \sum_i -2 \cdot (n_i - n_{est}) \\
\sum_i (n_i - n_{est}) &= 0 \\
\sum_i n_i - \sum_i n_{est} &= 0 \\
\sum_i n_i - m \cdot n_{est} &= 0 \\
m \cdot n_{est} &= \sum_i n_i \\
n_{est} &= \frac{1}{m}\sum_i n_i
\end{aligned}$$

The results is an old familiar friend. It is of course the mean!

(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}.

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

Granted that n_{est} is an instance of the dataset and the distance to itself is not included in the Loss Function we can derive and zeroise $L(n_{est})$ and find the optimal estimate.

$$\begin{aligned}
\frac{\partial}{\partial n_{est}}\mathcal{L}(n_{est}) &= 0 \\
&= \frac{\partial}{\partial n_{est}}\sum_i |d_i| \\
&= \sum_i \frac{\partial}{\partial n_{est}}|d_i| \\
&= -\sum_i sign(d_i)
\end{aligned}$$

Now if the amount of samples is odd, we pick the middle sample. If we now sort the summands and split the sum of distances into two parts, one containing all values lower than n_{est} and one with

all the values higher than n_{est} the sign function for the first sum will result in -1 and in 1 for the second sum. Each of these sums will have $\frac{n}{2}$ summands.

$$\begin{aligned}\sum_{i}^{\frac{n}{2}} sign(d_i) + \sum_{j=\frac{n}{2}+1}^{n} sign(d_j) &= 0 \\ \frac{n}{2} \cdot -1 + \frac{n}{2} \cdot 1 &= 0 \\ \frac{n}{2} - \frac{n}{2} &= 0\end{aligned}$$

Therefore in this case the middle object is the central tendency. If there is an even number of samples we can choose one of the two middle objects, because while both of them do not lead to the derivation being zero, both objects are the closest we can get to minimal given that we have to pick an existing sample. If we do not have to pick an object from the dataset, we can take the mean of the middle objects and the deriviation will again be zero. This is of course the concept of the median.

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

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

In this case each object either has a distance of one or zero from our estimate. If we now pick the most frequent color the least amount of distances will be one compared to the other colors. Therefore the sum will be minimal. This is of course the concept of the mode.

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.

The product is a distributive aggregation measure since it is an associative pairwise operation:

$$\begin{aligned}prod(D) &= \prod_{x \in D} x \\ &= \left(\prod_{x \in D_1} x\right) \cdot \left(\prod_{x \in D_2} x\right) \\ &= prod(prod(D_1), prod(D_2))\end{aligned}$$

(b) Mean.

Let $D = D_1 \cup D_2$ with $|D_1| = n_1$ and $|D_2| = n_2$ where D_1 is the data currently in the data warehouse and D_2 is the increment. It suffices to store two values for D_1 and D_2, the sum and $count$, since

$$\begin{aligned} mean(D) = \frac{1}{n_1 + n_2} \sum_{x \in D} x &= \frac{\sum_{x \in D_1} x + \sum_{x \in D_2} x}{n_1 + n_2} \\ &= \frac{sum(D_1) + sum(D_2)}{count(D_1) + count(D_2)}. \end{aligned}$$

Thus, the mean is an algebraic measure. It is not a distributive measure. Towards contradiction assume it would, i.e. for all databases D and partitions $D_1 \uplus D_2$ it holds $mean(D) = mean(mean(D_1), mean(D_2))$, i.e. in particular for $D = \{0, 2, 4, 6\}$, and the partition $D = D_1 \uplus D_2$ with $D_1 = \{0\}$, $D_2 = \{2, 4, 6\}$. Then

$$\begin{aligned} mean(D) &= mean(mean(D_1), mean(D_2)) \\ \frac{0+2+4+6}{4} &= \frac{1}{2}\left(\frac{0}{1} + \frac{2+4+6}{3}\right) \\ \frac{12}{4} &= \frac{1}{2} \cdot \frac{12}{3} \\ 3 &= 2 \end{aligned}$$

which is a contradiction.

To further derive the conditions when the distribution works, consider

$$\begin{aligned} mean(D) &= mean(mean(D_1), mean(D_2)) \\ \frac{1}{n_1 + n_2} \sum_{x \in D} x &= \frac{1}{2}\left(\frac{1}{n_1} \sum_{x \in D_1} x + \frac{1}{n_2} \sum_{x \in D_2} x\right) \\ \frac{1}{n_1 + n_2} \sum_{x \in D_1} x + \frac{1}{n_1 + n_2} \sum_{x \in D_2} x &= \frac{1}{2n_1} \sum_{x \in D_1} x + \frac{1}{2n_2} \sum_{x \in D_2} x \\ \left(\frac{1}{n_1 + n_2} - \frac{1}{2n_1}\right) \sum_{x \in D_1} x &= \left(\frac{1}{2n_2} - \frac{1}{n_1 + n_2}\right) \sum_{x \in D_2} x \\ \left(\frac{2n_1 - (n_1 + n_2)}{2n_1(n_1 + n_2)}\right) \sum_{x \in D_1} x &= \left(\frac{n_1 + n_2 - 2n_2}{2n_2(n_1 + n_2)}\right) \sum_{x \in D_2} x \\ \left(\frac{n_1 - n_2}{2n_1(n_1 + n_2)}\right) \sum_{x \in D_1} x &= \left(\frac{n_1 - n_2}{2n_2(n_1 + n_2)}\right) \sum_{x \in D_2} x \\ \left(\frac{n_1 - n_2}{n_1}\right) \sum_{x \in D_1} x &= \left(\frac{n_1 - n_2}{n_2}\right) \sum_{x \in D_2} x \\ \frac{1}{n_1} \sum_{x \in D_1} x &= \frac{1}{n_2} \sum_{x \in D_2} x \end{aligned}$$

The last operation is only an equivalence if $n_1 \neq n_2$. If $n_1 = n_2$, the statement holds trivially. Concluding, the mean can be computed in distributive manner if and only if the partitions have same size, or the same mean.

(c) Variance.

Similarly, the variance is also an algebraic measure:

$$
\begin{aligned}
var(D) &= \frac{1}{n_1+n_2-1}\left(\sum_{x\in D} x^2 - \frac{1}{n_1+n_2}\left(\sum_{x\in D} x\right)^2\right) \\
&= \frac{1}{n_1+n_2-1}\left(\sum_{x\in D} x^2 - \frac{1}{n_1+n_2}\left(\sum_{x\in D} x^2 + \sum_{x\in D_1, y\in D_2} xy + \sum_{x\in D_1, y\in D_2} yx\right)\right) \\
&= \frac{1}{n_1+n_2-1}\left(\sum_{x\in D} x^2 - \frac{1}{n_1+n_2}\left(\sum_{x\in D_1} x^2 + \sum_{x\in D_2} x^2 + 2\left(\sum_{x\in D_1} x\right)\left(\sum_{x\in D_2} x\right)\right)\right) \\
&= \frac{ss(D_1)+ss(D_2)-\frac{1}{count(D_1)+count(D_2)}\left(ss(D_1)+ss(D_2)+2\cdot sum(D_1)\cdot sum(D_2)\right)}{count(D_1)+count(D_2)-1}
\end{aligned}
$$

We need to store three values, the sum, $count$ and additionally the sum of squares (ss). Note that the variance is not distributive, since the information about central tendency is lost (the variance is shift-invariant). The variance $var(D)$ depends on where D_1 and D_2 are located in the data space and in general there is no way to infer that from $var(D_1)$ and $var(D_2)$ alone. However, if $mean(D_1) = mean(D_2) = 0$, one can show that

$$var(D) = \frac{n_1}{n_1+n_2} var(D_1) + \frac{n_2}{n_1+n_2} var(D_2).$$

(d) Median.

The median is a classical holistic measure which means intuitively that we need to look at the whole data at once in order to compute it. For the median to be an algebraic measure, we would need to be able to represent the median of D as an algebraic function of constant size aggregates of D_1 and D_2. Assume that we have computed such aggregates. Now the idea is that for any two sets D_1 and D_2, we can construct an example where the k-th element of D_1 (or D_2) is the median. That is, we potentially need to access every single element in D_1 (or D_2) from a constant size aggregate. This is clearly not possible. Thus, we need to look at the whole sets D_1 and D_2 together in order to find the median, i.e. the median is a holistic measure.

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.

The dataset is 2-anonymous, as there is no Quasi-Identifier-tuple which occurs only once. It is not 3-anonymous, as e.g. $(F, 24, 10000)$ occurs only twice.

(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?

5-anonymity can be achieved by suppressing the last digit of Age and the last digit of Zip. Hence, the minimal value is at most 2. It is not 1 as:

- Suppressing *Sex* leads to 2-anonymity, e.g. $(*, 24, 10000)$ occurs only twice.
- Suppressing the last digit of *Age* leads to 2-anonymity, e.g. $(F, 2*, 10001)$ occurs only twice. Suppressing the first digit does not give any benefit, as all age numbers begin with "2".
- Suppressing the last digit of Zip leads to 2-anonymity, e.g. $(F, 24, 1000*)$ occurs only twice. Suppressing any other digit does not give any benefit, as all zip codes begin with "1000".

(b) Distinct l-Diversity

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

k-anonymity only regards the quasi-identifiers, but does not investigate the distribution of the sensitive attribute within one equivalence-class w.r.t. the quasi-identifier. This can be exploited by the *Background-Knowledge Attack*.

(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.

The smallest equivalence-class w.r.t. to the Quasi-Identifier has size k. Hence, in this class there can only be at most k different values for the sensitive attribute. Thus, l can be bounded from above as $l \leq k$. Trivially, $1 \leq l$ holds as lower bound. As k-anonymity does not make any statement about the distribution of the sensitive attribute, we cannot guarantee a larger lower bound, i.e. the following bounds are tight: $1 \leq l \leq k$.

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

Let L be the number of different sensitive attribute values. Then, there can also be at most L different values within each equivalence class w.r.t. to an Quasi-Identifier. Thus, $l \leq L$.
Additional information: This bound is independent of the bound from (ii), as the former one operates only on the Quasi-Identifier columns and this one solely considers the sensitive attribute.

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

The dataset is distinct 1-diverse as $QI = (F, 21, 10001) \implies Disease = Flu$.

(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.

There are the following equivalence classes

Sex	Age	Zip	Diseases	l
F	$(-\infty, 22]$	$1000*$	{Flu, Heart Disease, Lung Cancer}	3
M	$(-\infty, 22]$	$1000*$	{Heart Disease, Lung Cancer}	2
F	$(22, \infty)$	$1000*$	{Breast Cancer, Heart Disease}	2

Hence, the table is now distinct 2-diverse.