Universität Bielefeld

CITEC / Faculty of Technology
Ambient Intelligence Group
Dr. Thomas Hermann

Sekretariat: S. Strunk
Office: CITEC-3.311
106-6891

Exercises of the lecture "Introduction to Data Mining"

WS 2018/2019 Excercise sheet 3

- Dependency analysis -

Problem 3.1, Linear correlation as optimization task

The linear correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

is connected to the optimal estimator a and b regarding the cost function

$$E(a, b) := \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

by

$$E_{\min} = (1 - r^2) \sum_{i=1}^{N} (y_i - \bar{y})^2$$

a) Calculate the optimal solution (Extremum of $E(a, b)$) by setting the gradient $\frac{\partial E}{\partial a}$, or $\frac{\partial E}{\partial b}$ to zero. What is the equation for minimal error that you get?

```
In [ ]:
```

b) Now examine the data set {(1,2), (2.5, 7), (3, 9), (2.8, 7), (1.4, 4), (3.5, 7.5), (4, 9), (3.2, 6.8)}.

- How big is r?
- Which are the optimal parameters a and b?
- How big is the error?
- Does the connection that we formulated in **(a)** hold?
- How does a measuring error "(4,.9) instead of (4,9)" influence r?

Plot the data set and compare the regression line.

```
In [ ]:
```

c) Analyze the data set in a similar manner using the non-parametric correlation.

- How big is r_{sp}? Can we reject the null hypothesis of uncorrelatedness?
- What's the meaning of the parameters a and b in the case of rank correlation?
- How does a transmission error "(4,-9) instead of (4,9)" influence the results now?

In [ ]:

d) Additional Task (optional, but a good exercise): Prove the above mentioned relation between r and E_{min}.

(Hint: You can assume $\bar{x} = 0$ without loss of generality and work with the vectors $\vec{m}_x$ and $\vec{m}_y$, that we discussed during the geometric interpretation of r during the lecture.)

In [ ]:

Thomas Hermann (thermann@techfak)	Lecture Wed 12-14, CITEC lecture hall	office building	public transportation
Ferdinand Schlatt:	Tutorial Tue 12:15-13:45, UHG H11	Universität Bielefeld	light rail 4
		Universitätsstraße 25	from Bahnhof and Jahnplatz
		33615 Bielefeld	

In [ ]: