Exam Winter Semester 2020-2021

Data Mining und Knowledge Discovery (IN2030) (Technische Universität München)

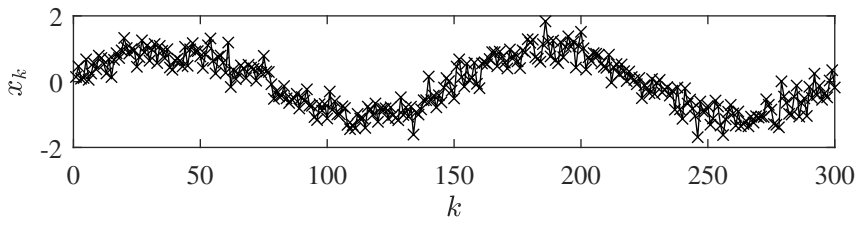
Problem 1: Errors (8 of 46 points)

Which kind of errors seem to be contained in the following time series data sets?

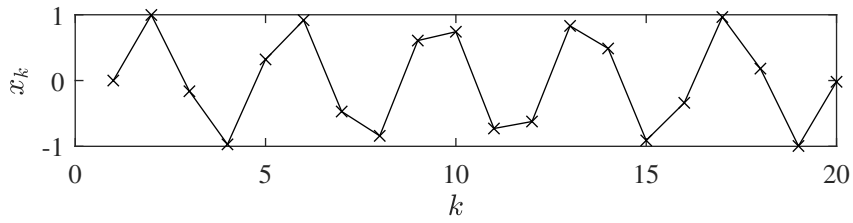
a) $X = (-1.3499, 3.0349, 0.7254, -0.0631, 0.7147, -2050, -0.1241)$

b) $X = (0.72, 0.72, 0.72, 1.08, 1.08, 1.44, 0.72, 0.72)$

c)



d)



Problem 2: Correlation (10 of 46 points)

Consider the data set $X = \{(1,1), (1,5), (2,2), (2,4), (3,3)\}$.

a) What is the value of Pearson correlation between these two features? Explain!

b) How many bins do you need for these two features, so that the chi–square test for independence will indicate the maximum possible correlation?

c) Using the bins from (b), which points would you add to X, so that the chi–square method will indicate the minimum possible correlation?

d) Now consider the data set $X = \{(-0.5, -0.5), (0.5, 0.5)\}$ and two bins $[-1, 0)$, $[0, 1)$ for each feature. Which points would you add to X, so that the Pearson correlation will be very high, but the chi–square method will indicate a very low correlation?

e) What do you learn from this?

Problem 3: Classification (16 of 46 points)

Consider a classifier C_1 with a true positive rate of $3/4$ and a false positive rate of $2/3$. We construct a probabilistic classifier C_2 that takes the output of classifier C_1 and inverts it with a probability $p \in [0, 1]$.

a) Compute the expected value of the true positive rate of C_2 and the expected value of the false positive rate of C_2.

b) The expected behavior of C_2 should be as close as possible to the top left corner of the receiver operator characteristic. For which value of $p = p^*$ will this be achieved?

c) For $p = p^*$, compute the expected value of the true positive rate of C_2 and the expected value of the false positive rate of C_2.

Problem 4: Clustering (12 of 46 points)

Are the following statements about clustering true or false?

a) true ☐ false ☐ In alternating optimization, termination on U is usually more efficient than termination on V.

b) true ☐ false ☐ Possibilistic clustering can be used to find ellipsoidal clusters.

c) true ☐ false ☐ In fuzzy c–means, the sum of memberships of one severe outlier in all clusters will be approximately equal to one.

d) true ☐ false ☐ In fuzzy c–means, the sum of memberships of c severe outliers in one cluster will be approximately equal to one.

e) true ☐ false ☐ For data without noise, noise clustering will perform poorly.

f) true ☐ false ☐ Alternating optimization will always converge to a local or a global optimum of the objective function .