Machine Learning and Data Mining

Assignment 1

Dr. Zeyd Boukhers
boukhers@uni-koblenz.de

Raphael Menges
raphaelmenges@uni-koblenz.de

Akram Sadat Hosseini
sadathosseini@uni-koblenz.de

Qusai Ramadan
qramadan@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: November 4, 2019, 09:00 a.m.
Tutorial on: November 7, 2019

The exercises in this assignment are of theoretical nature and may not be solved by execution of high-level Python commands but through **manual step-by-step calculations which must be included in submissions**. For this assignment it is also allowed to upload a single .pdf file generated from LATEX code or scanned and compressed (!) handwritten solutions

Team Name: omega

1. Stanislav Avramenko : 219202429
2. Vanessa Joan Chebet : 219202398
3. Mirza Shuja Mohiuddin : 219202626
4. Oleksandr Tarasov : 219202510

1 Statistics (18 Points)

a. **(6 points)** The following values for the attribute x are given: 5, 3, 4, 4, 6, 20, 2, 3, 3, 10. Determine mean, median, mode, variance, and standard deviation of x.

b. **(6) points** Let X be a random variable of a specific scenario. The probability distribution of X is given in the following table:

X = x	1	2	3	4	5
P(X = x)	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{5}$

Compute the values of mean (E[X]), variance (Var[X]), and standard deviation (SD[X])

c. **(6 points)** Information in random variables can be categorized according to their scale of measurement. State at least three scales of measurement,1 describe them shortly and provide a small real-world example.

Solution

a. We have array of numbers [5,3,4,4,6,20,2,3,3,10]

- Mean value
  $\mu = \frac{\sum_{i=1}^{n}(x_i)}{n}$
  $\mu = \frac{5+3+4+4+6+20+2+3+3+10}{10} = \frac{60}{10} = 6$
  Mean value is 6.

- Median
  Let's sort the array.
  Finally we have an array [2,3,3,3,4,4,5,6,10,20]
  We have paired number of elements(10).
  $median = \frac{x_{n/2}+x_{n/2+1}}{2}$
  $median = \frac{8}{2} = 4$
  Median value is 4.

- Mode
  Mode is the most common value in the sample. Let's make the frequency table :

number	2	3	4	5	6	10	20
frequency	1	3	2	1	1	1	1

  The biggest frequency(3) has number 3.
  The mode value is 3.

- Variance
  $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$
  , where μ is the mean value

$\sigma^2 = \frac{16+9+9+9+4+4+1+0+16+196}{10} = \frac{264}{10} = 26.4$
The variance value is 26.4

- Standard deviation
$\sigma = \sqrt{\sigma^2}$
$\sigma = \sqrt{26.4} = 5.1380930314660516$
Standard deviation value is 5.1380930314660516.

b. We have the table with X random discrete values and it‘s distribution.

- Mean
$E[X] = \sum_{i=1}^{n}(x_i p_i)$
$E[X] = \frac{2}{5} + \frac{2}{10} + \frac{3}{5} + \frac{4}{10} + 1 = \frac{26}{10} = 2.6$
Mean value is 2.6

- Variance
$Var[X] = \sum_{i=1}^{n}(x_i - E[X])^2 p_i$
$Var[X] = 2.44$
Variance value is 2.44

- Standard deviation $SD[X] = \sqrt{Var[X]}$
$SD[X] = \sqrt{2.44} = 1.562049935$
Standard deviation value is 1.562049935

c. There are four different scales of measurement: Nominal, Ordinal, Interval, Ratio.

- Nominal(categorical) scale of measurement is used when variables have no numeric values. As a result these variables can‘t be used for arithmetic operations or for ordering, but can be divided into categories.
As an example we can take the value of Sex property. It can take values "Male" or "Female". Usually for comfortable using of values we determine them numeric indexes("Male"=1, "Female"=2 or "Male"=2, "Female"=1). But it does not matter that "Male" variable is bigger than "Female". Their indexes have no empirical value.

- Ordinal scale is used for variables that can be submitted like an ordered list. As a result indexes are determined through the variable‘s rank. But empirical value does not depend on difference of neighbouring indexes. The equality of the index‘s difference in two pairs of neighbour‘s variables does not mean that real differences of measurements of these objects are the same.
As an example we can take graduation of people by smoking frequency. Non-smoker=1, rarely=2, often=3, always=3. The difference between non-smoker‘s and rarely smoker‘s indexes is 1. The same is for rarely smoker and often smoker. But that does not mean of real equality.

- Interval scale is used when variables have ordered values with significant intervals. If the difference between 1st and 2d value is the same like differences

that 3d and 4th values have, than the differences are equal. But the attitude of two variables means nothing.
As an example we can take I.Q. measurements. If two people have 40 and 80 IQ units, that does not matter, that second one is smarter twice than first.

- Ratio scale is the same like interval scale but also with absolutely zero point(measurement does not exist) and meaningful attitude.
As an example we can take age measurements. If two people are 10 and 20 years old, it matters that the second one is twice older than the first.

2 Error Calculation(12 points)

You are given n-many computed outputs y_i and desired outcomes $\widetilde{y}_i(\widetilde{y}1, 2, n)$. Provide the following error measures in regards to y_i and $\widetilde{y}_1$ by writing down their formula and a short description about their characteristics, i.e., the behavior in regard to the difference between computed and desired outcome.

a. (3 points) Sum of Square Error (SSE)

b. (3 points) Mean Square Error (MSE)

c. (3 points) Root Mean Square Error (RMSE)

d. (3 points) Mean Absolute Error (MAE)

Solution

a. Sum of Square Error (SSE)
$SSE = \sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2$
Common error is the sum of square errors. The less each error, the less common error is. So the closer real values to desired values the smaller common error is. When the computed and desired values are the same the sum function take 0 value(minimum). This measurement is also very sensitive to large errors because of square operation by this error calculation.

b. Mean Square Error (MSE)
$MSE = \frac{\sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2}{n}$
The mean square error always non-negative. It is used to evaluate estimator. The higher mean square error, the worse the desired data is. The closure desired variables to real values, the closure value of mean square error to 0. This error is sensitive to difference changes because of squared calculations in formula.

c. Root Mean Square Error (RMSE)
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2}{n}}$
The root mean square error always non-negative. If it take 0 value than the prediction is excellent, but it happens almost never. The better prediction of the estimator the smaller RMSE is.

d. Mean Absolute Error (MAE)
$MAE = \frac{\sum_{i=1}^{n}|y_i - \widetilde{y}_i|}{n}$
The mean absolute error characterizes the precision of prediction. The smaller difference between actual and desired data the closer MAE to 0. This valuation is always non-negative.

3 Statistical Evaluation (30 Points)

Imagine that we have already developed algorithms that approximate values of a real experiment.

a. (15 points) Some real binary outcomes and predicted binary outcomes of one algorithm are stated in the following table:

Index	Real Outcome	Predict Outcome
0	false	true
1	true	true
2	false	true
3	false	false
4	true	true
5	true	false
6	false	false
7	true	false
8	true	false
9	false	true

Provide for each predicted outcome, whether it is a true positive, a true negative, a false positive, or a false negative prediction. Compute then accuracy, precision, recall, and F1-score. In addition, state the general formulas to compute these values. Describe the intention of the F1-score shortly.

b. (15 points) Some real numbers and predicted numbers of one algorithm are stated in the following table:

Index	Real Outcome	Predict Outcome
0	1	4
1	2	4
2	5	3
3	4	9
4	2	3
5	6	3
6	1	1
7	4	4
8	3	2
9	3	3

3. Statistical Evaluation.

a)

Index	Real Outcome	Predicted Outcome	
0	false	true	False Positive
1	true	true	True Positive
2	~~true~~ false	true	False Positive
3	false	false	True Negative
4	true	true	True Positive
5	true	false	False Negative
6	false	false	True Negative
7	true	false	False Negative
8	true	false	False Negative
9	false	true	False Positive

Confusion matrix:

		Predicted outcome: Negative	Predicted outcome: Positive
Actual outcome	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive = 2
True Negative = 2
False Positive = 3
False Negative = 3

i) Precision $= \frac{\text{True Positive}}{\text{True Positive + False Positive}}$

$= \frac{\text{True Positive}}{\text{Total Predicted Positive}}$

$= \frac{2}{5} = \frac{4}{10} = 0.4$

ii) Recall $= \frac{\text{True Positive}}{\text{True Positive + False Negative}}$

$= \frac{\text{True Positive}}{\text{Total Actual Positive}}$

$= \frac{2}{5} = 0.4$

iii) Accuracy $= \frac{\text{True Positive + True Negative}}{\text{True Positive + True Negative + False Positive + False Negative}}$

$= \frac{2+2}{2+2+3+3} = \frac{4}{10} = 0.4$

iv) F1-score $= F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$

$= 2 \times \left(\frac{0.4 \times 0.4}{0.4 + 0.4}\right)$

$= 2 \left(\frac{0.16}{0.8}\right)$

$= 2 \cdot 0.2$

$= 0.4$

ii) F1-score

F1 score is used to ensure there's a balance between Precision and recall when evaluating models to be used to solve classification problems.

b)

Index	Real Outcome	Predicted outcome	$(\hat{y}_i - y_i)^2$
0	1	4	$(-3)^2 = 9$
1	2	4	$(-2)^2 = 4$
2	5	3	$(2)^2 = 4$
3	4	9	$(-5)^2 = 25$
4	2	3	$(-1)^2 = 1$
5	6	3	$(3)^2 = 9$
6	1	1	$(0)^2 = 0$
7	4	4	$(0)^2 = 0$
8	3	2	$(1)^2 = 1$
9	3	3	$(0)^2 = 0$

i) $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$= 9 + 4 + 4 + 25 + 1 + 9 + 0 + 0 + 1 + 0$

$= \underline{53}$

ii) $MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

$= \frac{53}{10} = \underline{5.3} = 5.3$

iii) $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$

$= \sqrt{5.3}$

$= 2.30$

iv) $MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$

$= \frac{|1-4| + |2-4| + |5-3| + |4-9| + |2-3| + |6-3| + |1-1| + |4-4| + |3-2| + |3-3|}{10}$

$= \frac{3+2+2+5+1+3+0+0+1+0}{10}$

$= 17/10 = 1.7$

4 Vector distances(10 points)

Given the two following vectors:
x = (1, 1, 3, 2, 1) y = (1, 2, 1, 1, 0)

a. (6 points) The distance between two vectors = $(x_1, , x_n)$ and $y = (y_1, , y_n)$ provides us information about the similarity of those vectors. By writing down step-by-step of your solution, calculate the Euclidean distance, Manhattan distance, and Chebyshev distance. For each case write also the general equation of the distance measurement.

b. (4 points) Normalize the vector w = (3, 0, 4) using (i) the norm of the vector and (ii) the standard deviation (z-scores).

Machine Learning & Data Mining

4 Vector Distances :-

$$x = (\underset{x_0}{-1}, \underset{x_1}{-1}, \underset{x_2}{3}, \underset{x_3}{2}, \underset{x_4}{-1})$$

$$y = (\underset{y_0}{-1}, \underset{y_1}{2}, \underset{y_2}{-1}, \underset{y_3}{1}, \underset{y_4}{0})$$

a) Euclidean distance $= \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$

$$= \sqrt{(-1+1)^2 + (-1-2)^2 + (3+1)^2 + (2-1)^2 + (-1+0)^2}$$

$$= \sqrt{0^2 + (-3)^2 + (4)^2 + (1)^2 + (-1)^2}$$

$$= \sqrt{9+16+1+1} = \underline{\underline{5.196}}$$

b) Manhattan distance $= \sum_{i=0}^{n} |x_i - y_i| = |-1+1| + |-1-2| + |3+1| + |2-1| + |-1-0|$

$$= 0+3+4+1+1 = \underline{\underline{9}}$$

Chebyshev distance $= \max [|x_i - y_i|]$

c) Chebysev distance $= \max [|-1+1| + |-1-2|, |3+1|, |2-1|, |-1-0|]$

$= \max[0, 3, 4, 1, 1] = \underline{4}$

b) Normalize the vector $w = (3, 0, 4)$ using (i) norm
(ii) standard deviation.

(i) Normalizing with norm

$w = (3, 0, 4)$ $\quad \|w\| = \sqrt{3^2 + 0^2 + 4^2}$

$= \sqrt{9+16} = \sqrt{25} = \underline{5}$

$$\bar{w} = \frac{w}{\|w\|} = \frac{(3, 0, 4)}{5} = (3/5, 0, 4/5)$$

$$\boxed{\bar{w} = (0.6, 0, 0.8)}$$

(ii) Normalizing with standard deviation. (Z scores)

$$Z_n = \frac{W_n - M}{\hat{s}}$$

M = Mean

$\hat{s}$ = Standard deviation.

$$\hat{s} = \sqrt{\frac{\Sigma (W_n - M)^2}{N - 1}}$$

$$= \sqrt{\frac{(3-2.33)^2 + (0-2.33)^2 + (4-2.33)^2}{3-1}}$$

$$M = \frac{3+0+4}{3} = 7/3 = \underline{2.33}$$

$$\hat{S} = \sqrt{\frac{(0.67)^2 + (-2.33)^2 + (1.67)^2}{2}} = \sqrt{\frac{0.4489 + 5.4289 + 2.7889}{2}}$$

$$= \sqrt{\frac{8.6667}{2}} = \sqrt{4.3334} = \sqrt{2.1667}$$

$$= \underline{\underline{1.472}}$$

$$Z_n = \frac{W_n - M}{\hat{S}} = \frac{\left[(3-2.33), (0-2.33), (4-2.33)\right]}{1.472}$$

$$= \frac{\left[0.67, -2.33, 1.67\right]}{1.472}$$

$$Z_n = \left[0.455, -1.583, 1.135\right]$$

5 Supervised VS Unsupervised Learning (30 Points)

The data pre-processing, learning via training, learning without training, and testing are main components in machine learning. However, depending on the machine learning type (i.e. supervised vs unsupervised machine learning) some of these components are required while some are not.

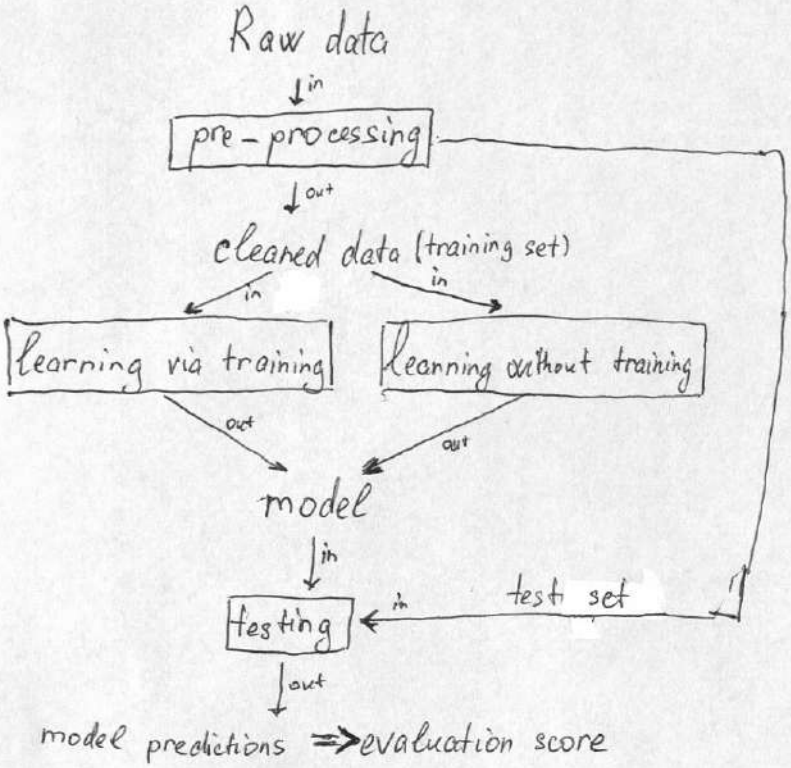
a. (13 points) What are the main components for a supervised machine learning process? Please illustrate the interactions between the components graphically and identify the input and the output of each component.

b. (7 points) What are the main components of an unsupervised machine learning process? Please illustrate the interactions between the components graphically and show the input and the output of each component.

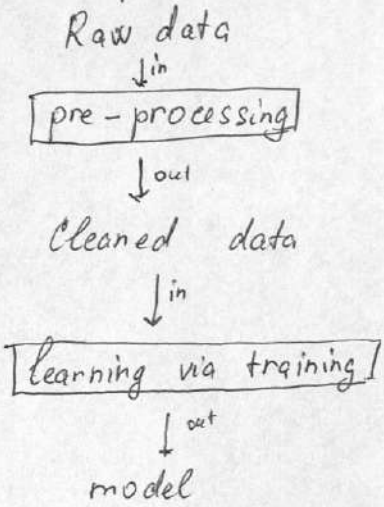
c. (10 points) Describe the differences between the supervised and unsupervised machine learning processes by referring to your graphics in (a) and (b).

5 Supervised VS Unsupervised Learning

a) Supervised



b) Unsupervised



c) There is a difference at the learning stage. In a supervised machine learning process, besides learning via training, we can also use learning without training so-called „instance based learning". Also, in supervised learning we split data into training and test set. In addition, in unsupervised machine learning process it is not required to do testing. Performing clustering, for example, we don't usually have ground truth labels to compare them with model's answers.