TU Delft

CYBER DATA ANALYTICS (CS4035)

Resit Exam, August 14 2018
9:00 - 12:00

Important: This exam consists out of 5 questions. Always give full explanations of your answers and number all steps of the asked algorithms. Do not forget to put your name and student number on every sheet of paper. Answers are required to be in English.

Question 1 - class imbalance

Boosting is a commonly used method to reduce the bias in machine learning methods. It updates sample weights D_t for learning hypothesis models h_t in several iterations t using for example the following formula:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

where α_t is the model importance term, $h_t(x_i)$ is a function returning the label assigned by the model learned in iteration t to sample i, y_i is the true label of sample i, and Z_t is a normalizing term.

(a) Give a reformulation of the above update equation that will perform better on imbalanced fraud data. Briefly explain the effect of this reformulation. (5 pt)

(b) Is there a downside to using boosting methods for fraud detection? Explain why (not). (5 pt)

(c) The Gini index ($2 \cdot \frac{n_+}{n} \cdot \frac{n_-}{n}$) and the square root Gini index ($\sqrt{\frac{n_+}{n} \cdot \frac{n_-}{n}}$) are two commonly used splitting criteria in decision tree learning. Suppose you have 10 positives and 10 negatives, and you need to choose between two splits $[8+, 2-][2+, 8-]$ and $[10+, 6-][0+, 4-]$. Computing the averages of these two criteria give the following conflicting results:

split	Gini	$\sqrt{\text{Gini}}$
$[8+, 2-][2+, 8-]$	0.32	0.4
$[10+, 6-][0+, 4-]$	0.375	0.387

Given that in your fraud detection problem, mistakingly classifying a positive (fraud) as negative (normal) is about 10 times as costly as misclassifying a negative, which split would you prefer? Explain why. (5 pt)

Question 2 - time series and anomaly detection

(a) Briefly explain how to use ARMA for anomaly detection in multi-dimensional time series. What type of correlation between signal values does it use for detection? (5 pt)

(b) Briefly explain how to use PCA for anomaly detection in multi-dimensional time series. What type of correlation between signal values does it use for detection? (5 pt)

(c) Briefly explain distance-based anomaly detection using the k-Nearest Neighbors algorithm. (5 pt)

(d) Briefly explain how to modify the algorithm under (c) to be density-based. Give one advantage of using a density-based anomaly detection algorithm over a distance-based algorithm. (5 pt)

Question 3 - hashing

Consider a stream W of length n. Each item i in the stream is an integer. A Bloom filter is a specific data structure that allows for efficient checking whether an element i is in W.

(a) Which guarantees does a Bloom filter provide for the correctness of such checks? (5 pt)

Suppose a Bloom filter of size 5 uses the following two hash functions:

$$h(i) = 2(i-1) \mod 5$$
$$h(i) = 3(i+1) \mod 5$$

(b) Compute the content of the filter after observing the following stream $W : 2, 3, 8, 15$. (5 pt)

(c) Find at least two items for which a membership check in the Bloom filter of (b) gives the correct answer, and two for which the check gives the wrong answer. (5 pt)

(d) Locality sensitive hashing is a technique that can speed up nearest neighbor testing using hash functions. Explain how this technique works for nearest neighbor search and what speed-up to expect when using m hashing bins to search over n data points, assuming uniform hash functions. (5 pt)

Question 4 - sequential data mining

Given the following subset of NetFlow data sampled from a computer network:

Time	Duration	Protocol	Src IP Addr	Dst IP Addr	Flags	Tos	Packets	Bytes	Flows
07.692	0.000	UDP	147.32.85.26	147.32.84.229	INT	0	1	76	1
07.692	0.000	UDP	83.200.200.38	118.161.179.143	INT	0	1	60	1
07.695	4.782	TCP	147.32.85.26	195.250.146.6	A	0	21	1260	1
07.718	0.831	TCP	80.13.129.154	147.32.84.229	FPA	0	4	417	1
07.724	0.000	UDP	83.200.200.38	147.32.86.165	INT	0	1	74	1
07.725	0.000	UDP	147.32.86.165	83.200.200.38	INT	0	1	61	1
07.747	0.000	UDP	192.168.88.150	192.168.95.255	INT	0	1	199	1
07.750	0.000	UDP	90.151.97.146	147.32.84.229	A	0	1	60	1
07.754	0.000	UDP	83.200.200.38	147.32.84.229	INT	0	1	76	1
07.755	0.000	UDP	147.32.84.229	78.240.61.78	INT	0	1	60	1
07.780	0.001	TCP	147.32.85.26	147.32.3.93	A	0	3	197	1
07.781	0.000	TCP	147.32.85.26	213.180.53.50	FPA	0	1	62	1
07.781	0.000	TCP	83.200.200.38	147.32.84.59	FPA	0	2	137	1

(a) Describe briefly the steps required to transform such NetFlow data to make it suitable for leaning discrete sequential models such as n-grams in order to fingerprint the different hosts in this network. (5 pt)

(b) Give the 3-gram model for the sequence of Protocol values (over the whole network, not per host), with probabilities, without smoothing. (5 pt)

(c) Explain the difference between behavioral fingerprinting and behavioral profiling using the feature values that occur for hosts (Src Ip Addr) 147.32.85.26 and 83.200.200.38. (5 pt)

(d) Another method for profiling network traffic uses state machines. Give two advantages and two disadvantages of using state-machines compared to using n-grams. (5 pt)

Question 5 - privacy

(a) Two methods for privacy aware data publishing are rank swapping and fixed size micro-aggregation. Compare these two methods in terms of privacy guarantees and computational complexities. (5 pt)

Given the following data sample:

Attribute 1	Attribute 2	Attribute 3	Attribute 4
4.6	3.6	1.0	0.2
4.8	3.4	1.9	0.2
5.1	3.7	1.5	0.4
5.5	2.3	4.0	1.3
5.8	2.7	5.1	1.9
6.4	3.2	4.5	1.5
6.3	2.9	5.6	1.8
7.0	3.2	4.7	1.4

(b) Show a result (sample) of applying rank swapping to this data with parameter $p = 2$ (do not reswap already swapped values). (5pt)

(c) Show a result (sample) of performing fixed size micro-aggregation with value $k = 3$. (5pt)