ISO/IEC TS 4213:2022-10 (E)

Information technology - Artificial intelligence - Assessment of machine learning classification performance

Contents

Page

Foreword v

Introduction vi

1 Scope 1

2 Normative references 1

3 Terms and definitions 1
3.1 Classification and related terms 1
3.2 Metrics and related terms 1

4 Abbreviated terms 3

5 General principles 4
5.1 Generalized process for machine learning classification performance assessment 4
5.2 Purpose of machine learning classification performance assessment 4
5.3 Control criteria in machine learning classification performance assessment 5
5.3.1 General 5
5.3.2 Data representativeness and bias 5
5.3.3 Preprocessing 5
5.3.4 Training data 5
5.3.5 Test and validation data 6
5.3.6 Cross-validation 6
5.3.7 Limiting information leakage 6
5.3.8 Limiting channel effects 6
5.3.9 Ground truth 7
5.3.10 Machine learning algorithms, hyperparameters and parameters 7
5.3.11 Evaluation environment 8
5.3.12 Acceleration 8
5.3.13 Appropriate baselines 8
5.3.14 Machine learning classification performance context 8

6 Statistical measures of performance 8
6.1 General 8
6.2 Base elements for metric computation 9
6.2.1 General 9
6.2.2 Confusion matrix 9
6.2.3 Accuracy 9
6.2.4 Precision, recall and specificity 9
6.2.5 F1 score 9
6.2.6 F 9
6.2.7 Kullback-Leibler divergence 10
6.3 Binary classification 10
6.3.1 General 10
6.3.2 Confusion matrix for binary classification 11
6.3.3 Accuracy for binary classification 11
6.3.4 Precision, recall, specificity, F1 score and F for binary classification 11
6.3.5 Kullback-Leibler divergence for binary classification 11
6.3.6 Receiver operating characteristic curve and area under the receiver operating characteristic curve 11

6.3.7 Precision recall curve and area under the precision recall curve 11
6.3.8 Cumulative response curve 12
6.3.9 Lift curve 12
6.4 Multi-class classification 12
6.4.1 General 12
6.4.2 Accuracy for multi-class classification 12
6.4.3 Macro-average, weighted-average and micro-average 12
6.4.4 Distribution difference or distance metrics 13
6.5 Multi-label classification 14
6.5.1 General 14
6.5.2 Hamming loss 14
6.5.3 Exact match ratio 15
6.5.4 Jaccard index 15
6.5.5 Distribution difference or distance metrics 15
6.6 Computational complexity 16
6.6.1 General 16
6.6.2 Classification latency 16
6.6.3 Classification throughput 17
6.6.4 Classification efficiency 17
6.6.5 Energy consumption 17

7 Statistical tests of significance 18
7.1 General 18
7.2 Paired Student's t-test 18
7.3 Analysis of variance 19
7.4 Kruskal-Wallis test 19
7.5 Chi-squared test 19
7.6 Wilcoxon signed-ranks test 19
7.7 Fisher's exact test 19
7.8 Central limit theorem 20
7.9 McNemar test 20
7.10 Accommodating multiple comparisons 20
7.10.1 General 20
7.10.2 Bonferroni correction 20
7.10.3 False discovery rate 21

8 Reporting 21

Annex A (informative) Multi-class classification performance illustration 22

Annex B (informative) Illustration of ROC curve derived from classification results 24

Annex C (informative) Summary information on machine learning classification benchmark tests 29

Annex D (informative) Chance-corrected cause-specific mortality fraction 31

Bibliography 32