Developing a TS on assessment of ML classification performance

Michael Thieme
Accenture Federal Services

Lingzhong Meng
Chinese Academy of Sciences
Why does “Machine Learning Classification Performance Assessment” matter?

Problem Statement
• To ensure fairness and reduce bias in ML systems, consistent approaches and methods can be applied to ML classification performance assessment
• We need to be able to compare results across different evaluation regimes

Approach
• SC 42 developed a Technical Specification - ISO/IEC 4213 - that provides methodologies for binary, multi-class, and multi-label classification use cases
• Anticipated publication in Q2 2022
• Builds on foundational concepts in ISO/IEC 22989
### ML classification types addressed in TS 4213

<table>
<thead>
<tr>
<th>Binary</th>
<th>Multi-Class</th>
<th>Multi-Label</th>
</tr>
</thead>
</table>
| • Each sample labelled as one of two mutually exclusive classes  
• Often positive or negative with reference to a categorization | • Each sample labelled as one of three or more mutually exclusive classes  
*Example – ML classification software learns to categorize images as “dog”, “cat” or “other” based on labels assigned by a human reviewer* | • Each sample can be labelled as one or more classes – labels not mutually exclusive  
• Sample can have multiple labels  
• ML model can predict some labels correctly, but fail to predict other correct labels  
• ML model can predict some labels correctly, but predict others incorrectly  
*Example - ML classification software learns to categorize text as one or more of opinion, news, hostile, misinformation or disinformation based on labels assigned by a human reviewer* |

*Example – ML classification software learns to mark email as “spam” or “not spam” based on feedback from email recipient*
How can you use TS 4213?

Specifies methodologies for measuring classification performance of machine learning models, systems and algorithms

Questions that TS 4213 helps answer

- How “good” is the model?
- How reliable are its predictions?
- What is the expected frequency and size of errors?
- What is the best-performing model out of N alternatives?
- Does the model perform well over time with noisy or new production data

Evaluation concepts that TS 4213 includes

- Data representativeness and bias
- Pre-processing
- Training data
- Test and validation data
- Cross-validation
- Limiting information leakage
- Limiting channel effects
- Ground truth
- ML algorithms, hyperparameters and parameters
- Evaluation environment
- Acceleration
- Appropriate baselines
- Putting performance in context
Informative Annexes

Annex A can help us better use TS 4213

Annex A
Multi-class classification performance illustration

Annex B
Illustration of ROC curve derived from classification results

Annex C
Summary information on ML classification benchmark tests

Annex D
Chance-corrected cause-specific mortality fraction
Thank you

Michael Thieme
Accenture Federal Services
michael.thieme@afs.com

Lingzhong Meng
Chinese Academy of Sciences
lingzhong@iscas.ac.cn