UNIVERSITY COMPUTER SCIENCE • EXERCISE SHEET

13.6 Evaluation and leakage

Name: _____ Class: _____ Date: _____ **Total: 24 marks**

KEY WORDS confusion matrix • precision • recall

Objective

Derive every metric from the matrix and design a leakage-proof split.

1 Key knowledge and worked examples

■ **The matrix**

Test 200: TP 36, FP 10, FN 6, TN 148 – accuracy 0.92, precision 0.783, recall 0.857, F1 0.818; base rate 21% ✓.

■ **Worked example: the folds**

5-fold CV: 0.905, 0.925, 0.915, 0.900, 0.935 – mean 0.916, sd 0.013 ✓.

2 Practice

2.1 What does spending the test set mean? [3]

2.2 Give the accuracy of the trivial always-on-time predictor. [3]

2.3 Name the three plumbing leakages this unit lists. [3]

3 Exam-style questions

3.1 A competitor reports: TP 40, FP 20, FN 2, TN 138 on the same 200-record test set.

Compute accuracy, precision, recall, and F1 for the competitor. Compare with the course model on each metric and state which model you would deploy for catching lates, with the reason. State what additional resampling evidence you would demand before trusting either. [8]

3.2 Design the leakage defence for a pipeline that scales features, imputes missing values, and selects features.

State the wrong order (all three before splitting) and what leaks at each step. Give the correct per-fold order. Explain the midnight status-code example in this vocabulary. State why fold metrics are averaged over per-fold counts, not over fold accuracies alone when folds are unequal. [7]

Solutions

2.1

- Using it for model choice: it becomes a second validation set, and its score is then optimistic ✓.

2.2

- Predicting every parcel on time: $158/200 = 0.79$ ✓.

2.3

- Scaling before splitting; imputing with test statistics; feature selection over the full data (and the label-encoding feature itself) ✓.

3.1

- Competitor: accuracy $(40 + 138)/200 = 0.89$ ✓.
- Precision $40/60 = 0.667$; recall $40/42 = 0.952$; F1 $= 2\times0.667\times0.952/(0.667+0.952) = 0.784$ ✓.
- Comparison: competitor wins recall (0.952 vs 0.857), loses precision (0.667 vs 0.783) and accuracy (0.89 vs 0.92) ✓.
- For catching lates (refunds depend on not missing them): the competitor – recall is the operative metric ✓.
- Demand: 5-fold CV with per-fold counts, and the same folds for both models ✓.

3.2

- Wrong order: fit scaler, imputer, and feature selector on ALL data before splitting – the test set's mean/variance, its missing-value distribution, and its label-feature correlations all leak into training ✓.
- Correct per-fold: split first; fit scaler, imputer, selector on the training fold only; transform the validation fold with those fitted objects ✓.
- The status code: a feature whose provenance is post-outcome – the data card's provenance column catches it before any split ✓.
- Unequal folds: fold accuracies weight small folds equally – pool TP/FP/FN/TN across folds and compute metrics on the pooled counts (or weight by fold size) ✓.

exercises.lol exercises.com.cn