COMP2200/COMP6200 Lecture 5a – k-Nearest Neighbours and Model Evaluation

Greg Baker

25th August 2025

Agenda

- The Red Rooster line
- Train/validation/test split
- k-NN vs logistic regression in Orange
- Classification metrics
- Cross-validation and tuning k

The Red Rooster line

- Red Rooster is an Australian roast chicken chain
- Urban myth: almost no stores appear east of a line through Sydney
- Data: locations of fast-food chains across NSW
- Goal: predict whether a store is Red Rooster based on location
- Illustrates classification and geographic bias

Modelling setup

- Features: latitude and longitude of each store
- Class label: `chain` (Red Rooster vs others)
- Compare logistic regression and k-NN
- Need separate training, validation and test data to tune and evaluate models

Prediction time

- Which algorithm will classify locations better: logistic regression or k-NN?
- Discuss with a neighbour and jot down your hypothesis

Build the workflow in Orange (1/2)

- Load `nsw-fast-food-chains-geocoded.csv`
- Inspect data; expect about 900 rows
- Select `chain` as the target
- Keep only latitude and longitude as features
- **Checkpoint:** confirm Orange shows the expected row count

Build the workflow in Orange (2/2)

- Add kNN and Logistic Regression learners
- Connect both to Test & Score and Confusion Matrix
- Run Test & Score to view accuracy, precision, recall and F1
- **Checkpoint:** metrics not appearing? Re-check your connections

Reflect on results

- Which model scored higher on accuracy?
- Did the outcome match your earlier prediction?
- If not, what might explain the difference?

Cross-validation

- Test & Score widget runs k-fold cross-validation on the training data
- Use validation results to select between logistic regression and k-NN
- Vary k or use **Optimize Parameters** for grid search
- Hold out a final test set for unbiased accuracy estimates