Model Evaluation

Why Evaluation Matters

- Evaluation determines which ML models perform best in practice.
- Training accuracy is **not** a reliable indicator of future performance.
- Limited data makes evaluation harder; careful techniques required.

Training vs Testing

- Training error = *resubstitution error* (optimistically biased).
- Test error = performance on **unseen** data.
- Three data roles:
    - **Training set**: builds the model.
    - **Validation set**: tunes hyperparameters.
    - **Test set**: estimates future error.

Holdout Method

- Split data into training and testing sets.
- Often $2/3$ for training, $1/3$ for testing.
- Stratification ensures class proportions remain consistent.

Confidence Intervals

- Accuracy measured on a test set is an *estimate*.
- Model test accuracy approximates a Bernoulli process.
- Confidence intervals derived using normal approximation.
- Useful for quantifying uncertainty of error estimates.

Cross-Validation (CV)

- **k-fold CV**: Divide data into k folds; each fold used once for testing.
- **10-fold CV** is standard.
- **Repeated CV**: increases reliability.

Leave-One-Out & Bootstrap

- **Leave-One-Out CV**: n-fold CV (n = number of instances).
- Uses maximum training data; computationally heavy.
- **0.632 Bootstrap**: samples with replacement.
- Training set ~63.2% unique instances; remaining used as test.

Comparing Models

- Use statistical significance tests.
- **Paired t-test** evaluates whether two algorithms differ reliably.
- Must consider dependence between datasets when using repeated sampling.

Predicting Probabilities

Two loss functions:

- **Quadratic loss**: penalizes probability distribution “shape”.
- **Information loss**: -log2(p(correct)). Strongly punishes zero probabilities.

Cost-Sensitive Evaluation

- Different errors have different costs.
- Confusion matrix elements: TP, FP, FN, TN.
- Techniques:
  - Adjust decision thresholds.
  - Weight instances.
  - Use cost matrices.

Reciever Operating Characteristic (ROC) Curves

- Plot True Positive Rate vs False Positive Rate.
- Independent of class balance and cost.
- Area Under Curve (AUC) summarizes performance.

Recall–Precision Curves

- Useful when positive class is rare.
- Precision = relevance of retrieved items.
- Recall = proportion of relevant items retrieved.

Cost Curves

- Show expected cost as class distribution varies.
- Straight-line representation for each classifier.
- Help determine which classifier is optimal for given distributions.

Numeric Prediction Evaluation

Metrics:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Relative Absolute Error (RAE)
- Correlation Coefficient

Minimum Description Length (MDL)

- Prefers simpler models plus cost of encoding errors.
- Equivalent to maximizing posterior probability.
- Avoids overfitting by penalizing complexity.

MDL for Clustering

- Choose clustering that compresses data best.
- Encode cluster centers and member deviations.
- Good clusters reduce description length.

Summary

- Training error not equal to test error.
- Use cross-validation for robust estimates.
- Costs, probabilities, and tradeoffs matter.
- ROC, lift, and precision-recall curves visualize tradeoffs.
- MDL offers principled model selection.