Wrap up

Final Lecture II

Nayel Bettache

Cross-Validation: Principles

Goal:

- Estimate a model's ability to generalize to unseen data.
- Helps select models, tune hyperparameters, and assess prediction accuracy.

Addresses:

- Overfitting: Avoid models that fit training data too closely.
- Model selection: Compare and select among candidate models.
- Performance estimation: Estimate out-of-sample error.

Basic Cross-Validation Procedure:

1. Split data into training and validation sets.
2. Train the model on the training set.
3. Evaluate model performance on the validation set.

Common Evaluation Metrics:

- Regression: MSE, RMSE, MAE, R^2
- Classification: Accuracy, AUC, F1-score, Log-loss

Cross-Validation Methods

Holdout Method:

- Single split (e.g., 70% training, 30% test).
- Simple but performance estimate has high variance.

k-Fold Cross-Validation:

1. Split data into k roughly equal parts (folds).
2. For each fold:
   - Train model on the $k-1$ other folds.
   - Validate on the held-out fold.
3. Average the performance across all k runs.

- Lower variance than holdout; good bias-variance trade-off.

Leave-One-Out Cross-Validation (LOOCV):

- Special case of k-fold with $k = n$.
- Each data point serves as its own validation set.
- Very low bias, but computationally expensive and high variance.

Repeated k-Fold:

- Run k-fold cross-validation multiple times with different random partitions.

Applications and Considerations

Hyperparameter Tuning: Nested Cross-Validation

- Avoids data leakage during model selection.
- **Inner loop:** Tune hyperparameters (e.g., regularization strength, tree depth).
- **Outer loop:** Estimate generalization performance.

Choosing k in k-Fold CV: Bias-Variance Trade-Off

- Higher k (e.g., LOOCV): Lower bias, higher variance, more computational cost.
- Lower k (e.g., 5-fold): Higher bias, lower variance, faster computation.

Data-Aware Considerations:

- **Stratified Sampling:**
  - Maintain class proportions in each fold.
  - Especially important in imbalanced classification problems.
- **Time Series Data:**
  - Use time-aware cross-validation (e.g., rolling windows, forward chaining).
  - Avoid random shuffling to prevent look-ahead bias.

Model Selection: Fundamentals

Goal:

- Choose the model that best balances prediction accuracy, generalization to new data, and interpretability.

Bias-Variance Trade-Off:

- **Simple models:** High bias (may miss important patterns), low variance (stable estimates).
- **Complex models:** Low bias (fit training data well), high variance (may overfit).

Parsimony Principle:

- Prefer simpler models if their performance is similar.
- Example: Linear model may be preferred over a neural network if both have similar prediction error.

Overfitting vs. Underfitting:

- **Overfitting:** Model too complex — fits noise in training data, poor generalization.
- **Underfitting:** Model too simple — misses patterns in data, poor fit.

Information Criteria for Model Selection

Goal: Balance model fit and complexity using penalized criteria.

Akaike Information Criterion (AIC): $\text{AIC} = -2\log(L) + 2p$

- L: Maximum likelihood of the model
- p: Number of estimated parameters
- Prioritizes predictive accuracy; less strict penalty

Bayesian Information Criterion (BIC): $\text{BIC} = -2\log(L) + p\log(n)$

- n: Number of observations
- Favors simpler models; stronger penalty on complexity
- Often preferred for model identification

Adjusted R^2: $R^2_{\text{adj}} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$

- Penalizes R^2 for model complexity
- Useful for comparing nested linear models

Interpretation:

- Lower AIC or BIC indicates a preferred model (relative comparison only).
- Higher adjusted R^2 suggests better balance of fit and complexity.
- BIC penalizes complexity more than AIC.

Validation-Based Methods for Model Selection

Cross-Validation Error: $\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{Error}_i$

- Average validation error over k folds.
- Lower CV error indicates better expected generalization performance.

Prediction Error Metrics:

- **Regression (continuous targets):** $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- **Classification (categorical targets):** Misclassification rate $= \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ where $I(\cdot)$ is the indicator function.

One-Standard-Error Rule:

- Among all candidate models, identify the model with the lowest CV error.
- Choose the simplest model whose CV error is within one standard error of the minimum.
- Helps prevent overfitting while preserving near-optimal performance.

Variable Selection Methods

Goal: Identify a subset of predictors that improves model interpretability without sacrificing predictive performance.

Subset Selection Methods:

- **Best Subset Selection:**
  - Try all 2^p possible combinations of predictors.
  - Computationally intensive; feasible only for small p.
- **Forward Selection:**
  - Start with no variables.
  - Add variables one at a time based on performance improvement.
- **Backward Elimination:**
  - Start with all variables.
  - Remove least useful variable at each step.
- **Stepwise Selection:**
  - Combines forward and backward steps to refine selection.

Regularization

- **Ridge Regression (ℓ_2 penalty):**

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Shrinks coefficients toward zero; helps with multicollinearity, but does not perform variable selection.
- **Lasso (ℓ_1 penalty):**

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Encourages sparsity; some coefficients exactly zero.
- **Elastic Net (Combined $\ell_1 + \ell_2$):**

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$$

- Useful when predictors are highly correlated or $p > n$.

Regularization and Modern Approaches

Regularization Path:

- Sequence of models obtained by varying the regularization parameter λ.
- Shows how coefficients evolve from full model to sparse model as λ increases.

Tuning Regularization via Cross-Validation:

- Select optimal λ by minimizing validation error.
- **Grid search:** Evaluate model on a fixed grid of λ values.
- **Random search:** Sample λ values randomly; more efficient in high dimensions.

Ensemble Methods: Combine multiple models to improve performance

- **Bagging (Bootstrap Aggregating):**
  - Reduces variance by averaging models trained on bootstrap samples.
  - Example: Random Forests.
- **Boosting:**
  - Sequentially corrects errors of previous models to reduce bias.
  - Example: Gradient Boosting Machines (GBM), XGBoost.

Linearity Assumption: Detailed View

- **Mathematical expression**: $E(Y|X) = X\beta$
- **Detection methods**:
  - Residual vs. fitted plots
  - Residual vs. individual predictor plots
- **Remedies for nonlinearity**:
  - Transform variables (log, square, square root)
  - Add polynomial terms
  - Use splines or local regression
  - Consider generalized additive models (GAMs)

Independence and Constant Variance: Detailed View

- **Independence**:
  - Mathematical expression: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
  - Detection: Durbin-Watson test, ACF/PACF plots
  - Remedies: Time series models, mixed-effects models
- **Homoscedasticity**:
  - Mathematical expression: $\text{Var}(\varepsilon_i) = \sigma^2$ for all i
  - Detection: Breusch-Pagan test, residual plots, scale-location plots
  - Remedies: Transformation, weighted least squares, robust standard errors

Coefficient Estimation and Properties

- **OLS estimator**: $\hat{\beta} = (X^T X)^{-1} X^T y$
- **Properties**:
  - Unbiased: $E(\hat{\beta}) = \beta$
  - Variance: $\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$
  - BLUE (Best Linear Unbiased Estimator)
  - Normally distributed if errors are normal
- **Covariance matrix**: $\hat{\sigma}^2 (X^T X)^{-1}$ where $\hat{\sigma}^2 = \frac{RSS}{n-p-1}$

Model Evaluation in Multiple Regression

- R^2 **and adjusted** R^2:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} \tag{1}$$

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \tag{2}$$

- F**-statistic**:

$$F = \frac{ESS/p}{RSS/(n-p-1)} \tag{3}$$

$$= \frac{R^2/p}{(1-R^2)/(n-p-1)} \tag{4}$$

Multicollinearity

- **Perfect multicollinearity**: Linear dependence among predictors
- **Near multicollinearity**: High correlation among predictors
- **Consequences**:
  - Inflated coefficient variances
  - Unstable coefficient estimates
  - Reduced power of tests
- **Detection**:
  - Correlation matrix
  - Variance Inflation Factor (VIF): $\text{VIF}_j = \frac{1}{1-R_j^2}$
  - Condition number of $X^T X$
- **Remedies**:
  - Remove redundant predictors
  - Use principal components
  - Ridge regression
  - Centered variables for interactions

Sampling Distribution of Regression Coefficients

- **Under classical assumptions**:

$$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1}) \tag{5}$$

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-p-1} \tag{6}$$

- **For nonlinear functions of parameters**:
  - Delta method: Uses first-order Taylor expansion
  - Bootstrap: Empirical sampling distribution

Weighted Least Squares: Theory

- **Model**: $y = X\beta + \varepsilon$, where $\text{Var}(\varepsilon) = \sigma^2\Omega$
- **WLS estimator**:

$$\hat{\beta}_{WLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y \tag{7}$$

- **Properties**:
  - Unbiased: $E(\hat{\beta}_{WLS}) = \beta$
  - Variance: $\text{Var}(\hat{\beta}_{WLS}) = \sigma^2(X^T\Omega^{-1}X)^{-1}$
  - BLUE among all linear unbiased estimators
- **Challenge**: Estimating Ω in practice

Heteroskedasticity-Consistent Standard Errors

- **White's estimator (HC0)**:

$$\hat{\text{Var}}(\hat{\beta}) = (X^T X)^{-1} X^T \text{diag}(e_i^2) X (X^T X)^{-1} \tag{8}$$

- **Improved estimators**:
  - HC1: Correction for degrees of freedom
  - HC2: Leverage adjustment
  - HC3: Further leverage adjustment
  - HC4, HC5: For high leverage points
- **HAC (Heteroskedasticity and Autocorrelation Consistent)**:
  - Newey-West estimator
  - Accounts for both heteroskedasticity and autocorrelation

Final Review: Connecting Theory to Practice

- Statistical inference is built on a foundation of assumptions
- Model diagnosis is as important as model building
- When assumptions are violated, alternatives exist:
  - Robust methods
  - Resampling approaches
  - Transformation techniques
- Model selection should consider:
  - Statistical criteria (AIC, BIC, CV error)
  - Domain knowledge
  - Interpretability
  - Practical constraints
- Remember that statistical models serve to approximate reality

Exam Preparation Tips

- Focus on understanding concepts, not just formulas
- Practice applying methods to real data
- Be able to interpret output from statistical software
- Understand when each method is appropriate
- Be prepared to diagnose problems and recommend solutions
- Connect topics across the course:
  - How do regression assumptions affect inference?
  - How does cross-validation improve model selection?
  - How do bootstrap methods complement classical inference?