Machine Learning for Omics Integration - Day 1 Notes

Course Notes

December 17, 2024

Contents

1 Key Concepts and Principles

1.1 Sample Size Considerations

- **General Recommendation**: Sample size should be more than the number of features, ideally 10x larger
- **Data-Model Trade-off**: The less data you have, the more modeling effort you need to invest
- **Context-Dependent**: Sample size adequacy varies by application
  - Some cases: 100 samples may be sufficient
  - Other cases: Millions of samples may still be inadequate

1.2 Mathematical Framework

1.2.1 Linear Model Representation

$$Y = \alpha + \beta X$$

Where: - **Y** = Phenotype variable (outcome/response variable) - **X** = Omics data (e.g., gene expression, protein levels, metabolite concentrations) - (alpha) = Intercept term - (beta) = Slope coefficient/effect size

1.2.2 Variable Definitions

- **Phenotype variable (Y)**: The biological trait or clinical outcome being studied
- **Omics data (X)**: High-dimensional molecular measurements
  - Examples: Gene expression profiles, protein abundance, metabolomics data

2 Important Notes

- Quality and relevance of data often matter more than sheer quantity
- Model complexity should be balanced with available sample size
- Feature selection becomes crucial when dealing with high-dimensional omics data

3 Statistical Approaches

3.1 Frequentist Statistics

Core Principle: Based on maximum likelihood estimation (MLE)

3.1.1 Key Characteristics:

- **Summary Statistics Focus**: Emphasizes point estimates and confidence intervals
- **P-value Emphasis**: Heavy reliance on p-values for hypothesis testing
- **Limitations in Omics**:
  - Over-focus on p-values can be problematic with high-dimensional data
  - Multiple testing corrections become critical
  - May not capture uncertainty effectively in complex models

3.1.2 Frequentist vs. Other Approaches:

- **Advantages**: Well-established, computationally efficient for simple models
- **Challenges**: Less flexibility for incorporating prior knowledge
- **In Omics Context**: Can struggle with high-dimensional, low-sample-size scenarios

3.2 Bayesian Statistics

Core Principle: Incorporates prior knowledge and quantifies uncertainty through probability distributions

3.2.1 Bayes' Rule (Mathematical Foundation)

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Where: - **P(H|E)** = Posterior probability (probability of hypothesis H given evidence E) - **P(E|H)** = Likelihood (probability of evidence E given hypothesis H) - **P(H)** = Prior probability (initial belief about hypothesis H) - **P(E)** = Marginal probability (total probability of evidence E)

3.2.2 Bayesian vs. Frequentist in Omics Context

Aspect	Bayesian	Frequentist
Prior Knowledge	Incorporates biological prior information	Ignores prior knowledge
Uncertainty	Full probability distributions	Point estimates + confidence intervals
Multiple Testing	Natural shrinkage via hierarchical priors	Requires explicit corrections (FDR, Bonferroni)
Computational Cost	Higher (MCMC, sampling methods)	Lower (closed-form solutions)
Interpretation	Intuitive probability statements	Abstract long-run frequency

3.2.3 Applications in Omics

- **Regularization**: Bayesian priors naturally provide regularization (e.g., Bayesian LASSO)
- **Hierarchical Modeling**: Accounts for multiple levels of biological organization
- **Uncertainty Quantification**: Essential for clinical decision-making with omics data

4 Missing Heritability Problem

4.1 Definition and Context

Missing heritability refers to the gap between heritability estimates from family studies and the variance explained by identified genetic variants in genome-wide association studies (GWAS).

4.1.1 The Heritability Gap

- **Family-based heritability**: Often 40-80% for complex traits (estimated from twin/family studies)
- **GWAS-explained variance**: Typically only 5-15% of phenotypic variance
- **Missing component**: The unexplained 60-70% represents "missing heritability"

4.1.2 Examples

1. **Height**: Family studies suggest ~80% heritability, but known variants explain only ~45%
2. **Type 2 Diabetes**: Family risk is substantial, but identified variants explain <10% of risk
3. **Schizophrenia**: High familial aggregation (~80% heritability) vs. ~30% explained by common variants

4.2 Proposed Explanations

4.2.1 1. Rare Variants with Large Effects

- **Hypothesis**: Many rare variants (MAF $< 1\%$) have large effect sizes
- **Detection Challenge**: GWAS underpowered for rare variants
- **Solution**: Whole-genome sequencing in large cohorts

4.2.2 2. Structural Variants

- **Copy Number Variants (CNVs)**: Large insertions/deletions not well-captured by SNP arrays
- **Inversions and Translocations**: Complex rearrangements missed by standard GWAS

4.2.3 3. Epistatic Interactions

- **Gene $\times$ Gene interactions**: Non-additive effects between loci
- **Statistical Challenge**: Requires very large sample sizes to detect
- **Example**: Variants may only be pathogenic in specific genetic backgrounds

4.2.4 4. Gene × Environment Interactions

- **Phenotypic expression**: Genetic effects may depend on environmental context
- **Examples**: Diet-gene interactions in metabolic traits, stress-gene interactions in psychiatric disorders

4.2.5 5. Epigenetic Factors

- **DNA methylation**: Heritable but not captured by genetic sequence
- **Histone modifications**: Can be inherited across generations
- **Challenge**: Tissue-specific and environmentally responsive

4.3 Multi-Omics Solutions

4.3.1 Integrative Approaches

- **Genomics + Transcriptomics**: Expression QTL (eQTL) analysis
- **Genomics + Metabolomics**: Metabolite QTL (mQTL) studies
- **Multi-tissue analysis**: GTEx-style approaches across tissues

4.3.2 Advantages of Multi-Omics for Missing Heritability

1. **Functional annotation**: Links genetic variants to molecular phenotypes
2. **Pathway analysis**: Identifies biological mechanisms
3. **Tissue specificity**: Captures context-dependent genetic effects
4. **Regulatory elements**: Identifies non-coding variant effects

5 LASSO Regression and Regularization

5.1 LASSO Mathematical Formulation

5.1.1 Standard Linear Regression (OLS)

$$\min_{\beta} ||Y - X\beta||_2^2$$

5.1.2 LASSO Objective Function

$$\min_{\beta} ||Y - X\beta||_2^2 + \lambda||\beta||_1$$

Where: - **||Y - X ||^2__2** = Residual Sum of Squares (RSS) - = Regularization parameter (penalty strength) - **|| ||__1** = L1 penalty (sum of absolute values of coefficients)

5.1.3 L1 vs L2 Penalties

Penalty Type	Mathematical Form	Effect	Use Case
L1 (LASSO)	$\lambda\Sigma\lvert\beta_j\rvert$	**Feature Selection -** drives coefficients to exactly zero	Sparse solutions, interpretable models
L2 (Ridge)	$\lambda\Sigma\beta_j^2$	**Shrinkage -** shrinks coefficients toward zero	When all features potentially relevant

5.2 LASSO in High-Dimensional Omics

5.2.1 Why LASSO for Omics Data?

1. **p » n problem**: More features than samples
2. **Sparsity assumption**: Only subset of genes/proteins truly associated with phenotype
3. **Multicollinearity**: Omics features often highly correlated
4. **Interpretability**: Need to identify specific biomarkers

5.2.2 Elastic Net Extension

$$\min_{\beta} ||Y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2$$

Combines advantages: - L1 penalty: Feature selection **- L2 penalty**: Handles correlated features better than LASSO alone

5.3 Cross-Validation and Model Selection

5.3.1 Parameter Selection

Critical Principle: **must be selected using cross-validation BEFORE any model training**

5.3.2 Proper Cross-Validation Workflow

```
1. Split data into training and test sets
2. Within training set only:
   a. Perform k-fold cross-validation
   b. For each fold:
      - Fit LASSO with different  values
      - Evaluate performance on validation fold
   c. Select optimal  * with minimum CV error
3. Train final model on full training set using  *
4. Evaluate final performance on held-out test set
```

5.3.3 Common Mistakes to Avoid

Wrong: Select on full dataset, then evaluate performance **Correct**: Select using only training data via cross-validation

Wrong: Use same data for feature selection and performance evaluation
Correct: Separate feature selection from final model validation

5.3.4 Cross-Validation Metrics for Omics

- **Regression tasks**: MSE, MAE, R^2
- **Classification tasks**: AUC, Accuracy, F1-score
- **Multi-class**: Balanced accuracy, macro-averaged metrics

6 Multi-Omics Integration Strategies

6.1 Types of Integration Approaches

6.1.1 1. Early Integration (Data-level fusion)

- **Concept**: Concatenate different omics datasets into single feature matrix
- **Advantages**: Simple implementation, can use standard ML algorithms
- **Challenges**: Different scales, dimensions, missing data patterns

- **Example**: [Genomics | Transcriptomics | Proteomics] → Combined matrix

6.1.2 2. Intermediate Integration (Feature-level fusion)

- **Concept**: Extract features from each omics layer, then combine features
- **Examples**: Principal components from each omics type
- **Advantages**: Dimensionality reduction, captures layer-specific patterns

6.1.3 3. Late Integration (Decision-level fusion)

- **Concept**: Build separate models for each omics type, combine predictions
- **Advantages**: Accounts for different data characteristics
- **Methods**: Ensemble methods, weighted voting, stacking

6.2 Advanced Integration Methods

6.2.1 Multi-Omics Factor Analysis (MOFA)

- **Principle**: Identifies latent factors explaining variance across omics layers
- **Advantage**: Handles missing data, identifies shared vs. specific factors
- **Applications**: Single-cell multi-omics, cancer studies

6.2.2 DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents)

- **Purpose**: Supervised integration for classification
- **Strength**: Identifies correlated features across omics types
- **Output**: Multi-omics signature for disease prediction

6.2.3 Network-Based Integration

- **Approach**: Model interactions within and between omics layers
- **Examples**: Pathway analysis, protein-protein interaction networks
- **Advantage**: Incorporates biological knowledge

6.3 Validation and Evaluation

6.3.1 Key Principles for Multi-Omics Validation

1. **Biological Validation**
   - Literature support for identified biomarkers
   - Pathway enrichment analysis
   - Functional validation experiments
2. **Statistical Validation**
   - Independent test cohorts
   - Cross-study validation
   - Temporal validation (if longitudinal data available)
3. **Clinical Validation**
   - Association with clinical outcomes
   - Comparison with existing clinical markers
   - Cost-benefit analysis for clinical implementation

6.3.2 Performance Metrics

- **Discrimination**: How well does the model separate classes?
- **Calibration**: Do predicted probabilities match observed frequencies?
- **Clinical Utility**: Does the model improve clinical decision-making?

6.3.3 Common Pitfalls in Multi-Omics

1. **Data leakage**: Information from test set influencing model selection
2. **Overfitting**: Complex models memorizing noise rather than signal
3. **Batch effects**: Technical variation confounding biological signal
4. **Population stratification**: Genetic ancestry effects in association studies

7 Practical Implementation Considerations

7.1 Data Preprocessing

- **Normalization**: Between-sample standardization
- **Scaling**: Between-omics standardization

- **Missing data**: Imputation strategies specific to omics type
- **Batch correction**: Computational methods (ComBat, limma) or experimental design

7.2 Computational Resources

- **Memory requirements**: Increase dramatically with multi-omics
- **Parallel processing**: Essential for large-scale analyses
- **Cloud computing**: Often necessary for population-scale studies

7.3 Reproducibility

- **Version control**: Track analysis code and software versions
- **Containerization**: Docker/Singularity for computational reproducibility
- **Documentation**: Detailed methods for replication