Regularized regression II

Tuning and Cross Validation

Tuning and Cross-Validation

Lasso regression modeling process

1. Choose a series of λ values
2. Estimate a sequence of penalized regression models
   - Since we are interested in the best prediction model for new data...
   - ...this sequence is estimated in a Cross-Validation loop
3. Choose the best λ based on step 2
4. Re-fit model with chosen λ on full training data

$\rightarrow$ Data is split into training and validation set(s) for **model tuning**

Tuning and Cross-Validation

Cross-Validation with the Lasso

1. Split the data into k sets at random
2. Fit a sequence of regularized models using $k-1$ parts of the data
3. Estimate model performances on the holdout set
4. Repeat step 2 & 3 k times

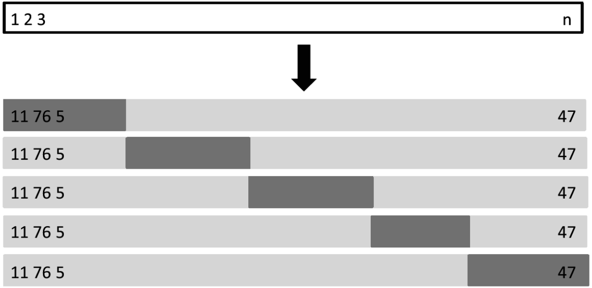
Cross-validated errors (κ indicates data partitions)

$$CV(\hat{f}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_\lambda^{-\kappa(i)}(x_i))$$

with $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ for regression problems.

Tuning and Cross-Validation

Figure: 5-Fold Cross-Validation with training set and validation set (example)



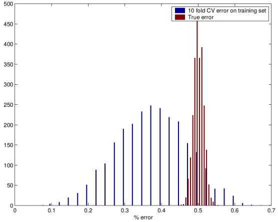
James et al. (2013)

Tuning and Cross-Validation

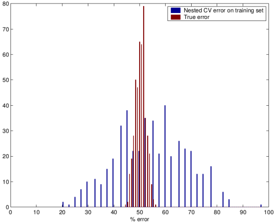
- Repeated Cross-Validation
  1. Run e.g. 10-fold CV five times
  2. Average performance scores over repetitions
  3. Different splits into folds increases robustness
- Nested Cross-Validation
  1. Split data into outer and inner folds
  2. Inner folds: Run CV within inner training fold(s) for tuning
  3. Outer folds: Evaluate best model on the outer test fold(s)
  4. Separates model selection and model assessment

Figure: Bias in CV error (Varma and Simon 2006)

(a) 10-fold CV



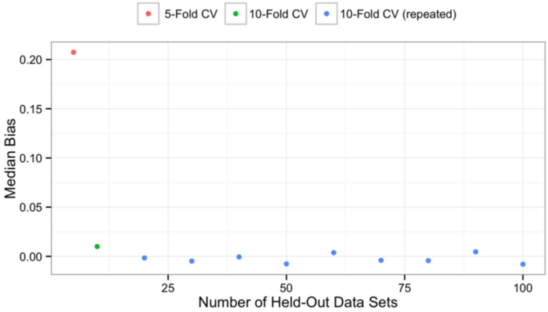
(b) Nested CV



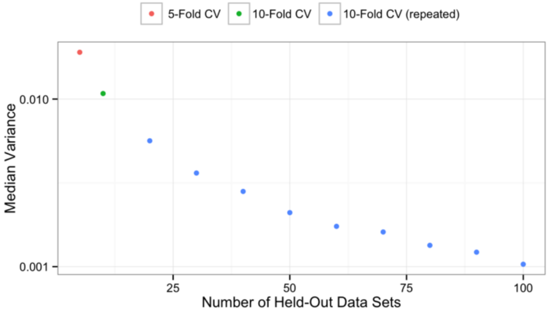
Tuning and Cross-Validation

Figure: "Accuracy" and "precision" of estimating model performance with different types of CV

(a) Bias



(b) Variance



http://appliedpredictivemodeling.com/blog/

Summary

- Ridge, lasso and elastic net penalize complexity
  - Can be used to fit sparse and stable models
  - Typically applied in large p, small n situations
  - Utilize Cross-Validation for parameter tuning
- Statistical inference after feature selection?
  - Selection needs to be taken into account (Taylor & Tibshirani 2015)

Software Resources

Resources for R

- Standard package for ridge regression, lasso and elastic net: `glmnet`
- Group lasso penalization implemented in `grpreg` and `gglasso`
- Tools for post-selection inference: `selectiveInference`

References

Efron, B. and Hastie, T. (2016). Sparse Modeling and the Lasso. In Efron, B. and Hastie, T. (Eds.), *Computer Age Statistical Inference: Algorithms, Evidence and Data Science* (pp. 298–324). New York, NY: Cambridge University Press

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: Chapman & Hall/CRC.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Taylor, J. and Tibshirani, R. (2015). Statistical Learning and Selective Inference. *Proceedings of the National Academy of Sciences, 112*(25), 7629–7634.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics 7*, 91.