Regularized regression I

Stepwise Variable Selection

Stepwise Variable Selection

Algorithm 1: Classical forward selection

Set p-value threshold τ;
Initialize empty model;
repeat
    for *each predictor not in the model* **do**
        Add the predictor to the current model;
        Estimate the statistical significance of the new term;
    end
    if *the smallest p is less than* τ **then**
        Include the corresponding predictor in the model;
    end
until *no significant predictor remains outside the model*;

Stepwise Variable Selection

There are a number of **serious** problems here!

1. Multiple testing issue
2. Objective function does not focus on prediction accuracy
3. Prone to performance evaluation bias

Adjusting stepwise selection approaches

1. Usage of performance measures instead of p-values
2. Implement feature selection in a (proper) **resampling setting**
3. Interweave feature selection in the model-building process

Stepwise Variable Selection

Algorithm 2: Forward selection with resampling

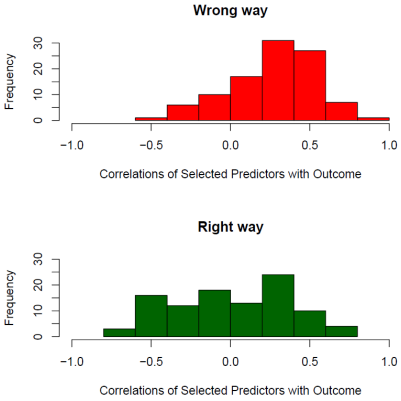
```
Set the number of resampling iterations;
Set the number of features p;
Initialize empty model;
for each resampling iteration do
    Partition data into training and hold-out set;
    for k = 0, ..., p − 1 do
        Consider all p − k models that add an additional predictor to the current model;
        Choose the best among these models in terms of loss in the training data;
        Evaluate the chosen model in the hold-out set;
    end
end
Determine the best number of predictors over all hold-out sets;
```

Stepwise Variable Selection

Cross-Validation done wrong

- **Never** separate feature selection and CV
  - CV *after* selection on full data *biases* performance measures
  - Hold-out samples are no longer independent test sets
- Include feature selection within the CV loop
- Unsupervised screening on full data is valid

Figure: Correlations of y with unrelated x's with incorrect and correct CV



Hastie et al. (2009)