ECE5984 – Applications of Machine Learning
Lecture 17 – Variable Selection

Creed Jones, PhD

Course update

- Project I is due TODAY
  - Don’t forget
  - If your team wishes to make a change for Project II, email me
- HW 4 is posted
  - April 5
- Quiz this Thursday, March 24
  - Lectures 14-17

Today’s Objectives

Variable Selection

- Concept
- Procedures
  - Forward selection
  - Backward selection
  - Stepwise selection
  - Exhaustive (all subsets) selection

VARIABLE SELECTION

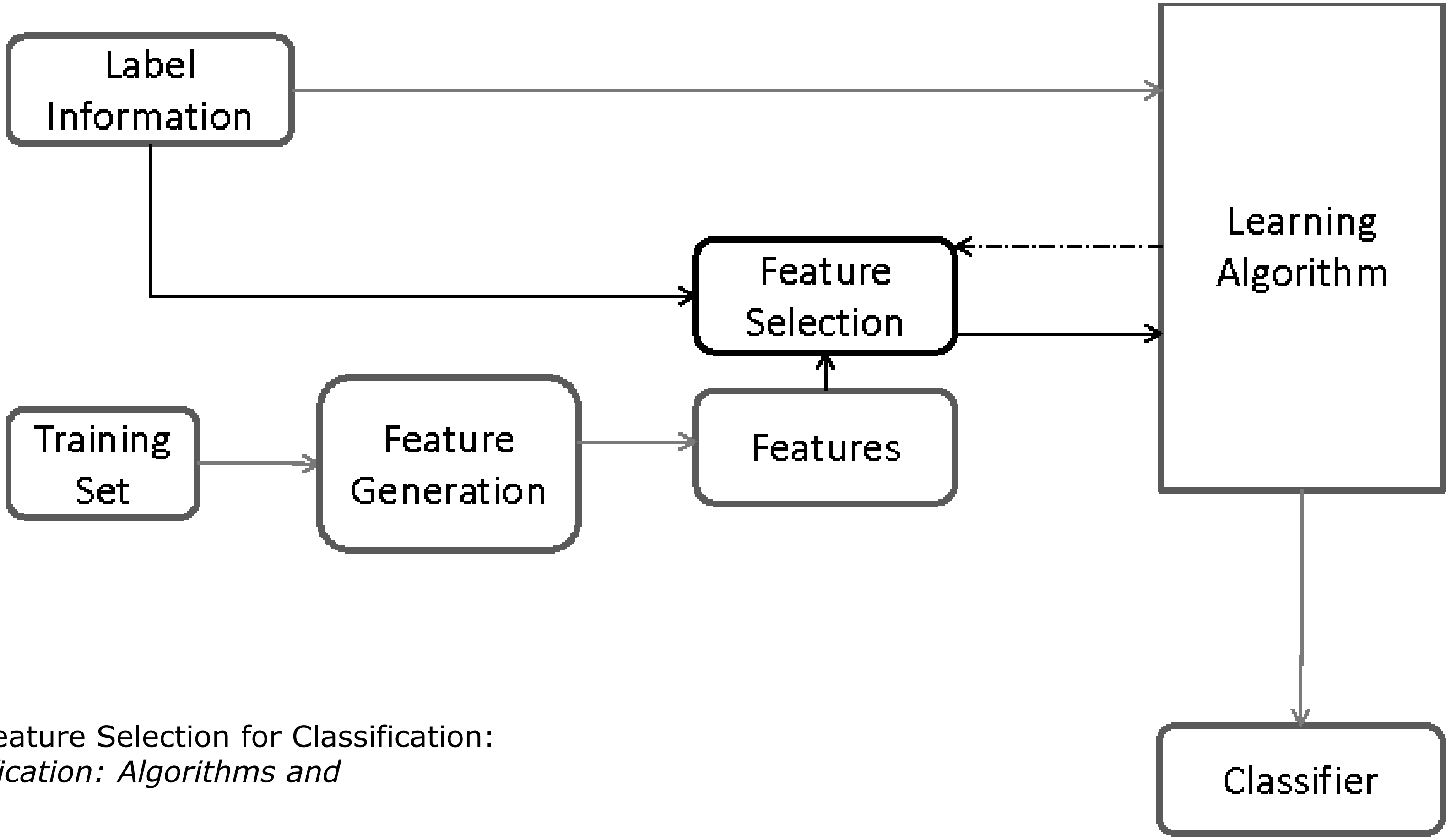
We want to limit the number of variables used in a model – especially a regression model – for several reasons

1. Generalization
   1. Remember Occam's Razor – we want the simplest possible model that will train well, <u>because</u> this model will perform best on new data

2. Comprehension
   1. Simpler models are easier to understand and can give clear insights

3. Efficiency
   1. Less storage required
   2. Faster training time
   3. Sometimes variables cost money to acquire

Variable selection is an iterative process, successively testing performance of modified data sets



Tang, Alelyani and Liu. Feature Selection for Classification: A Review, in *Data Classification: Algorithms and Applications*, 2014

We use a strategy of selecting the variables for our model using statistical measures of their relevance

PROBLEM: Find a set of predictor variables which gives a good fit, predicts the dependent value well and is as small as possible.

There are four popular variable selection methods:

- Forward
- Backward
- Stepwise
- All Subsets

Variable selection processes are usually based on computing the F-test statistic (from ANOVA: Analysis of Variances)

- The F statistic is a measure of the goodness of division of a set into two classes for a particular variable
  - If the variance within the classes is small compared to the variance across classes, then that variable does a good job of discriminating the classes
- The *One-Way ANOVA Test Statistic* for variable d_k is given by:

$$F_{STAT} = \frac{MeanSquaresAcross}{MeanSquaresWithin} = \frac{\sum_{j=1}^{J} n_j \left(\mu_{jk} - \mu_k\right)^2 \Big/ (J-1)}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(d_{ik} - \mu_{jk}\right)^2 \Big/ (N-J)}$$

where N is the dataset size, J is the number of classes for the target, μ_{jk} is the mean of the j^{th} variable in the k^{th} class and μ_k is the overall mean for that variable.

For more info, see section 13.2 in the Illowsky Statistics book posted earlier

Forward selection builds the model from nothing, choosing the best next variable at each step

- Start with a model with no predictors.
- Add variable with <u>largest</u> F statistic (provided p less than some cut-off).
- Refit with this variable. Recompute all F statistics for adding one of the remaining variables and add variable with <u>largest</u> F statistic.
- Continue until no variable is significant at cut-off level.

- Remember, the F statistic is large when the variance across classes is much higher than the variance within classes

Backward selection starts with a *full* model (all variables used), eliminating the least useful variable at each step

- Start with a model with all predictors.
- Delete variable with smallest F statistic (provided p more than some cut-off).
- Refit with this variable deleted. Recompute all F statistics for deleting one of the remaining variables and delete variable with the smallest F statistic.
- Continue until every remaining variable is significant at cut-off level.

Stepwise Selection attempts to add new significant variables to the model, checking each time to see if any variables can now be *excluded*

- Start with model with no predictors.
- Add variable with <u>largest</u> F statistic (provided p less than some cut-off).
- Refit with this variable added. Recompute all F statistics for adding one of the remaining variables and add variable with <u>largest</u> F statistic.
- At each step after adding a variable try to eliminate any variable not significant at some level (that is, do BACKWARD elimination till that stops).
- After doing the backwards steps take another FORWARD step.
- Continue until every remaining variable is significant at cut-off level and every excluded variable is insignificant OR until variable to be added is same as last deleted variable.

"All Subsets" attempts to use all possible (or all reasonably possible) combinations of variables

- For each subset of the set of predictors, fit the model and compute some summary statistic of the quality of the fit.
- Pick model which optimizes this summary statistic.
- With k predictors fit 2k models; impractical for k too large. Special Best subsets algorithms work without looking at all 2k models
- Possible summary statistics:
  - R^2: but NOTE — adding a variable increases R^2 so this is most useful for comparing models of the same size.
  - Adjusted R^2: This method adjusts R^2 to try to compensate for the fact that more variables produces larger R^2 even when the extra variables are irrelevant.
  - Cp : Like Adjusted R^2 but based on a trade off of bias and variance.
  - PRESS: The sum of squares of the PRESS residuals.
- NOTE: I have never used this…

I want to show a few examples of variable selection in action, using t-statistic to include or exclude modeling variables

- These outputs are from SAS, see www.sas.com
- SAS is used by many large organizations for statistical modeling and machine learning

- If you use SAS to develop regression models, you will see many scores, metrics and tests for model performance
  - A course in advanced statistics would be required to understand them all

An example of *Forward Selection*

```
   Forward Selection Proc for Dependent Variable RISK
Step 1 Var CULTURE Entered R-sq=0.3127 C(p)=47.48
            DF  Sum Sq    Mean Sq    F    Prob>F
Regression   1   62.9631   62.9631  50.49 0.0001
Error      111  138.4167    1.2470
Total      112  201.37982301
           Par   Std    Type II
Variable    Est  Error     Sum Sq      F     Prob>F
INTERCEP 3.1979 0.1938  339.6491 272.37 0.0001
CULTURE  0.0733 0.0103   62.9631  50.49 0.0001
-----------------------------------------------------
```

An example of *Forward Selection*

```
  Forward Selection Proc for Dependent Variable RISK
Step 1 Var CULTURE Entered R-sq=0.3127 C(p)=47.48
            DF  Sum Sq   Mean Sq   F    Prob>F
Regression   1   62.9631   62.9631  50.49 0.0001
Error      111  138.4167    1.2470
Total      112  201.37982301
           Par   Std   Type II
Variable    Est  Error    Sum Sq      F    Prob>F
INTERCEP  3.1979 0.1938  339.6491 272.37 0.0001
CULTURE   0.0733 0.0103   62.9631  50.49 0.0001
-----------------------------------------------------
```

```
Step 2 Var STAY Entered R-sq=0.450 C(p)=18.12
            DF  Sum Sq  Mean Sq    F   Prob>F
Regression   2  90.7020   45.3510 45.07 0.0001
Error      110 110.6778    1.0061
Total      112 201.37982301
           Par     Std Type II
Variable    Est      Error   Sum Sq      F Prob>F
INTERCEP  0.80549 0.48776   2.7440  2.73 0.1015
CULTURE   0.05645 0.00980 33.3969 33.19 0.0001
STAY      0.27547 0.05246 27.7388 27.57 0.0001
------------------------------------------------
```

An example of *Forward Selection*

```
  Forward Selection Proc for Dependent Variable RISK
Step 1 Var CULTURE Entered R-sq=0.3127 C(p)=47.48
            DF  Sum Sq   Mean Sq    F    Prob>F
Regression   1   62.9631  62.9631  50.49 0.0001
Error       111  138.4167    1.2470
Total       112  201.37982301
            Par   Std    Type II
Variable    Est  Error    Sum Sq      F     Prob>F
INTERCEP 3.1979 0.1938  339.6491 272.37 0.0001
CULTURE  0.0733 0.0103   62.9631  50.49 0.0001
-----------------------------------------------------
```

```
Step 2 Var STAY Entered R-sq=0.450 C(p)=18.12
            DF Sum Sq Mean Sq   F   Prob>F
Regression   2  90.7020   45.3510 45.07 0.0001
Error       110 110.6778    1.0061
Total       112 201.37982301
           Par      Std Type II
Variable    Est      Error  Sum Sq      F Prob>F
INTERCEP  0.80549 0.48776  2.7440  2.73 0.1015
CULTURE   0.05645 0.00980 33.3969 33.19 0.0001
STAY      0.27547 0.05246 27.7388 27.57 0.0001
------------------------------------------------
```

```
Step 3 Var FACIL Entered R-sq=0.493 C(p)=10.33
            DF Sum of Sq Mean Sq    F   Prob>F
Regression   3   99.3608 33.1203 35.39 0.0001
Error       109  102.0190  0.9360
Total       112  201.3798
            Par   Std      Type II
Variable    Est  Error    Sum Sq     F   Prob>F
INTERCEP 0.4913 0.4816  0.9740   1.04 0.3099
CULTURE  0.0542 0.0095 30.5982 32.69 0.0001
STAY     0.2239 0.0534 16.4766 17.60 0.0001
FACIL    0.0196 0.0065  8.6588   9.25 0.0029
-----------------------------------------------
```

An example of *Forward Selection*

```
  Forward Selection Proc for Dependent Variable RISK
Step 1 Var CULTURE Entered R-sq=0.3127 C(p)=47.48
            DF Sum Sq   Mean Sq   F   Prob>F
Regression   1   62.9631   62.9631  50.49 0.0001
Error       111  138.4167    1.2470
Total       112  201.37982301
           Par   Std   Type II
Variable    Est  Error    Sum Sq      F    Prob>F
INTERCEP  3.1979 0.1938  339.6491 272.37 0.0001
CULTURE   0.0733 0.0103   62.9631  50.49 0.0001
-----------------------------------------------------
```

```
Step 2 Var STAY Entered R-sq=0.450 C(p)=18.12
            DF Sum Sq Mean Sq   F   Prob>F
Regression   2  90.7020   45.3510 45.07 0.0001
Error       110 110.6778    1.0061
Total       112 201.37982301
            Par     Std Type II
Variable    Est       Error  Sum Sq      F Prob>F
INTERCEP  0.80549 0.48776  2.7440  2.73 0.1015
CULTURE   0.05645 0.00980 33.3969 33.19 0.0001
STAY      0.27547 0.05246 27.7388 27.57 0.0001
-----------------------------------------------------
```

```
Step 3 Var FACIL Entered R-sq=0.493 C(p)=10.33
            DF Sum of Sq Mean Sq    F   Prob>F
Regression   3   99.3608 33.1203 35.39 0.0001
Error       109  102.0190  0.9360
Total       112  201.3798
           Par    Std      Type II
Variable    Est  Error    Sum Sq     F  Prob>F
INTERCEP  0.4913 0.4816  0.9740  1.04 0.3099
CULTURE   0.0542 0.0095 30.5982 32.69 0.0001
STAY      0.2239 0.0534 16.4766 17.60 0.0001
FACIL     0.0196 0.0065  8.6588  9.25 0.0029
----------------------------------------------
```

```
Step 4 Var NRATIO Entered R-sq=0.525 C(p)= 5.03
             DF Sum of Sq Mean Sq   F   Prob>F
Regression    4 105.8210 26.4552 29.90 0.0001
Error        108  95.5589  0.8848
Total        112 201.3798
             Par    Std       Type II
Variable      Est  Error    Sum Sq    F    Prob>F
INTERCEP  -0.4951 0.5938  0.6151  0.70 0.4063
CULTURE    0.0482 0.0095 22.8451 25.82 0.0001
STAY       0.2676 0.0543 21.4500 24.24 0.0001
NRATIO     0.7926 0.2933  6.4601  7.30 0.0080
FACIL      0.0175 0.0063  6.7535  7.63 0.0067
---------------------------------------------
```

An example of *Stepwise Selection*

```
   Stepwise Procedure for Dependent Var RISK
Step 1 Var CULTURE Entrd R-sq=0.313 C(p)=47.48
           DF  Sum Sq  Mean Sq    F  Prob>F
Regression  1  62.9631 62.9631 50.49 0.0001
Error     111 138.4167  1.2470
Total     112 201.3798
          Par    Std   Type II
Variable  Est   Error   Sum Sq    F    Prob>F
INTERCEP 3.1979 0.1938 339.6491 272.37 0.0001
CULTURE  0.0733 0.0103  62.9631  50.49 0.0001
```

An example of *Stepwise Selection*

Stepwise Procedure for Dependent Var RISK

Step 1 Var CULTURE Entrd R-sq=0.313 C(p)=47.48

	DF	Sum Sq	Mean Sq	F	Prob>F
Regression	1	62.9631	62.9631	50.49	0.0001
Error	111	138.4167	1.2470		
Total	112	201.3798			

Variable	Par Est	Std Error	Type II Sum Sq	F	Prob>F
INTERCEP	3.1979	0.1938	339.6491	272.37	0.0001
CULTURE	0.0733	0.0103	62.9631	50.49	0.0001

steps 2 & 3 hidden…

Step 4 Var NRATIO Entered R-sq=0.525 C(p)=5.0278

	DF	Sum Sq	Mean Sq	F	Prob>F
Regression	4	105.8210	26.4552	29.90	0.0001
Error	108	95.5589	0.88480418		
Total	112	201.37982301			

Variable	Par Est	Std Error	Type II Sum Sq	F	Prob>F
INTERCEP	-0.4951	0.5938	0.6151	0.70	0.4063
CULTURE	0.0482	0.0095	22.8451	25.82	0.0001
STAY	0.2676	0.0543	21.4500	24.24	0.0001
NRATIO	0.7926	0.2933	6.4601	7.30	0.0080
FACIL	0.0175	0.0063	6.7535	7.63	0.0067

An example of *Stepwise Selection*

Stepwise Procedure for Dependent Var RISK

Step 1 Var CULTURE Entrd R-sq=0.313 C(p)=47.48

	DF	Sum Sq	Mean Sq	F	Prob>F
Regression	1	62.9631	62.9631	50.49	0.0001
Error	111	138.4167	1.2470		
Total	112	201.3798			

Variable	Par Est	Std Error	Type II Sum Sq	F	Prob>F
INTERCEP	3.1979	0.1938	339.6491	272.37	0.0001
CULTURE	0.0733	0.0103	62.9631	50.49	0.0001

steps 2 & 3 hidden…

Step 4 Var NRATIO Entered R-sq=0.525 C(p)=5.0278

	DF	Sum Sq	Mean Sq	F	Prob>F
Regression	4	105.8210	26.4552	29.90	0.0001
Error	108	95.5589	0.88480418		
Total	112	201.37982301			

Variable	Par Est	Std Error	Type II Sum Sq	F	Prob>F
INTERCEP	-0.4951	0.5938	0.6151	0.70	0.4063
CULTURE	0.0482	0.0095	22.8451	25.82	0.0001
STAY	0.2676	0.0543	21.4500	24.24	0.0001
NRATIO	0.7926	0.2933	6.4601	7.30	0.0080
FACIL	0.0175	0.0063	6.7535	7.63	0.0067

Step 5 Var CHEST Entered R-sq=0.538 C(p)=4.19

	DF	Sum Sq	Mean Sq	F	Prob>F
Regression	5	108.3272	21.6654	24.91	0.0001
Error	107	93.0527	0.8697		
Total	112	201.3798			

Variable	Par Est	Std Error	Type II Sum Sq	F	Prob>F
INTERCEP	-0.7680	0.6102	1.3776	1.58	0.2109
CULTURE	0.0432	0.0098	16.7198	19.23	0.0001
STAY	0.2339	0.0574	14.4381	16.60	0.0001
NRATIO	0.6724	0.2993	4.3888	5.05	0.0267
CHEST	0.0092	0.0054	2.5062	2.88	0.0925
FACIL	0.0184	0.0063	7.4571	8.57	0.0042

An example of *Stepwise Selection*

```
  Stepwise Procedure for Dependent Var RISK
Step 1 Var CULTURE Entrd R-sq=0.313 C(p)=47.48
           DF   Sum Sq  Mean Sq     F   Prob>F
Regression  1   62.9631 62.9631 50.49 0.0001
Error      111 138.4167  1.2470
Total      112 201.3798
          Par      Std     Type II
Variable  Est     Error    Sum Sq     F     Prob>F
INTERCEP 3.1979 0.1938 339.6491 272.37  0.0001
CULTURE  0.0733 0.0103  62.9631  50.49  0.0001
```

steps 2 & 3 hidden...

```
Step 4 Var NRATIO Entered R-sq=0.525 C(p)=5.0278
              DF    Sum Sq   Mean Sq      F  Prob>F
Regression     4     105.8210  26.4552  29.90 0.0001
Error        108      95.5589   0.88480418
Total        112     201.37982301
               Par         Std        Type II
Variable       Est          Error     Sum Sq      F  Prob>F
INTERCEP  -0.4951   0.5938      0.6151    0.70 0.4063
CULTURE    0.0482   0.0095     22.8451   25.82 0.0001
STAY       0.2676   0.0543     21.4500   24.24 0.0001
NRATIO     0.7926   0.2933      6.4601    7.30 0.0080
FACIL      0.0175   0.0063      6.7535    7.63 0.0067
```

```
Step 5 Var CHEST Entered R-sq=0.538 C(p)=4.19
           DF Sum Sq    Mean Sq     F   Prob>F
Regression  5 108.3272 21.6654 24.91 0.0001
Error      107   93.0527   0.8697
Total      112 201.3798
            Par    Std     Type II
Variable    Est   Error   Sum Sq      F   Prob>F
INTERCEP -0.7680 0.6102   1.3776   1.58 0.2109
CULTURE   0.0432 0.0098 16.7198 19.23 0.0001
STAY      0.2339 0.0574 14.4381 16.60 0.0001
NRATIO    0.6724 0.2993   4.3888   5.05 0.0267
CHEST     0.0092 0.0054   2.5062   2.88 0.0925
FACIL     0.0184 0.0063   7.4571   8.57 0.0042
---------------------------------------------
```

```
Step 6 Var CHEST Removed R-sq=0.525 C(p)=5.03
            DF   Sum Sq    Mean Sq     F   Prob>F
Regression   4  105.8210 26.4552 29.90 0.0001
Error       108   95.5589   0.8848
Total       112  201.3799
            Par      Std    Type II
Variable    Est     Error   Sum Sq     F    Prob>F
INTERCEP -0.4951 0.5938   0.6151   0.70 0.4063
CULTURE   0.0482 0.0095 22.8451 25.82 0.0001
STAY      0.2676 0.0543 21.4500 24.24 0.0001
NRATIO    0.7926 0.2933   6.4601   7.30 0.0080
FACIL     0.0175 0.0063   6.7535   7.63 0.0067
---------------------------------------------
```

An example of *Stepwise Selection*: the final model result (note chest being removed in the last step)

```
All variables left in the model are
significant at the 0.0500 level.  The
stepwise method terminated because the next
variable to be entered was just removed.
Summary of Stepwise Proc for Dependent Var RISK
      Variable Num Partl Model
Step Entd   Rem In   R**2 R**2 C(p)     F   Prob>F
 1   CULTURE     1 0.313 0.313 47.48 50.49 0.0001
 2   STAY        2 0.138 0.450 18.12 27.57 0.0001
 3   FACIL       3 0.043 0.493 10.33  9.25 0.0029
 4   NRATIO      4 0.032 0.526  5.03  7.30 0.0080
 5   CHEST       5 0.012 0.538  4.19  2.88 0.0925
 6        CHEST  4 0.012 0.526  5.02  2.88 0.0925
```

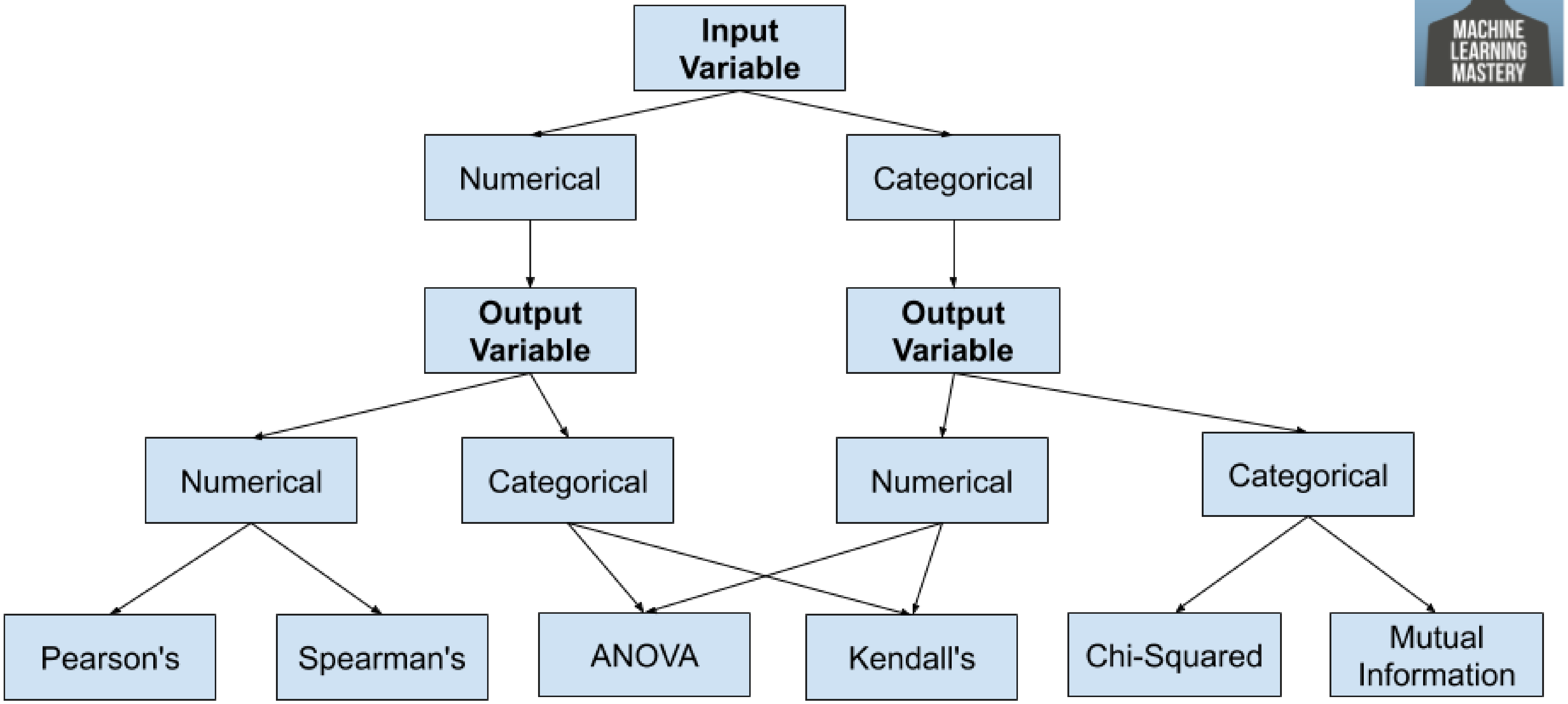
```
def tryVariableSelection(pred, targ, sel, dir, labels):
    ranseed = 98043
    xtrain, xtest, ytrain, ytest = skmodelsel.train_test_split(pred, targ, test_size=0.3, random_state=ranseed)
    model = sklinear_model.LinearRegression()
    if sel == 'sequential':
        selector = featsel.SequentialFeatureSelector(model, direction=dir, n_features_to_select=6)
    elif sel == 'RFE':
        selector = featsel.RFE(model, step=1, n_features_to_select=6)
    elif sel == 'RFECV':
        selector = featsel.RFECV(model, step=1, cv=5)
    selector.fit(xtrain, ytrain)
    newxtrain = selector.transform(xtrain)
    newxtest = selector.transform(xtest)
    model.fit(newxtrain, ytrain)
    print("\nUsing: {0}".format(labels[selector.get_support() == True]))
    print("Method {0}: Training set R-sq={1:8.5f},  test set MSE={2:e}".format(dir, model.score(newxtrain, 
ytrain),sk.metrics.mean_squared_error(ytest, model.predict(newxtest))))
```

```
xf = df[featureLabels]
yf = df[targetLabel]
newpred = imputer.fit_transform(xf.to_numpy(np.float64))
scaler = skpreproc.MinMaxScaler()
normpred = scaler.fit_transform(newpred)
target = yf.to_numpy(np.float64)
xtrain, xtest, ytrain, ytest = skmodelsel.train_test_split(normpred, target, test_size=0.3, random_state=ranseed)

model = sklinear_model.LinearRegression()
xtraintrim = xtrain[:,0:6]
xtesttrim = xtest[:,0:6]
regr = model.fit(xtraintrim, ytrain)
print("\nUsing: {0}".format(featureLabels[0:6]))
print("First 6: Training set R-sq={0:8.5f},  test set MSE={1:e}".format(regr.score(xtraintrim, 
ytrain),sk.metrics.mean_squared_error(ytest, regr.predict(xtesttrim))))

tryVariableSelection(normpred, target, 'sequential', 'forward', featureLabels)
tryVariableSelection(normpred, target, 'sequential', 'backward', featureLabels)
tryVariableSelection(normpred, target, 'RFE', 'RFE', featureLabels)
tryVariableSelection(normpred, target, 'RFECV', 'RFECV', featureLabels)
```

```
Using: ['yearID' 'G' 'AB' 'R' 'H' '2B']
First 6: Training set R-sq= 0.20288,  test set MSE=1.039803e+13

Using: ['yearID' 'G' 'R' 'HR' 'SO' 'GIDP']
Method forward: Training set R-sq= 0.24124,  test set MSE=1.025478e+13

Using: ['yearID' 'G' 'AB' 'H' 'HR' 'GIDP']
Method backward: Training set R-sq= 0.24515,  test set MSE=1.033069e+13

Using: ['yearID' 'G' 'AB' 'H' 'HR' 'RBI']
Method RFE: Training set R-sq= 0.24106,  test set MSE=1.021289e+13

Using: ['yearID' 'G' 'AB' 'R' 'H' '2B' 'HR' 'RBI' 'SB' 'CS' 'SF' 'GIDP']
Method RFECV: Training set R-sq= 0.25421,  test set MSE=1.018128e+13
```

```
xf = df[featureLabels]
yf = df[targetLabel]
newpred = imputer.fit_transform(xf
scaler = skpreproc.MinMaxScaler()
normpred = scaler.fit_transform(ne
target = yf.to_numpy(np.float64)
xtrain, xtest, ytrain, ytest = skm                                   seed)

model = sklinear_model.LinearRegre
xtraintrim = xtrain[:,0:6]
xtesttrim = xtest[:,0:6]
regr = model.fit(xtraintrim, ytrain)
print("\nUsing: {0}".format(featureLabels[0:6]))
print("First 6: Training set R-sq={0:8.5f},  test set MSE={1:e}".format(regr.score(xtraintrim,
ytrain),sk.metrics.mean_squared_error(ytest, regr.predict(xtesttrim))))

tryVariableSelection(normpred, target, 'sequential', 'forward', featureLabels)
tryVariableSelection(normpred, target, 'sequential', 'backward', featureLabels)
tryVariableSelection(normpred, target, 'RFE', 'RFE', featureLabels)
tryVariableSelection(normpred, target, 'RFECV', 'RFECV', featureLabels)
```

How to Choose a Feature Selection Method



Today's Objectives

Variable Selection

- Concept
- Procedures
  - Forward selection
  - Backward selection
  - Stepwise selection
  - Exhaustive (all subsets) selection