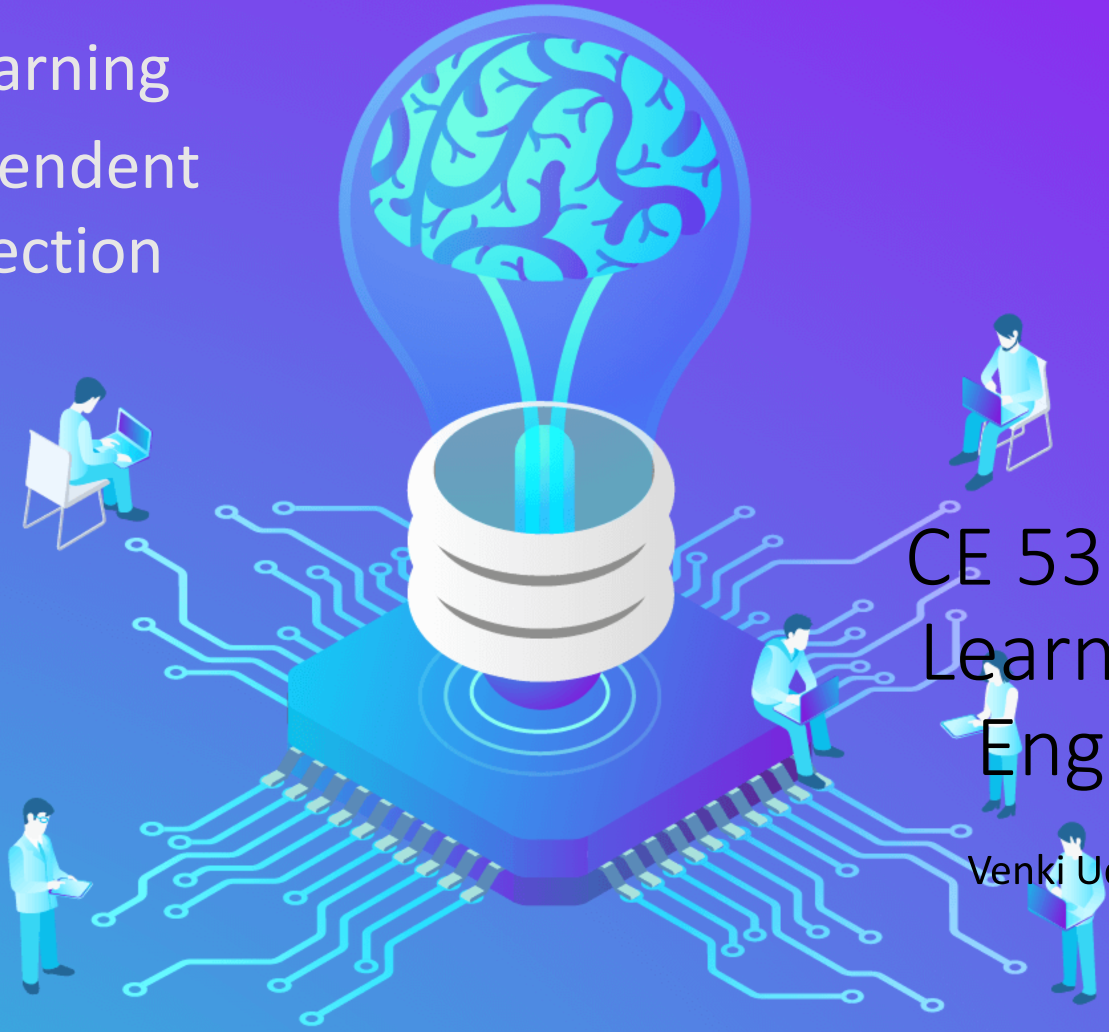
Machine Learning

Model Independent Feature Selection



CE 5331 Machine Learning for Civil Engineers

Venki Uddameri, Ph.D. , P.E.

What is Feature Selection?

When you build machine learning models you compile data from multiple sources

You collect many variables that could potentially describe the output of interest

- A parameter may have several surrogates
- Soil properties related to sand fraction, clay fraction

Which parameters should you use as inputs?

Feature Selection deals with selecting features

Approaches to Feature Selection

- There are many approaches to selecting features
- They can be classified into three broad categories
  - Filter Methods
    - Filter out redundant or non important inputs
  - Wrapper Methods
    - Make use of a model and optimize for best set of subsets
    - Best set of subsets maximize the performance of the model
    - Searching over all features is computationally expensive
      - Make use of heuristics
  - Embedded Methods
    - Embed feature selection as part of the model training

Both Wrapper & Embedded methods require a model to select optimal subsets

Filter Methods

Filter methods are the first step to feature selection

Can be used to identify a smaller subset of features

Can be used to eliminate variables for use with wrapper or embedded methods

Filter methods are model independent so useful with any model

They use statistical concepts but often in an *ad hoc* manner.

Filter Methods

Filter Methods seek to “remove” irrelevant variables

They may rank or order variables according to their relevance

A suitable cut-off can be used to select a subset

How do we define relevance?

- The Feature should be correlated to the output variable
- The Feature should explain the variability noted in the output
- The Feature should contain information about the output variable

Feature selection methods use different relevance criterion

Feature Relevance

Spearman Rank Correlation and Kendall-Tau are two common rank correlation measures

The Pearson Product Moment Correlation (R) is widely used when the inputs (features) and output are continuous

R varies between -1 to +1

Sign of R denotes the nature of the relationship

Magnitude of |R| denotes the strength of the relationship

When the output is not continuous then other correlation measures have to be used

- Integer data can be treated as continuous data

Rank Correlation coefficients can be used as well

- Less sensitive to outliers
- Measures ordinal correlation and not linear correlation
- Useful when the relationship is suspected to be nonlinear

Binary Variable Correlations

Serial, Biserial and Tetrachoric correlations

- Sometimes the output is binary
  - Input can be continuous
  - Input can be binary
- If the output is truly binary and input is continuous then “point biserial correlation” can be used
  - This value can be greater than 1
- If the output is made binary (by discretizing a continuous variable) and the input is continuous then “biserial correlation” can be used
- If both input and output are binary, then “tetrachoric Correlation” can be used

Multinomial Variables

- The correlations can only used for multinomial variables that are ordered
- When the output is ordered multinomial and input is continuous then ‘polyserial’ correlation coefficient is used
  - Generalization of the biserial correlation
- If both input and output variables are ordinal multinomial
  - Polychoric correlation
  - Assumed the underlying variables (latent variables) are continuous

Other Approaches

Contingency table measures can be used to evaluate variables

- Most useful for categorical variables
- Continuous variables can be categorized
  - The values may be affected by binning
- Accuracy is the ratio of sum of diagonal elements to the total elements

Entropy Based Methods

- Entropy is a measure of information content (or lack thereof) in a variable
- Conditional Entropy of y is the information content of y given x is known
- Mutual Information (MI) is the reduction in uncertainty of y due to knowledge of x

$$H(y|x) = -\sum_{y} p(y)log(p(y))$$

$$H(y|x) = -\sum_{x}\sum_{y} p(x,y)log(p(y|x))$$

$$MI(y,x) = H(y) - H(y|x)$$

Use MI to rank variables

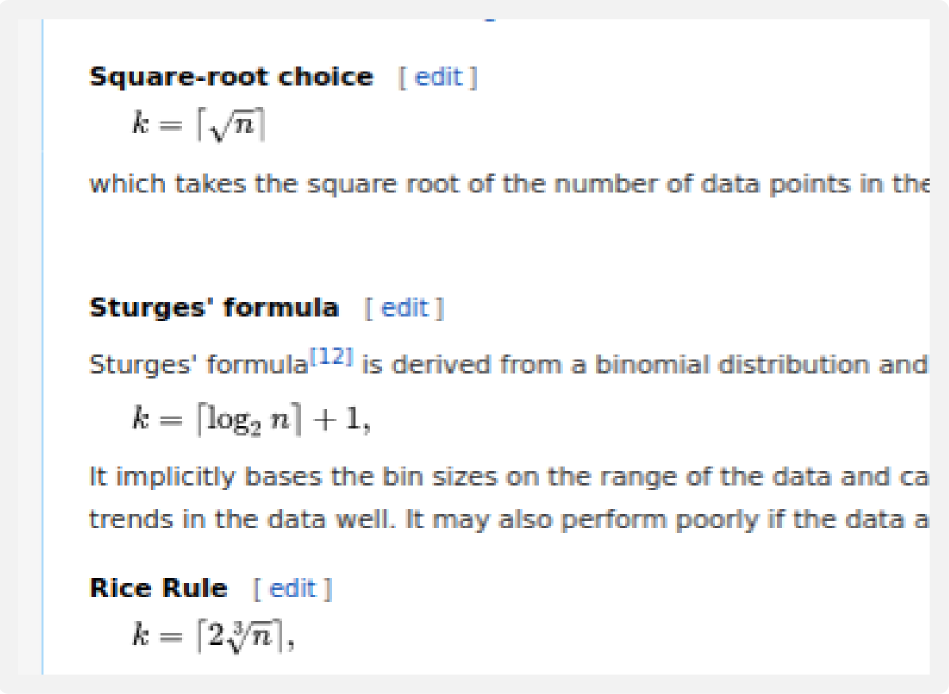
While Entropy can be computed using continuous variables it is common to use discrete versions presented here

Computing Entropy

Entropy computations require discretization

Continuous variables have to be discretized

- Use one of the binning method for discretization
  - Sturges rule for symmetric data
  - Doane’s formula or Freedman-Diaconis choice for skewed data



Square-root choice [edit]

$$k = \lceil \sqrt{n} \rceil$$

which takes the square root of the number of data points in the

Sturges' formula [edit]

Sturges' formula[12] is derived from a binomial distribution and

$$k = \lceil \log_2 n \rceil + 1,$$

It implicitly bases the bin sizes on the range of the data and ca
trends in the data well. It may also perform poorly if the data a

Rice Rule [edit]

$$k = \lceil 2\sqrt[3]{n} \rceil,$$

Both R and Python Provide Methods to compute Entropy & Mutual Information

MultiCollinearity

While an input must be strongly correlated with an output it should not be with other inputs

Inputs that are strongly correlated to each other are redundant

- Add little additional information

Correlation and MI can be used to remove highly correlated variables

Physical understanding of the system and other practical considerations must be borne in mind to decide which variables to remove

- Ease of measurement, data availability

Points to Consider

Use

Use physical basis to identify features

Identify

Identify those features that have strong correlation to the output

- Use appropriate correlation measures
- Use mutual information measures
- Remove variables that cause multicollinearity

Model Free Feature Selection

It is simple and model agnostic

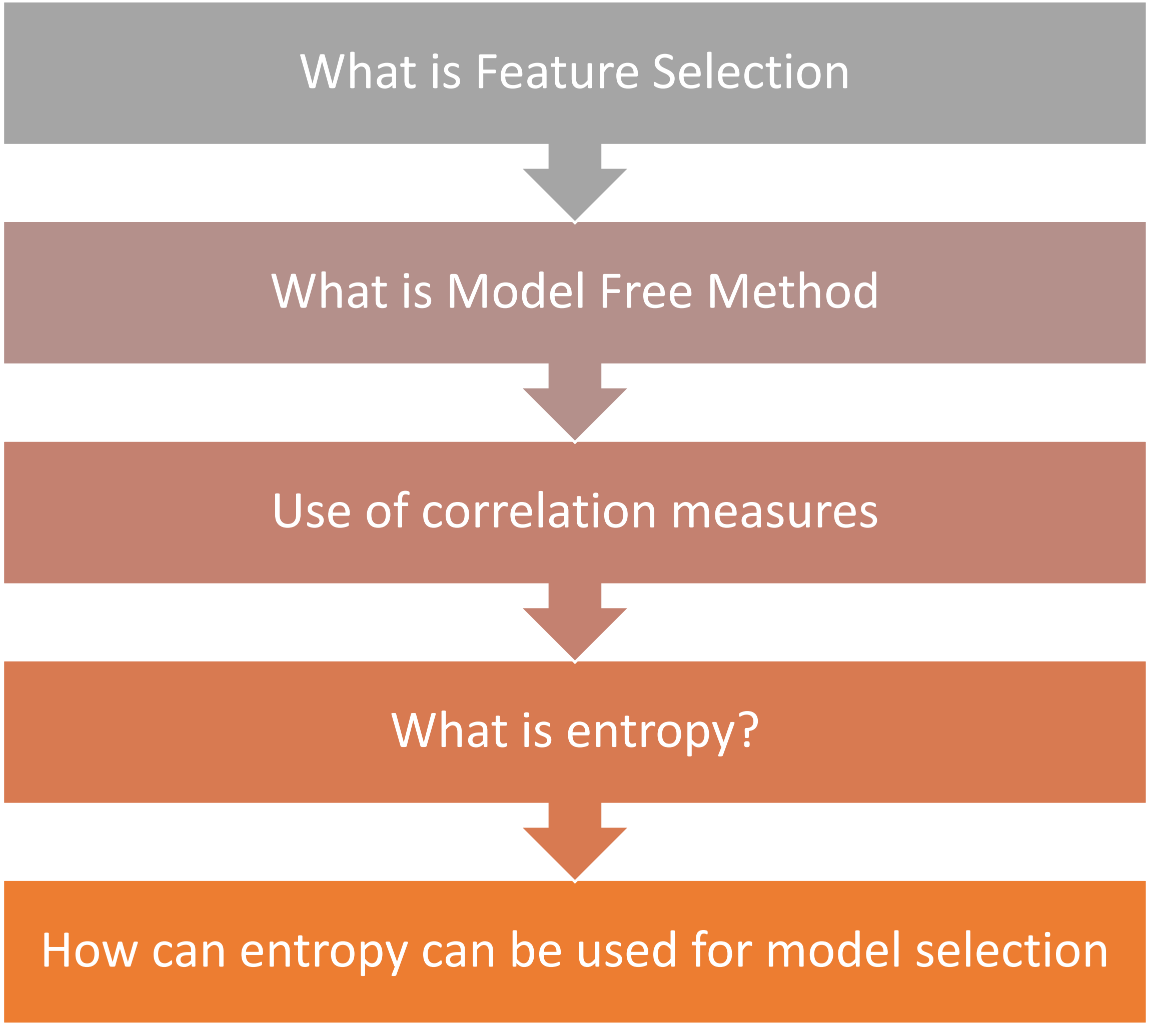
However it is subjective

- The cut-offs for correlation and MI has to be subjectively assigned

Can be used for initial model selection or exclusion of certain variables

Always use a combination of physical considerations and statistical criteria

You Should Know



What is Feature Selection

What is Model Free Method

Use of correlation measures

What is entropy?

How can entropy can be used for model selection