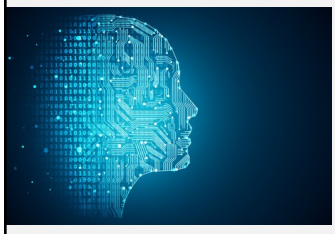


Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

1

PennState Institute for Computational and Data Sciences
Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory
PennState Clinical and Translational Science Institute

Predictive modeling in practice

- Model accuracy depends on the data!
  - Are the samples representative?
  - Do we have enough samples?
  - Do we have an informative set of features?
  - Do you have to gather the features yourself?
    - Different features may have different measurement cost
  - Do we have enough labeled (as opposed to unlabeled) data?
  - Do you have homogenous features e.g., all numeric or all categorical or do you have a mix of heterogeneous features?

PennState College of Information Sciences And Technology
Data Science for Researchers and Scholars
Vasant Honavar, Fall 2023

2

Predictive modeling in practice

- If you want to improve the model performance, you may need
  - More data samples
  - Better (more informative) features
  - Different machine learning algorithms (or hyperparameters)
- One way to decide if you need more/better data
  - Estimate model performance on training and test set
  - If performance on the training data is unsatisfactory, you may need
    - More data samples
    - More informative features
    - Different learning algorithm
  - If performance on the training data is good but performance on test data is unsatisfactory, you may want to
    - Gather more samples
    - Reduce model complexity
      - Regularization e.g., SVM
      - Feature selection

3

Predictive modeling in practice

- Data Types
  - Categorical
    - Nominal – No natural ordering
      - Nominal encoding (e.g., for decision trees)
      - One hot encoding (e.g., for linear classifiers)
    - Ordered/Ordinal – integers that preserve ordering
    - Continuous
      - Normalize using z-score (transform data by subtracting the mean and then dividing by the standard deviation
  - Look at the data to make these and other decisions!

4

What if your data has heterogeneous features?

- Find models that can handle all of the feature types
- Transform features in a sensible way into the form that the chosen machine learning algorithm can handle
  - One-hot-encode the categorical features
  - Nominal – No natural ordering
    - Nominal encoding (e.g., for decision trees)
    - One hot encoding (e.g., for linear classifiers)
  - Ordered/Ordinal and continuous
    - Continuous encoding followed by z-score normalization
    - Mapping continuous to ordered data using binning
    - Thermometer code
- Build different sub-models using different types of features and combine them using a second model

5

Derived features

- Features derived using by domain knowledge
  - BMI calculated from height and weight
  - Force calculated from mass and acceleration
- Features constructed by systematically combining existing features
  - Products of existing numerical features
  - Logical combinations categorical feature values (e.g., (color = red) and (shape = circle)

6

Uninformative features

- Avoid features which
  - Have the same value across almost all data samples
  - Have unique nominal value for each sample
    - E.g., Social security number, phone-number
    - An abstraction of the feature (such as area code) might be useful
  - Are highly correlated with one or more other features
    - Such features are redundant and only one is needed

7

Feature selection and dimensionality reduction

- Given n original features $F = \{f_1, \cdots, f_n\}$ select a subset $F' \subset F$ that results in optimal performance of the trained predictive model
  - Retain informative features, discard uninformative ones
  - Reduces training time and prediction time
  - Reduces sample size needed
  - Mitigates overfitting
- Dimensionality Reduction maps the data represented using the n original features into a new space defined by m derived features
  - Offers similar advantages to feature selection
  - Example: Autoencoders, Principal Component Analysis ..

8

Feature selection

- Given n original features $F = \{f_1, \cdots, f_n\}$ select a subset $F' \subset F$ that results in optimal performance of the trained predictive model
  - You can preselect a size $m < n$
  - Search for the optimal m
    - Exhaustive search over all possible 2^n subsets of F not feasible for large values of n
    - Need efficient compromises

9

Feature selection: Simple Strategy

- Given n original features $F = \{f_1, \cdots, f_n\}$ select a subset $F' \subset F$ that results in optimal performance of the trained predictive model
  - Filters – independent of the learning algorithm used to build the predictive model
    - Use a scoring function to score each feature, e.g., correlation between feature and class label
    - Greedily select m highest scoring features
    - Ignores correlation between features
    - Efficient – complexity $O(n)$

10

Feature selection: Second Order Methods

- Given n original features $F = \{f_1, \cdots, f_n\}$ select a subset $F' \subset F$ that results in optimal performance of the trained predictive model
  - Use a scoring function to score each feature
  - Until m features have been selected or we have run through all n features
    - Pick the highest scoring feature that hasn't already been considered
    - If none of the already chosen features is redundant with the current candidate
      - Add the candidate to the selected list

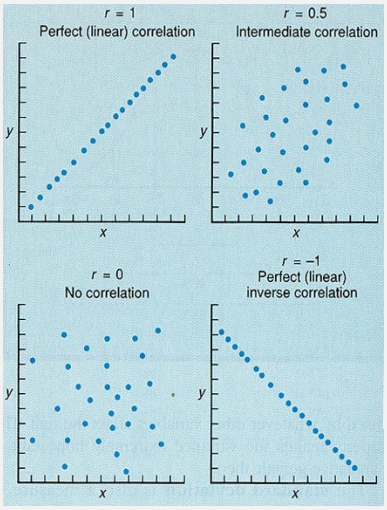
Feature scoring - Correlation

Example:

Pearson correlation coefficient

$$R(f_i, y) = \frac{Cov(f_i, y)}{\sqrt{Var(f_i)Var(y)}}$$

- We estimate the correlation coefficients from the training data
- Correlation lies between -1 and 1
- Higher magnitude of correlation means higher score
- So scoring is done using $|R(f_i, y)|$
- Other scoring functions may be used for other types of features



Variable Ranking – Single Var Classifier

- Select variables according to individual predictive power
- Performance of a classifier built with 1 variable
  - E.g., the value of the variable itself (set threshold on the values)
  - Usually estimated in terms of standard measures on the training data

13

Variable Ranking – Mutual Information

- Empirical estimates of mutual information between features and the label
  - Continuous case
    - $I(f_i,y) = \int_{f_i} \int_y p(f_i, y) \log \frac{p(f_i,y)}{p(f_i)p(y)} df_i dy$
  - Discrete Case
  - $I(f_i,y) = \sum_{f_i} \sum_y P(f_i, y) \log \frac{P(f_i,y)}{P(f_i)P(y)}$

14

Feature Subset Selection

- Requirements:
  - Scoring function to asses the features
  - Strategy to search the space of possible feature subsets
    - Finding the optimal feature subset is general hard
- Methods:
  - Filters
  - Wrappers
  - Embedded

15

PennState Institute for Computational and Data Sciences
Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory
PennState Clinical and Translational Science Institute

Feature Subset Selection - Filters

- Select subsets of variables as a pre-processing step, independently of the predictive model to be trained
  - Variable ranking with score function is a filter method
- Fast and efficient
- Works for training any predictive model
- Not optimized for any specific model
- Sometimes used as a pre-processing step for other methods

PennState College of Information Sciences And Technology
Data Science for Researchers and Scholars
Vasant Honavar, Fall 2023

16

Feature Selection - Wrappers

- Different feature subsets may perform optimally with different learning algorithms
- The feature subset selection algorithm is a "wrapper" around the learning algorithm
- Until some termination criterion is met
  1. Pick a feature subset and pass it to learning algorithm
  2. Train a model on a training set using the selected features
  3. Evaluate the model on a validation set (not the test set)
- There are many variations based on how to select the feature subsets
  - Greedy forward selection
  - Greedy backward selection
  - Etc.

PennState Institute for Computational and Data Sciences
Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory
PennState Clinical and Translational Science Institute

Feature Selection - Wrappers

- Exhaustive Search – Not feasible except for small n
- Forward Search – $O(n^2)$ – Greedy
  - Score each feature by itself and add the best feature to the initially empty set S
  - Try each subset consisting of the current S plus one remaining feature and add the best feature to S
  - Continue until stop getting significant improvement
- Backward Search – $O(n^2)$ – Greedy
  - Score the initial set S of all n features
  - Try each subset consisting of the current S minus one feature in S and drop the feature from S causing least decrease in performance
  - Continue until dropping a feature causes a significant decrease in performance
- Branch and Bound and other heuristic approaches available
- Pro – selected features are customized for the learning algorithm
- Con – computational overhead

Feature Subset Selection – Embedded Methods

- Performs feature selection during training
  - Nested Subset Methods
  - Direct Objective Optimization
    - Formulate the objective function of variable selection and optimize
      - goodness-of-fit (to be maximized)
      - number of variables (to be minimized)
  - Example: Lasso

Feature Selection - Summary

- Feature selection can improve the performance of learning algorithms
  - Predictive performance
  - Cost of training and prediction
- Don’t automatically discard variables with small scores
- Filters, Wrappers, Embedded Methods
  - How to search the space of all feature subsets?
  - How to asses performance of learner that uses a given feature subset?

Dimensionality Reduction

- Goal: reduce data dimensionality
  - Feature grouping – collapse feature values with similar class conditional distributions into single values
  - Transformation of feature space
    - Linear methods (PCA/SVD, LDA)
    - Matrix factorization of variable subsets
    - Kernel methods (kernel PCA)
    - Representation learning (autoencoders and deep autoencoders)