Introduction to Data Science

Metro Health Department

NASHVILLE SOFTWARE SCHOOL

Goals for today

- **Intro to Machine Learning**
- **Learn about Python libraries for statistical analysis and machine learning**

NASHVILLE SOFTWARE SCHOOL

Machine Learning

A “set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data”.

Can be applied to problems for which classical methods are not well-suited (eg. large, high-dimensional data sets).

Often the focus is on **prediction** instead of hypothesis-driven **inference**.

Machine Learning

Supervised Learning:

- Goal: Predict a target variable, given a set of predictors
- Requires a labeled set of data in order to fit a model
- Examples: Linear/Logistic Regression, Decision Trees, Support Vector Machines, Neural Networks

Unsupervised Learning:

- Goal: Uncover natural relationships/groupings within the data
- Examples: Clustering, Mixture Models

Example - Predicting Type 2 Diabetes

Statistical Modeling Approach

Machine Learning Approach

Example - Predicting Type 2 Diabetes

Statistical Modeling Approach

Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Ann Intern Med 2002; 136:575–581.

Data was collected on 5,158 participants.

Fit a multiple logistic regression model using age; sex; ethnicity; fasting and 2-hour glucose levels; systolic and diastolic blood pressures; total, LDL, and HDL cholesterol levels; triglyceride level; body mass index; and parental or sibling history of diabetes.

AUC = 0.75

Machine Learning Approach

NASHVILLE SOFTWARE SCHOOL

Example - Predicting Type 2 Diabetes

Statistical Modeling Approach

Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Ann Intern Med 2002; 136:575–581.

Data was collected on 5,158 participants.

Fit a multiple logistic regression model using age; sex; ethnicity; fasting and 2-hour glucose levels; systolic and diastolic blood pressures; total, LDL, and HDL cholesterol levels; triglyceride level; body mass index; and parental or sibling history of diabetes.

AUC = 0.75

Machine Learning Approach

Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015 Dec;3(4):277-87.

Used administrative claims, pharmacy records, healthcare utilization, and laboratory results of 4.1 million individuals.

Fit a sparse logistic regression model on 42,000 predictor variables.

AUC = 0.80

NASHVILLE SOFTWARE SCHOOL

Common Supervised Learning Algorithms

(Penalized) Linear/Logistic Regression

Sacrifice unbiasedness for greater generalizability

Artificial Neural Networks

Useful when there are potentially complex interactions between predictors. Effective for image recognition tasks.

Support Vector Machines

Finds an optimal hyperplane between target classes.

Common Supervised Learning Algorithms

Decision Trees

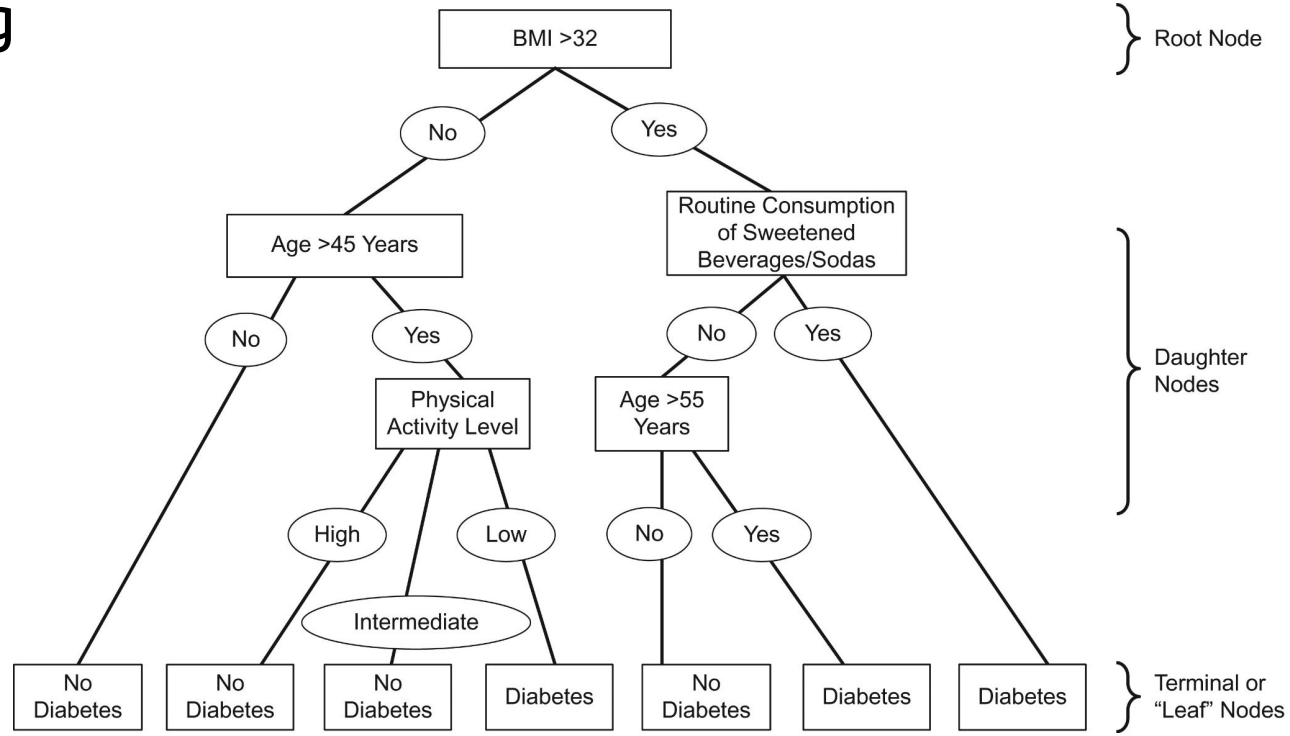
Partition the predictor space using one feature at a time

Ensemble/Superlearner Approach

Combine the predictions of multiple models to produce more accurate predictions

Examples:

- Random Forests
- Gradient Boosted Trees
- Stacking



https://academic.oup.com/aje/article/188/12/2222/5567515

Ensemble/Super Learner Approach

“Standard” Workflow

Ensemble/Super Learner Workflow

Ensemble/Super Learner Approach

“Standard” Workflow

Select from candidate models based on some some utility function.

Examples:

- Choose a simpler (more parsimonious) model
- Choosing a model that maximizes a metric for model fit
- Choosing a model consistent with prior literature

Ensemble/Super Learner Workflow

NASHVILLE SOFTWARE SCHOOL

Ensemble/Super Learner Approach

"Standard" Workflow

Select from candidate models based on some some utility function.

Examples:

- Choose a simpler (more parsimonious) model
- Choosing a model that maximizes a metric for model fit
- Choosing a model consistent with prior literature

Ensemble/Super Learner Workflow

Focus on predictions rather than specifying a single model

Include each candidate model

Either allow the algorithm to select the best model or take a weighted average of all candidate models

Maximize predictive accuracy without overfitting to the training data

NASHVILLE SOFTWARE SCHOOL

Python Libraries for Statistical Analysis and Machine Learning

statsmodels - Statistical tests and building and evaluating statistical models

***scipy stats* module** - probability distributions, descriptive statistics, statistical tests

pymc3 - Bayesian Analysis

scikit-learn - Predictive analysis and machine learning

Questions?