AUBURN
UNIVERSITY

Introduction to Predictive Modeling

Dr. Pei Xu
Auburn University
Thursday, August 22, 2019

Predictive Modeling

- Suppose we observe Y_i and $X_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$
- We believe that there is a relationship between Y and at least one of the X's.
- We can model the relationship as

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

- Where f is an unknown function and ε is a random error with mean zero.

Prediction vs. Inference

Why Do We Estimate f?

- Predictive Modeling, and this course, are all about how to estimate f.
- The term learning refers to using the data to “learn” f.
- Why do we care about estimating f?
- There are 2 reasons for estimating f,
  - **Prediction**
  - **Inference**

Prediction

- If we can produce a good estimate for f (and the variance of ε is not too large) we can make accurate predictions for the response, Y, based on a new value of **X**.

Example: Direct Mailing Prediction

- Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Don’t care too much about each individual characteristic.
- Just want to know: For a given individual should I send out a mailing?

Inference

- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

Classification vs. Regression

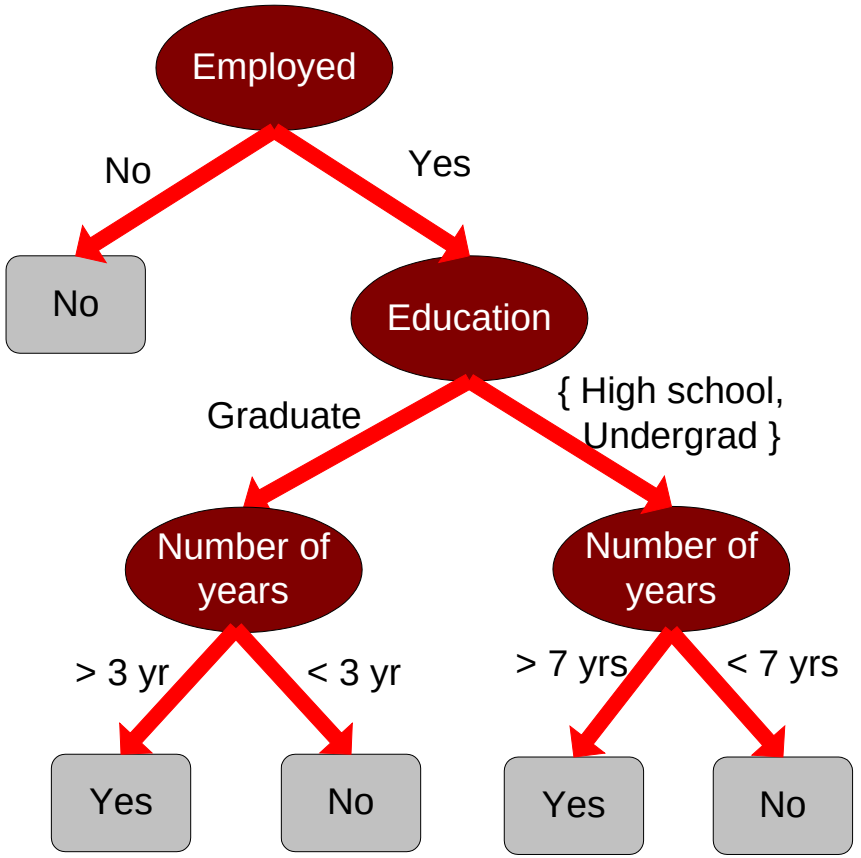
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

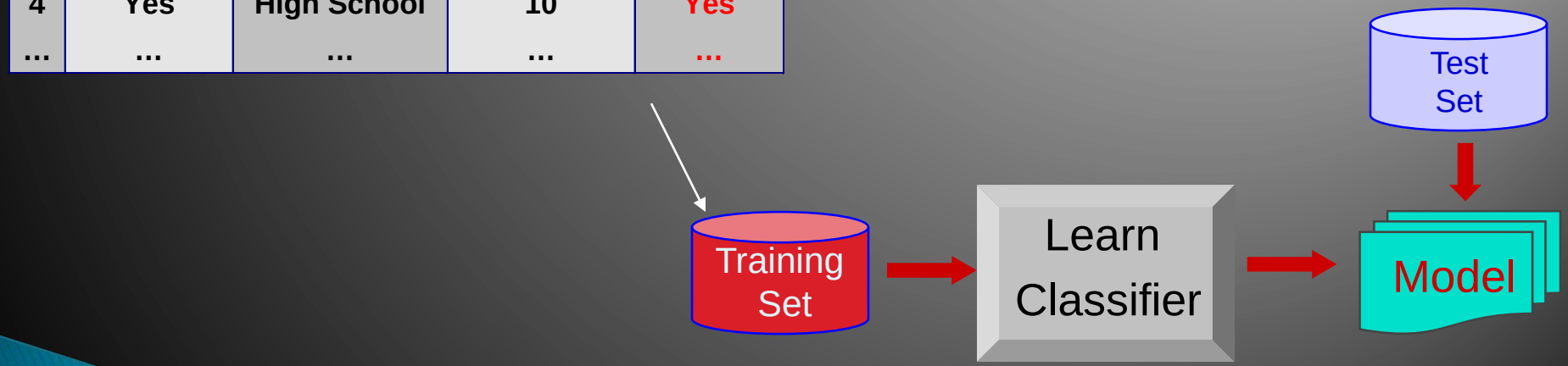


Classification Example

categorical categorical quantitative class

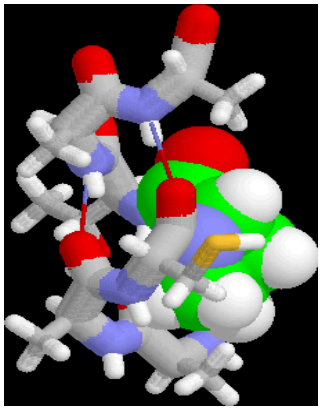
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Review Questions

Explain whether the following scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, describe the dataset (n) and feature set (p) for each case.

- We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Review Questions

Explain whether the following scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, describe the dataset (n) and feature set (p) for each case.

- We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- **Answer**: regression. inference. quantitative output of CEO salary based on CEO firm's features.
- n - 500 firms in the US
- p - profit, number of employees, industry