Large Scale Machine Learning: Decision Trees

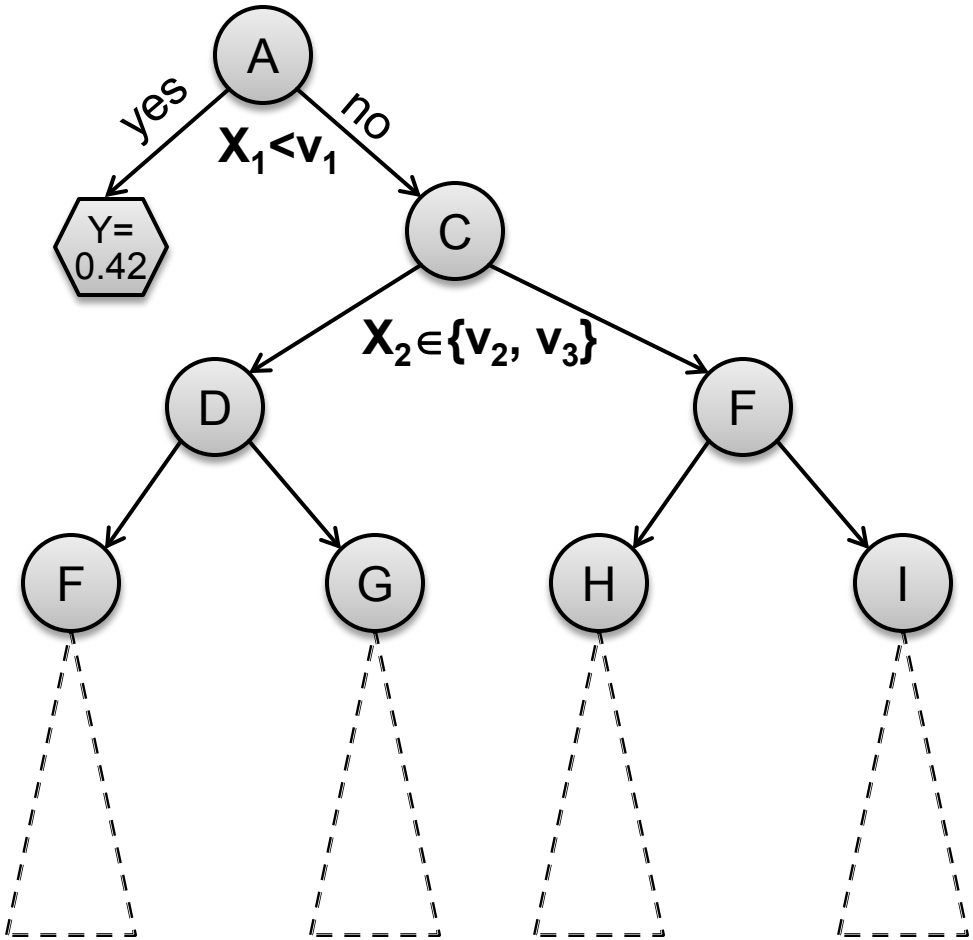
Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Decision Tree Learning

- **Give one attribute (e.g., wealth), try to predict the value of new people's wealths by means of some of the other available attribute**
- **Input attributes:**
  - **d** features/attributes: $\mathbf{X_1}$, $\mathbf{X_2}$, … $\mathbf{X_d}$
  - Each $\mathbf{X_j}$ has **domain** $\mathbf{O_j}$
    - **Categorical:** $\mathbf{O_j}$ = {red, blue}
    - **Numerical:** $\mathbf{H_j}$ = (0, 10)
  - $\boldsymbol{Y}$ is output variable with domain $\boldsymbol{O_Y}$:
    - **Categorical:** Classification, **Numerical:** Regression
- **Data D:**
  - $\boldsymbol{n}$ examples $\boldsymbol{(x_i, y_i)}$ where $\boldsymbol{x_i}$ is a $\boldsymbol{d}$-dim feature vector, $\boldsymbol{y_i} \in \boldsymbol{O_Y}$ is output variable
- **Task:**
  - Given an input data vector $\boldsymbol{x}$ predict $\boldsymbol{y}$

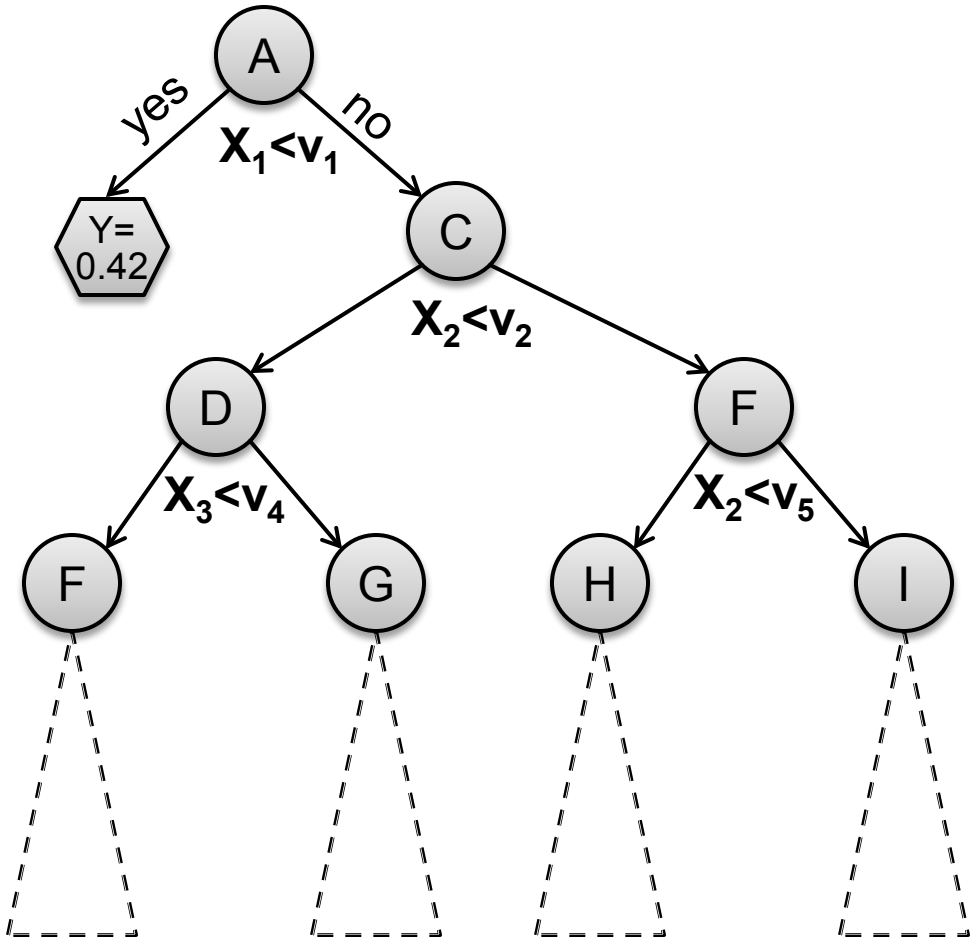
Decision Trees

- A Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output



Decision Trees (1)

- **Decision trees:**
  - Split the data at each internal node
  - Each leaf node makes a prediction
- **Lecture today:**
  - Binary splits: $X_j < v$
  - Numerical attrs.
  - Regression

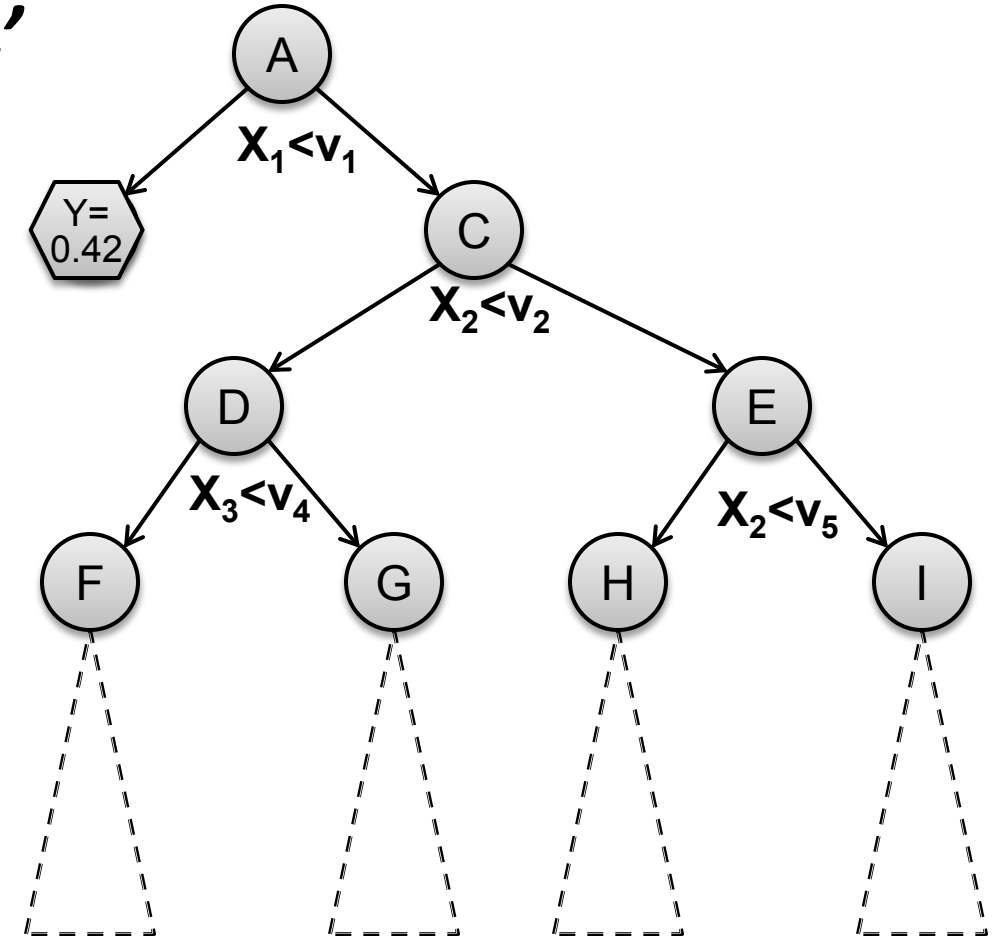


How to make predictions?

- **Input:** Example x_i
- **Output:** Predicted y_i'

- "Drop" x_i down the tree until it hits a leaf node

- Predict the value stored in the leaf that x_i hits



Decision Trees Vs. SVM

- **Alternative view:**