Classification Methods

Tozammel Hossain

Data Science & Analytics
University of Missouri

What are classification methods?

- **A subclass of supervised learning methods**
  - (Input, output) pairs are given for learning
  - Learn a map from input → output
- **Prediction Task**
  - Given a new instance/example, label/tag/annotate the instance/example with a class label
- **Easier to evaluate as ground truth is available**

Classification setting

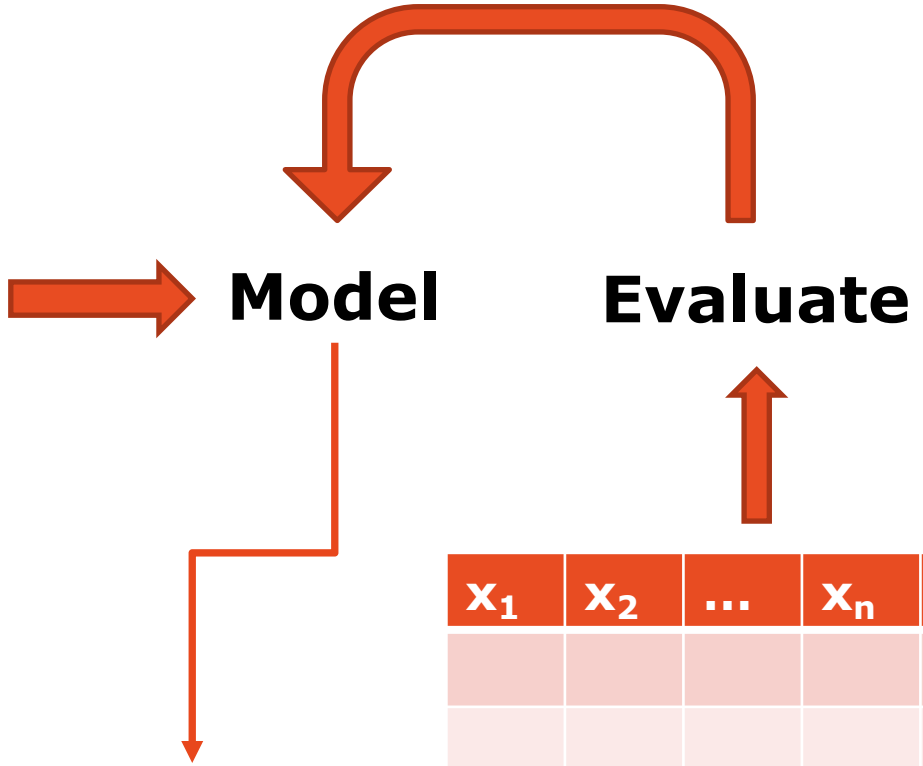
- **Given**
  - A dataset D
    - D = {d1,d2,..} is a set of rows/instances/examples
    - Each instance in D is described by values for a set of features/attributes X = {x1,x2,...}
    - Each instance in D is labeled with a class, from a mutually exclusive set y = {y1,y2,...}
  - Learning
    - class_label = f(input instance)
  - Classify
    - New instances (test dataset) into a class

x_1	x_2	...	x_n	y
				y_1
				y2
				y2
				Y3
				y1

Schematic

x_1	x_2	...	x_n	y
				y_1
				y_2
				y2
				Y3
				y1

x_1	x_2	...	x_n	y
				?
				?
				?
				?
				?



x_1	x_2	...	x_n	y
				y3
				y1
				y1
				y2
				y2

Applications of classification

- **Classifying**
  - Emails: spam/not spam
  - Reviews: good/bad/nasty
  - Commentaries: liberal/conservative
  - Galaxies: elliptical/spiral/bentdouble
  - Genes: cancer-implicated vs not
  - Archeological findings: different eras
  - Federalist papers: Madison/Hamilton
  - Facebook acquaintances: friends/family/frenemies
  - Twitter account: real/fake

Common theme

- **Inputs**
  - Can be any mix of continuous and discrete variables
  - Can be nominal, ordinal, interval, ratio
- **Outputs**
  - Classes are assumed to nominal, discrete

- Good idea for you to characterize the variables in your dataset before attempting any data mining
- Most of the hard work goes into properly encoding the features!

Some typical, wrong, assumptions

- **Assumption 1:**
  - Data from the training and test sets are drawn from the same distribution
- **Assumption 2:**
  - Each instance is labeled with only one class
- **Assumption 3:**
  - Penalty for misclassification the same either way

Classification of classifiers

- **Rule-based**
  - Decision tree (M1)
- **Probabilistic**
  - Naïve Bayes classifier (M2)
- **Max-margin classifier**
  - SVM (M5)
- **Neural network (M8)**

Baseline or Dummy Classifiers

- **Classifiers with simple rules**
- **May not use any features for prediction**
  - Use only the class variable to make prediction
- **A proposed model should perform better than baselines**

Some classification baselines

- **Most frequent**
  - Pick the dominant class value
- **Prior**
  - Estimate the prior probability distr. of class
  - Predict each instance with the class that has max prior
- **Uniform**
  - Each class is equally probable: Pr(a_class_val) = 1/num_class
- **Stratified**
  - Get prior distribution of class variable
  - Sample class value from the distribution

Baselines Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- **Most Frequent**
  - Yes
- **Prior**
  - P(Yes) > P(No)
- **Stratified**
  - P(Yes) = 9/14; P(No) = 5/14
- **Uniform**
  - P(Yes) = P(No) = 1/2

sklearn.dummy.DummyClassifier(*, *strategy='prior', random_state=None, constant=None*)