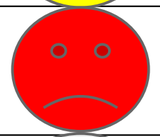
DLI Accelerated Data Science Teaching Kit

Lecture 14.1 - Overview

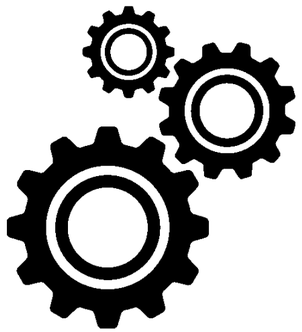
How will I rate "Chopin's 5th Symphony"?

Songs	Like?
Some nights	
Skyfall	
Comfortably numb	
We are young	
...	...
...	...
Chopin's 5th	???

Classification

What tools do you need for classification?

1. **Data** $S = \{(x_i, y_i)\}_{i = 1,...,n}$
   - x_i : data example with d attributes
   - y_i : label of example (what you care about)



2. Classification **model** $f_{(a,b,c,....)}$ with some parameters a, b, c,...
3. **Loss function** L(y, f(x))
   - how to penalize mistakes

Terminology Explanation

data example = data instance

attribute = feature = dimension

label = target attribute

Data S = $\{(x_i, y_i)\}_{i = 1,...,n}$

$x_i = (x_{i1}, \ldots, x_{id})$

- x_i : data example with d attributes
- y_i : label of example

Song name	Artist	Length	...	Like?
Some nights	Fun	4:23	...	
Skyfall	Adele	4:00	...	
Comf. numb	Pink Fl.	6:13	...	
We are young	Fun	3:50	...	
...	...	...	...	...
Chopin's 5th	Chopin	5:32	...	??

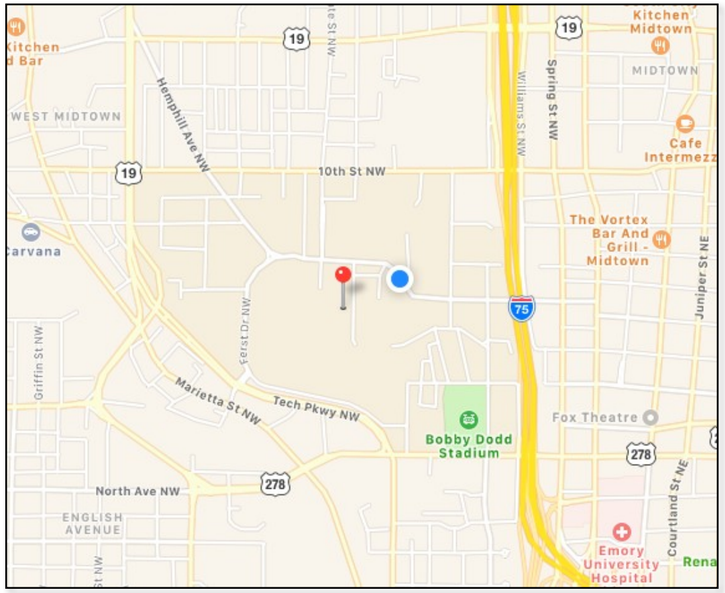
What is a “model”?

“a simplified representation of reality created to serve a purpose” Data Science for Business

Example: maps are abstract models of the physical world

There can be many models!!

(Everyone sees the world differently, so each of us has a different model.)



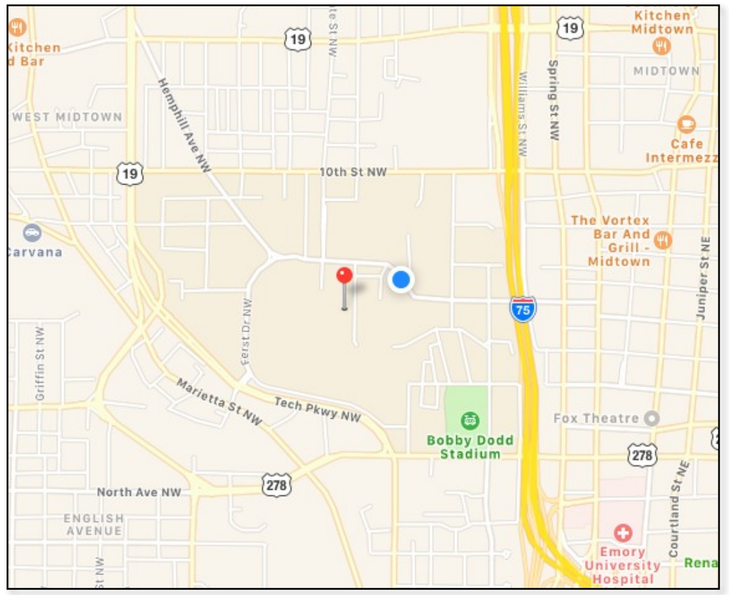
Screenshot from Apple Maps

In data science, a model is formula to estimate what you care about. The formula may be mathematical, a set of rules, a combination, etc.

Training a classifier = building the “model”

How do you learn appropriate values for parameters a, b, c, ... ?

Analogy: how do you know your map is a “good” map of the physical world?



Screenshot from Apple Maps

Classification Loss Function

Most common loss: 0-1 loss function

$$L_{0\text{-}1}(y, f(x)) = I(y \neq f(x))$$

More general loss functions are defined by a m x m cost matrix C such that

$$L(y, f(x)) = C_{ab}$$

where y = a and f(x) = b

Class	T0	T1
P0	0	C_{10}
P1	C_{01}	0

T0 (true class 0), T1 (true class 1)

P0 (predicted class 0), P1 (predicted class 1)

DLI Accelerated Data Science Teaching Kit

Thank You