Introduction to Data Science

SUPERVISED LEARNING PRIMER

BRIAN D'ALESSANDRO
ADJUNCT PROFESSOR, NYU

SUPERVISED VS. UNSUPERVISED

Supervised Learning: the process of inferring a function from labeled data. In SL, we have a target (dependent) variable Y and features (independent variables) X, and our goal is to learn a function Y=f(X).

Unsupervised Learning: the process of finding hidden structure in data that has no label.

Hint: If no label/target/dependent var, then it is probably unsupervised!

TYPES OF LABELS IN SL

SL can be further broken down by the type of target variable.

In regression problems, the labels can be any real valued number.

$$f(x) = y, \text{ where } y \in \mathbb{R}$$

In classification problems, the labels are discrete choices called 'classes', and one either estimates a particular class or the probability of being in a particular class.

$$f(x) = c_i, \text{ where } c_i \in C = [c_1, \ldots, c_k]$$

or

$$f(x) = P(c_i), \text{ where } c_i \in C = [c_1, \ldots, c_k] \text{ and } \sum_{c_i \in C} P(c_i) = 1$$

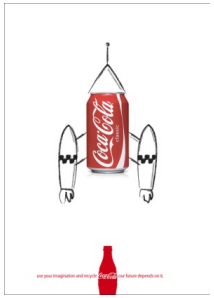
EXAMPLE REGRESSION PROBLEMS

What will the price of IBM stock be tomorrow?

How much will a new customer spend In the next year?

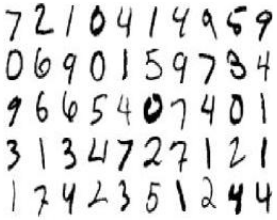
EXAMPLE CLASSIFICATION PROBLEMS

Will someone click on an ad?:
C=[No,Yes]

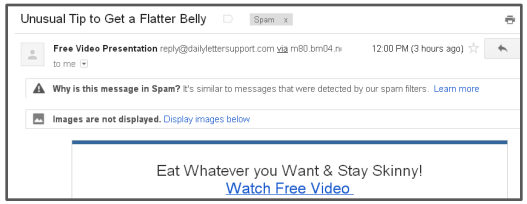


Is this pill good for headaches?:
C=[No,Yes]

What number is this?:
C=[0,1,2,3,4,5,6,7,8,9]



Is this e-mail spam?: *C=[No,Yes]*



What is this news article about?:
C=[Politics, Sports, Finance …]