Introduction to Data Science

Data Mining for Business Analytics

BRIAN D'ALESSANDRO

VP – DATA SCIENCE, DSTILLERY

ADJUNCT PROFESSOR, NYU

FALL 2014

INTRO TO SUPERVISED LEARNING

NYU – Intro to Data Science

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 <u>regression</u> problems, the labels can be any real valued number.

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

In <u>classification</u> 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

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

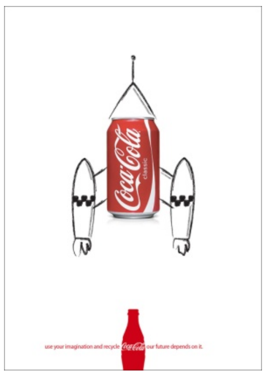
EXAMPLE REGRESSION PROBLEMS

What will the price of IBM stock be tomorrow?

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

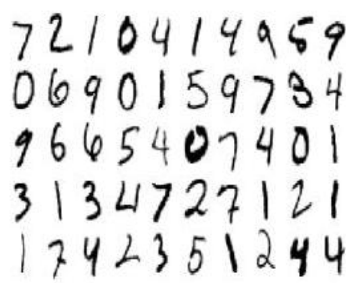
EXAMPLE CLASSIFICATION PROBLEMS

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

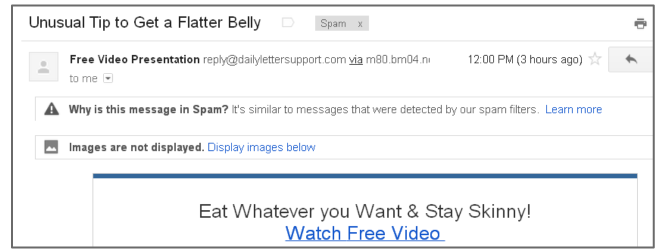


Is this pill good for headaches?:
C=[No, Yes]

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



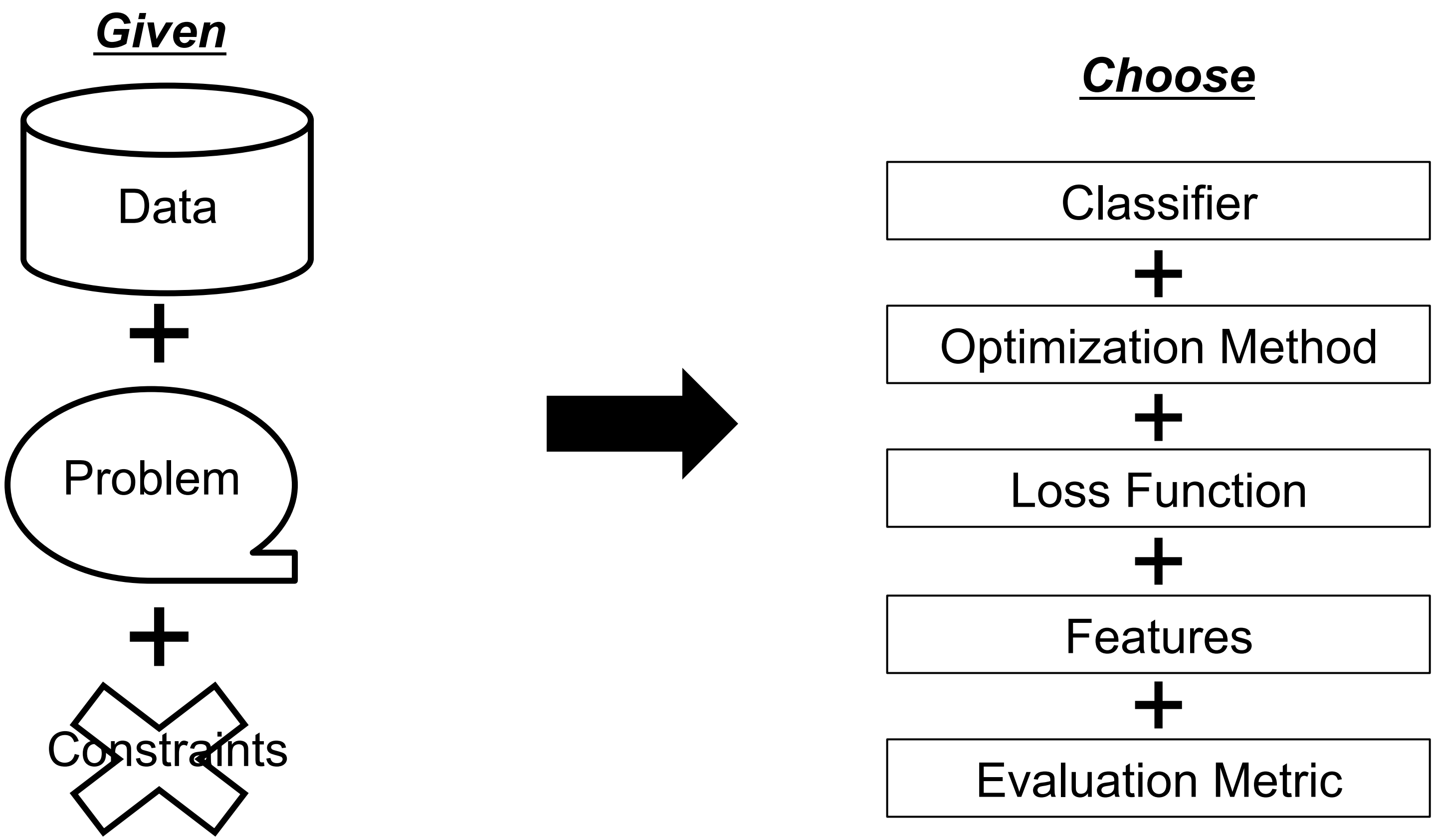
Is this e-mail spam?: *C=[No, Yes]*



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

A COMMON THEME

Few problems have out of the box solutions



The Data Scientist has to navigate these choices

CLASSIFICATION ALGORITHMS

The following is a non-exhaustive list of popular algorithms used in classification problems:

<u>Classic & Simpler Methods</u>

- Decision Tree
- Naïve Bayes
- K- Nearest Neighbors
- Linear Hyperplane

<u>Black Box but Powerful Methods</u>

- Random Forests
- Non-Linear SVM
- Neural Networks

We will NOT discuss each of these algorithms in detail in this course, but we will cover the process of how to choose one.

BUT WHICH ONE SHOULD I USE?

If world free of constraints, then (e.g. a data mining competition):

Try them all, choose best performer

Else:

Consider all constraints on your problem.

Choose best performer subject to constraints

TRY THEM ALL???

Train = Training Data
Val = Validation Data

For each Algorithm in <set of all algorithms>:

Build a classifier, $F^A(X)$ using **Train**

Get out-of-sample error of $F^A(X)$ using **Val**

Choose the Algorithm with the best out-of-sample error.

BAKEOFF RULES

1. Training data must always be disjoint from validation data.

2. Use the same training data and validation data for each hypothesis being tested.

3. Given a tie (statistical or exact), choose the simpler model (sometimes this is subjective).

4. Use this methodology for all design decisions (feature selection, hyper-parameter selection, model selection, etc.)

MODEL SELECTION

This is a generic term that has many flavors

1. The type of algorithm used (Naïve Bayes vs. Decision Tree)

2. The number of features used

3. The definition of the features used

4. The hyper-parameters used (usually related to regularization)

Regardless of how it is defined, use a rigorous validation process to choose. We will study this more in future lectures.

CONSTRAINTS TO CONSIDER

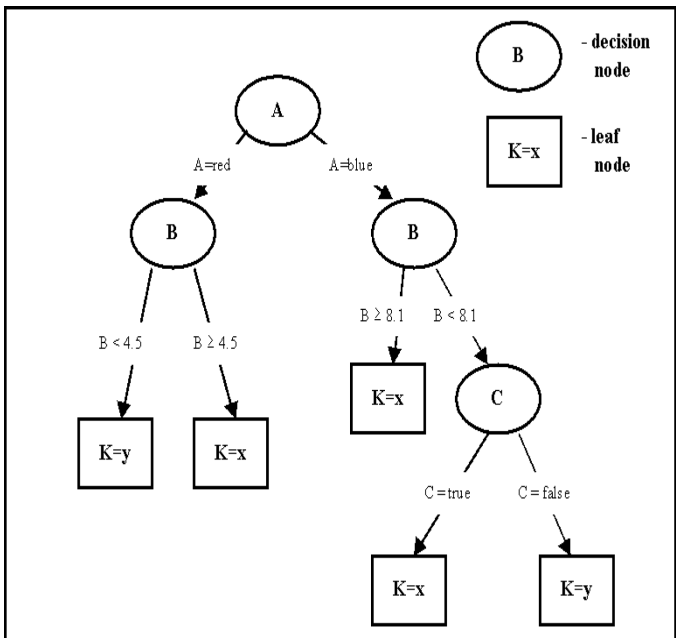
1. Do you understand it?

- Your own personal knowledge is a constraint worth admitting to
- You don't have to master every algorithm to be a good data scientist
- Getting the "best-fit" of an algorithm often requires intimate knowledge of said algorithm

CONSTRAINTS TO CONSIDER

2. Do you need to interpret it?

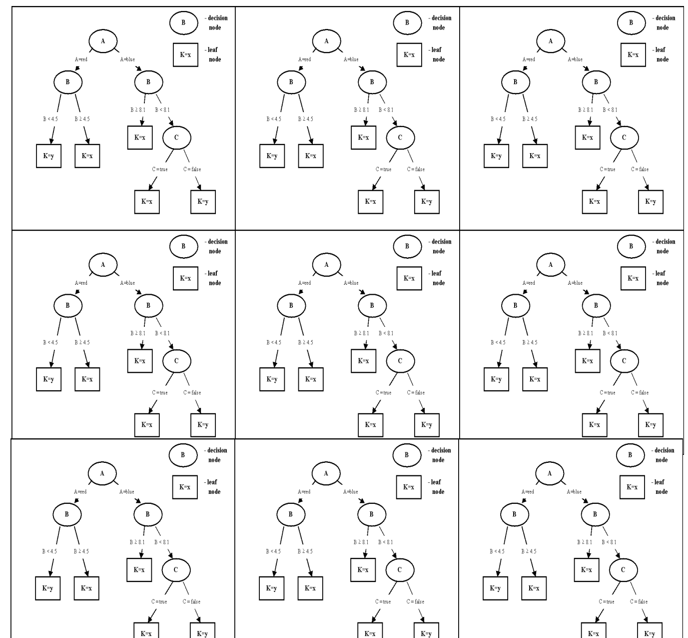
Decision Tree



Very Easy

Vs.

Random Forest

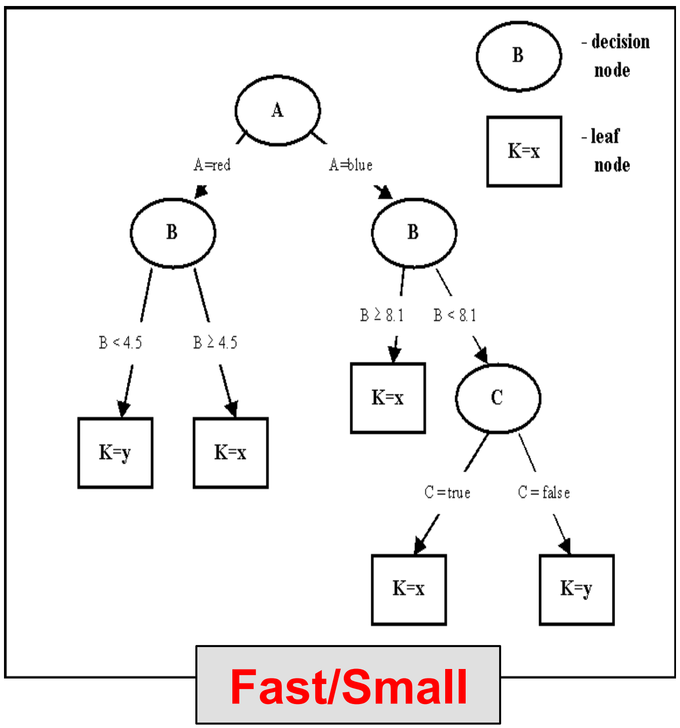


Very Difficult

CONSTRAINTS TO CONSIDER

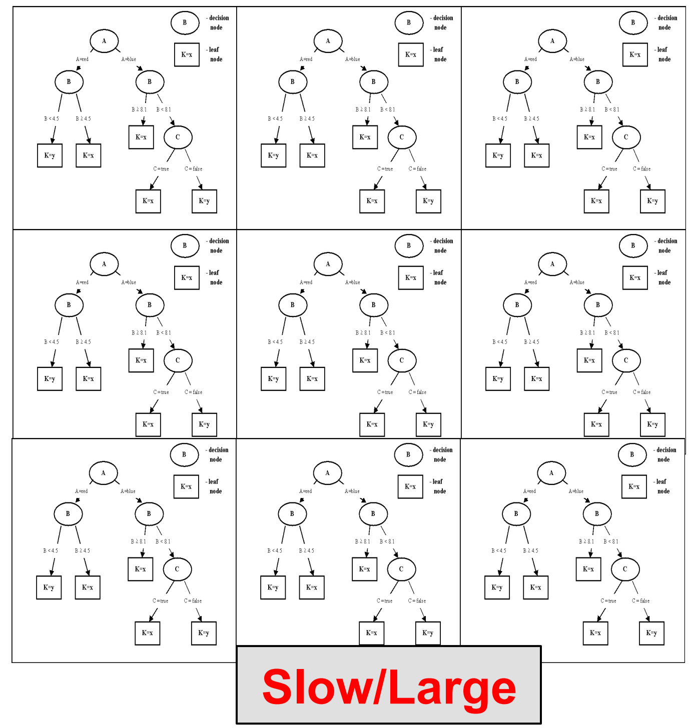
3. Does scalability matter (learning time, scoring time, model storage)?

Decision Tree



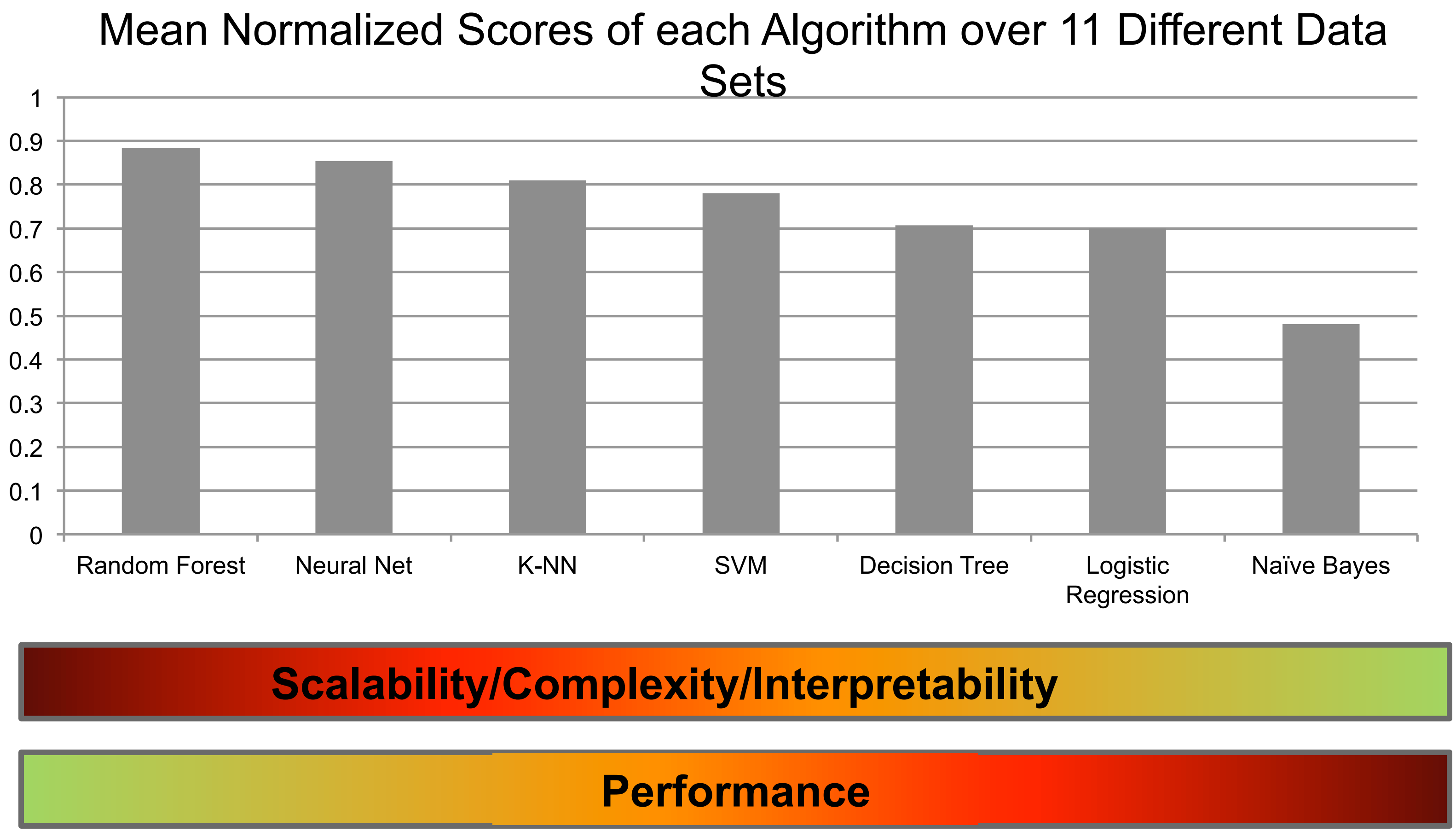
Vs.

Random Forest



NYU – Intro to Data Science

AN EMPIRICAL COMPARISON OF CLASSIFICATION ALGORITHMS



Source: An Empirical Comparison of Supervised Learning Algorithms http://www.niculescu-mizil.org/papers/comparison.tr.pdf