GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 6: CLASSIFICATION - KNN

Paul Burkard

07/07/2015

LAST TIME:

- WHAT IS LINEAR REGRESSION?
    - INPUTS/OUTPUTS?
    - USE CASES?
- WHAT IS CROSS-VALIDATION?
    - TYPES?
- WHAT IS REGULARIZATION?
    - TYPES FOR LINEAR REGRESSION?

TODAY:

I. CLASSIFICATION
II. K-NEAREST NEIGHBORS CLASSIFICATION
HANDS-ON: KNN

- What is **Classification**?
  - What are the inputs and outputs?
  - What are some potential use cases?
- What is **K-Nearest Neighbors**?

I. CLASSIFICATION

*Q: What is a **Classification** model/problem?*
A: A functional relationship between input & response variables...

*Where the target variables are **categorical**!*

$$y = f(\mathbf{X})$$

*The function we seek in a classification problem maps feature vectors to **qualitative/categorical target classes***

Here's (part of) an example dataset:

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

independent variables

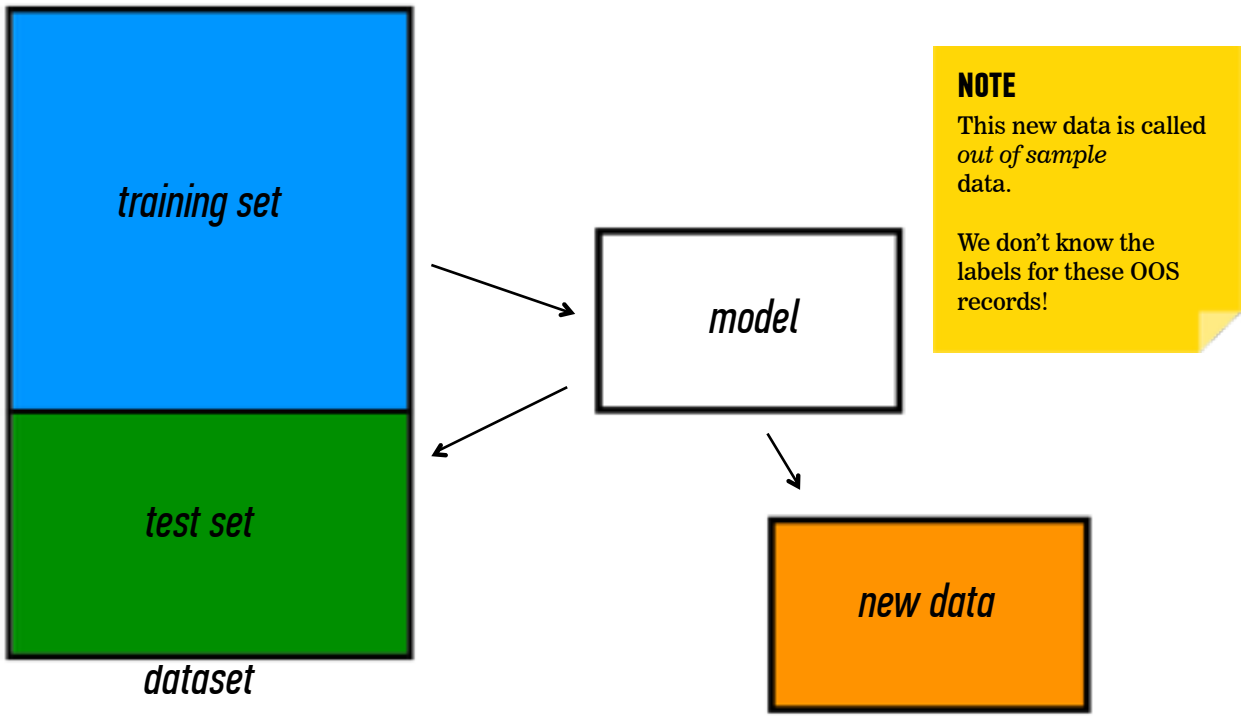
class labels (qualitative)

	continuous	*categorical*
supervised	*???*	*???*
unsupervised	*???*	*???*

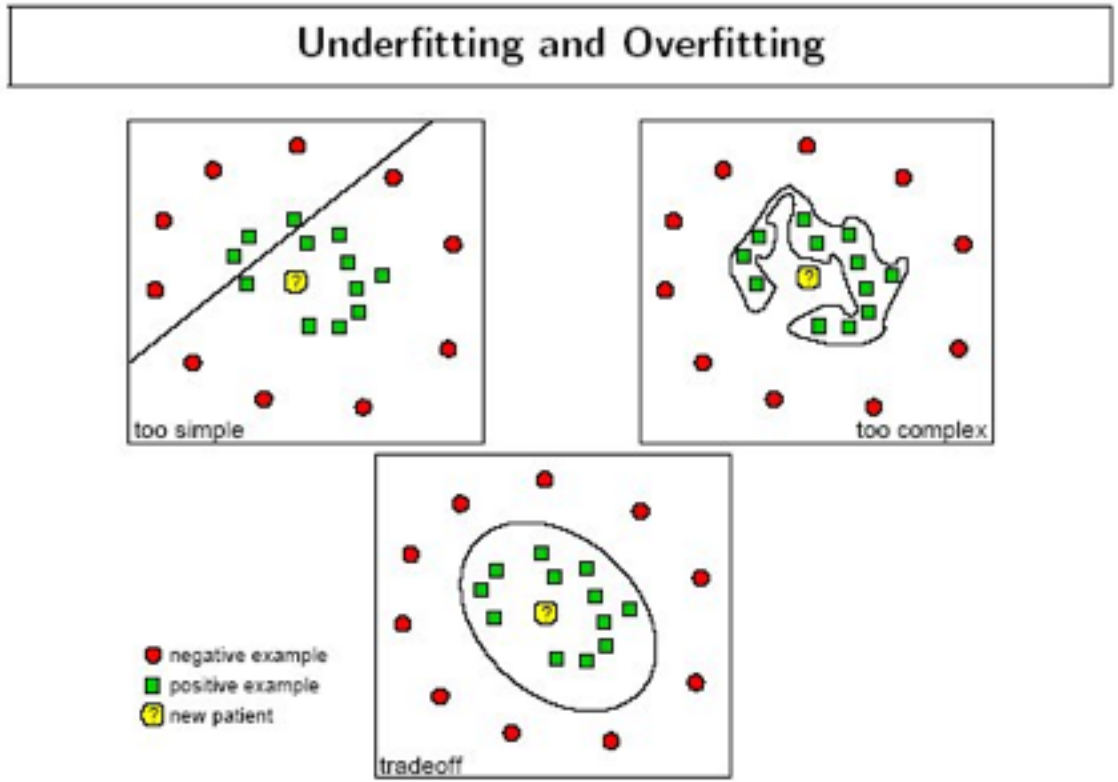
	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dim reduction*	*clustering*

Q: What steps does a supervised learning problem require?

1) split dataset
2) train model
3) test model
4) make predictions

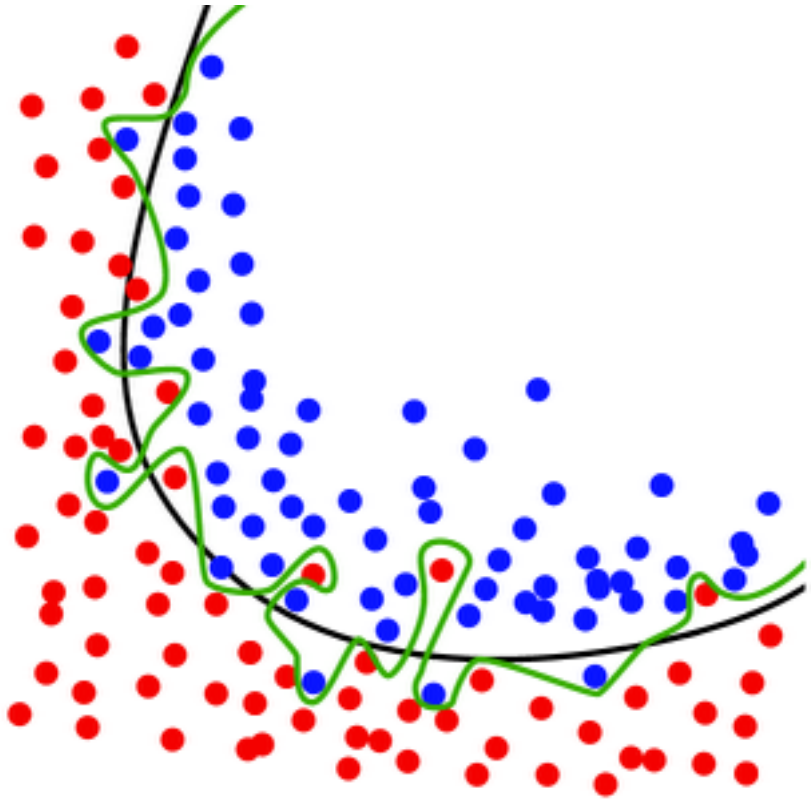


CLASSIFICATION OVERFITTING – EXAMPLE



source: http://www.dtreg.com

CLASSIFICATION OVERFITTING – EXAMPLE

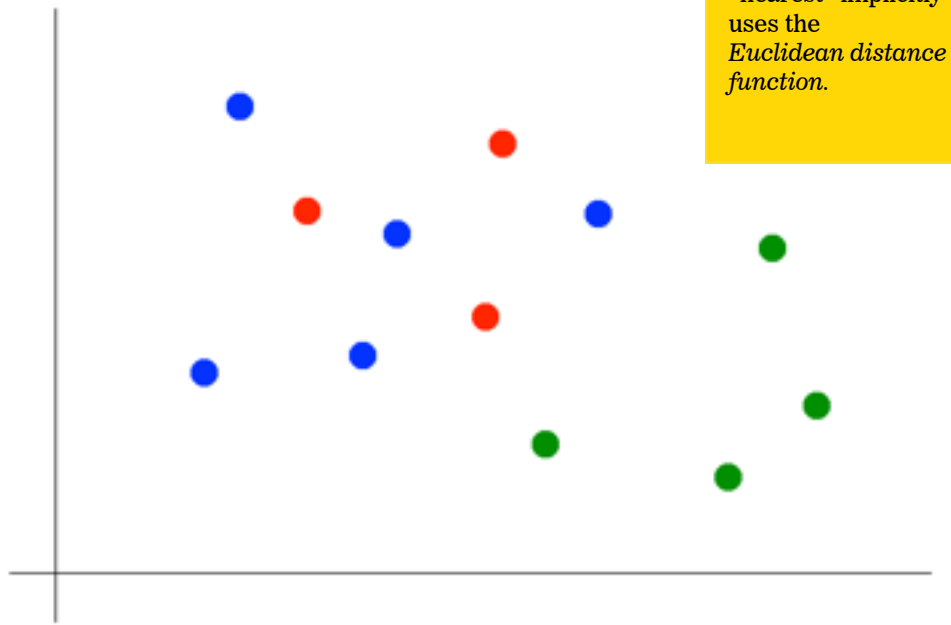


source: http://www.dtreg.com

II. K-NEAREST NEIGHBORS

Suppose we want to predict the color of the grey dot.

1) Pick a value for k.
2) Find colors of k nearest neighbors.
3) Assign the most common color to the grey dot.



OPTIONAL NOTE

Our definition of "nearest" implicitly uses the *Euclidean distance function.*

HANDS-ON: KNN