

# **INTRO TO DATA SCIENCE**

## **LECTURE 6: CLASSIFICATION – KNN**

**Paul Burkard**

**07/09/2015**

## **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**!*

$$y = f(X)$$

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

# CLASSIFICATION PROBLEMS

7

*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>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>

*independent  
variables*

*class  
labels  
(qualitative)*

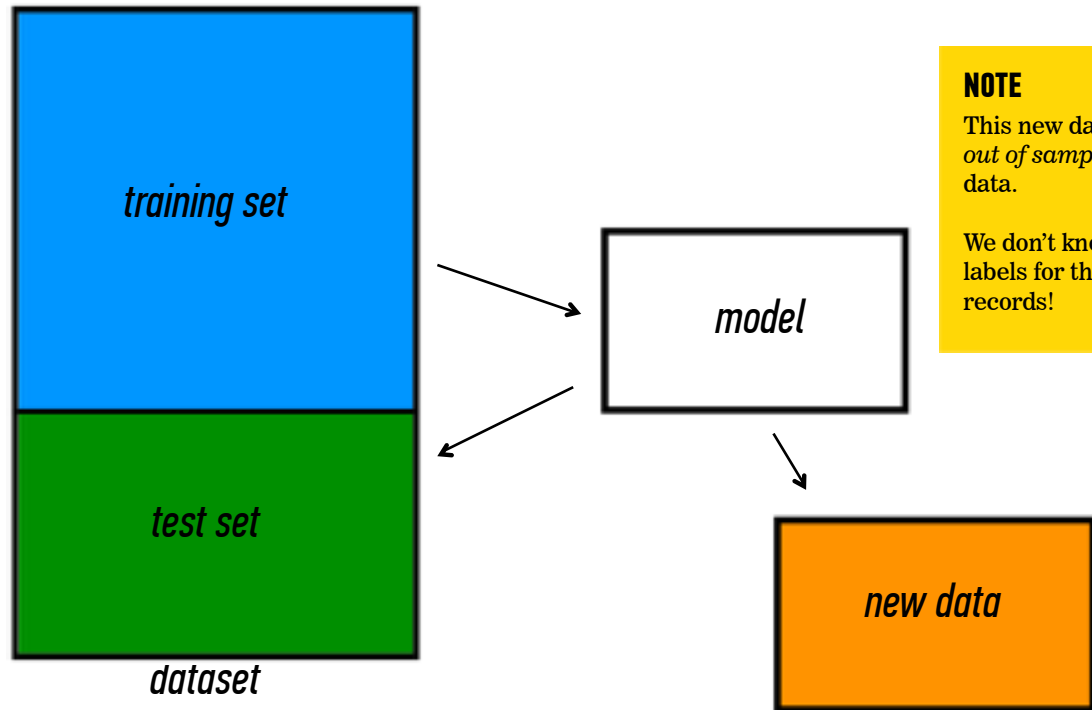
	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	???	???
<i>unsupervised</i>	???	???



	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dim reduction</i>	<i>clustering</i>

*Q: What steps does a supervised learning problem require?*

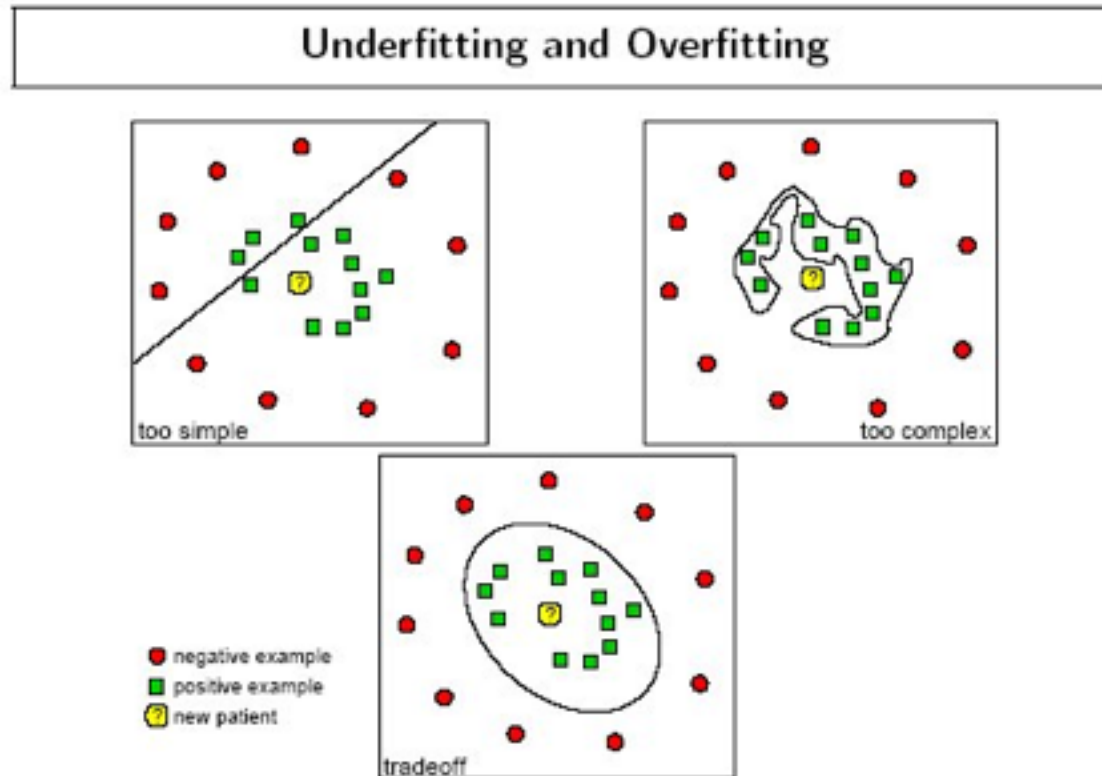
- 1) split dataset*
- 2) train model*
- 3) test model*
- 4) make predictions*

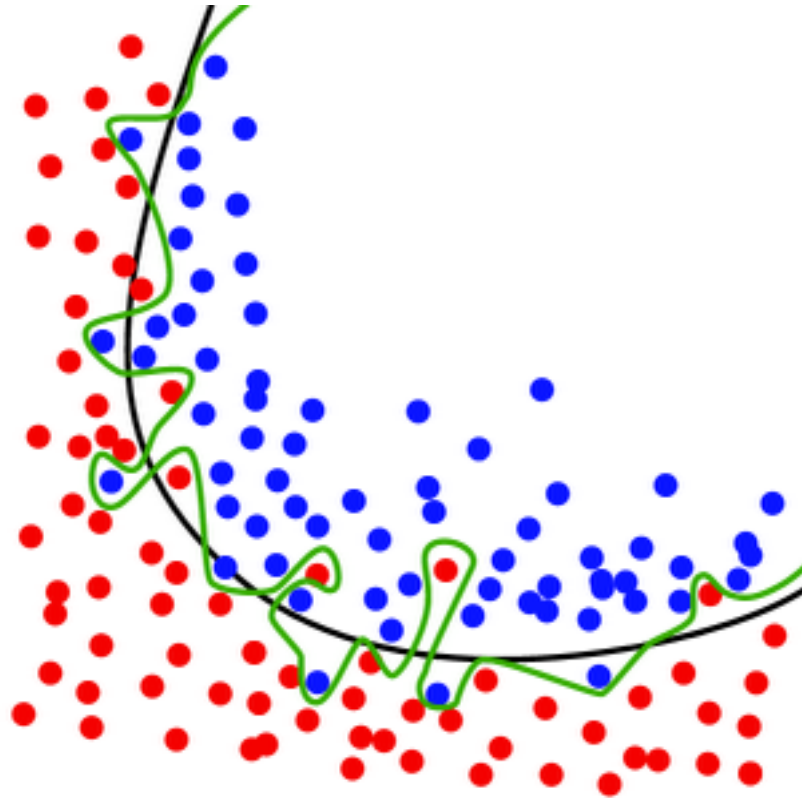


**NOTE**

This new data is called *out of sample* data.

We don't know the labels for these OOS records!

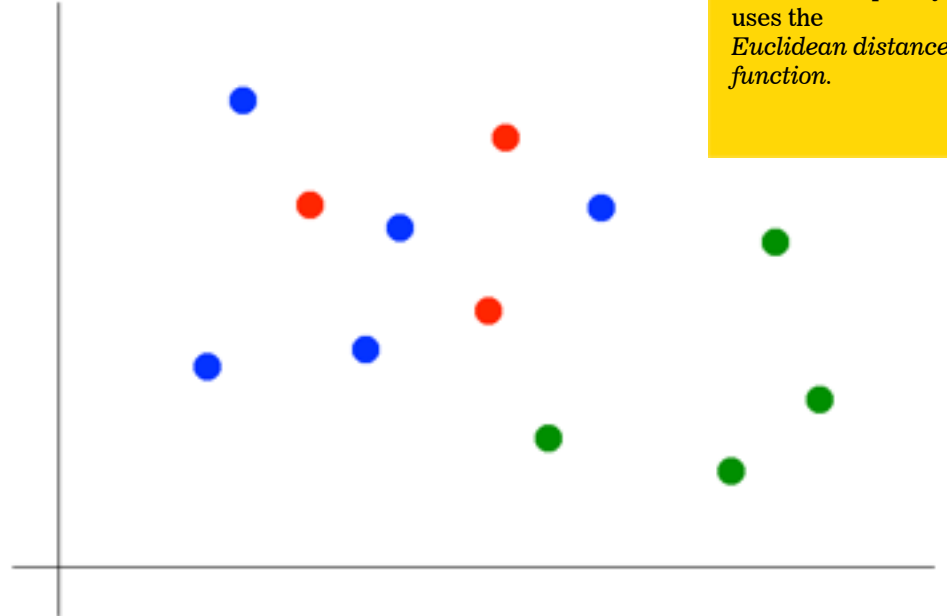




# **II. K-NEAREST NEIGHBORS**

*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*.

---

**INTRO TO DATA SCIENCE**

---

# **HANDS-ON: KNN**