
DATA SCIENCE

MACHINE LEARNING / KNN

AGENDA

- I. WHAT IS MACHINE LEARNING?
- II. SUPERVISED LEARNING
- III. UNSUPERVISED LEARNING
- IV. CLASSIFICATION WITH K-NEAREST NEIGHBORS

I. WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?

Machine Learning is a class of algorithms which are data-driven. Unlike classical algorithms, it is the data that defines a “good” answer and NOT the human

Example:

A **Non**-Machine Learning algorithm might “define” a face as having a roundish structure, two eyes, hair, nose, etc. The algorithm then looks for these “hard-coded” features in test cases.

A Machine Learning algorithm might only be given several pictures of faces and non-faces that are labeled as such. From the examples (called training set) it would “figure out” its own definition of a face.

NOT PERFECT

Machine Learning is generally *semi-automatic* meaning that intelligent decisions by humans are still necessary

EXAMPLE: FACIAL RECOGNITION: MACHINE LEARNING

Training set



Face



Not Face



Face

Test



Face?

TYPES OF MACHINE LEARNING PROBLEMS

supervised
unsupervised

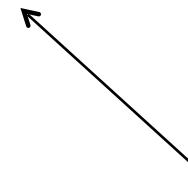
making predictions
extracting structure

TYPES OF MACHINE LEARNING PROBLEMS

supervised
unsupervised

Supervised: labeled data
Unsupervised: unlabeled data

generalization
making predictions
extracting structure



representation

TYPES OF MACHINE LEARNING PROBLEMS

supervised
unsupervised

generalization
making predictions
extracting structure

Previous
example was
supervised!

II. SUPERVISED LEARNING

SUPERVISED LEARNING

- All about making *predictions*
- List of “Predictors” X
 - Also known as features, independent variables, inputs, regressors, covariates, attributes
- “Response” y
 - Also known as outcome, label, target, dependent variable
- If y is continuous: **Regression**
 - e.g., price, blood pressure
- If y is categorical (values in a finite, unordered set): **Classification**
 - e.g., spam/ham, digit 0-9, cancer class of tissue sample
- Data are composed of “observations” (predictors and the associated response)
 - Also known as samples, examples, instances, records

EXAMPLE #1: PREDICTING NEONATAL INFECTION

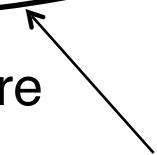
Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick



Goal: Detect subtle patterns in the data that predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc...

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear



predictors

Sample response: Did the child develop an infection? True/False

SUPERVISED LEARNING – FISHER'S IRIS

150
observations
 $(n = 150)$

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>

4 predictors ($p = 4$)

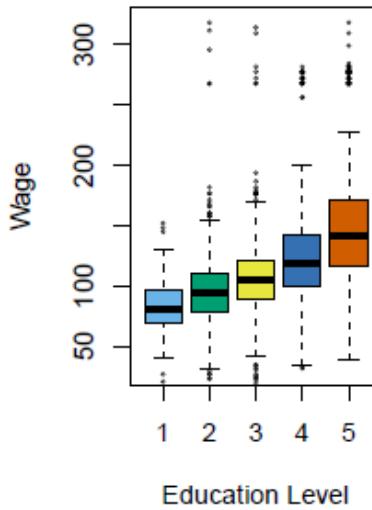
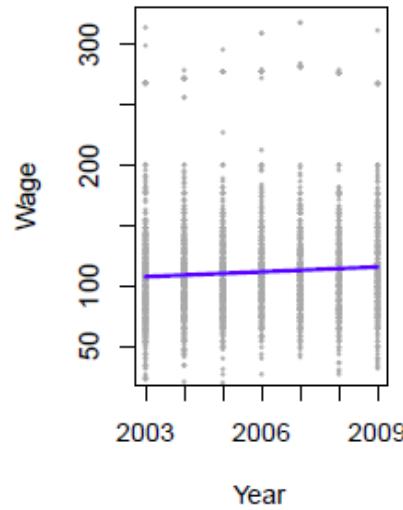
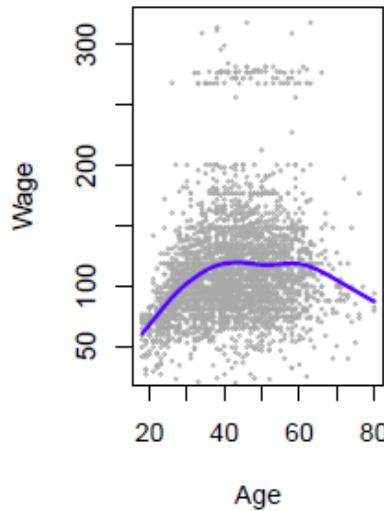
response

SUPERVISED LEARNING

- Supervised Learning uses known/labeled “**training** cases” in order to:
 - Accurately predict unseen **test** cases
 - Understand which predictors affect the response, and how
 - *Use the past to predict the future*

REGRESSION EXAMPLE

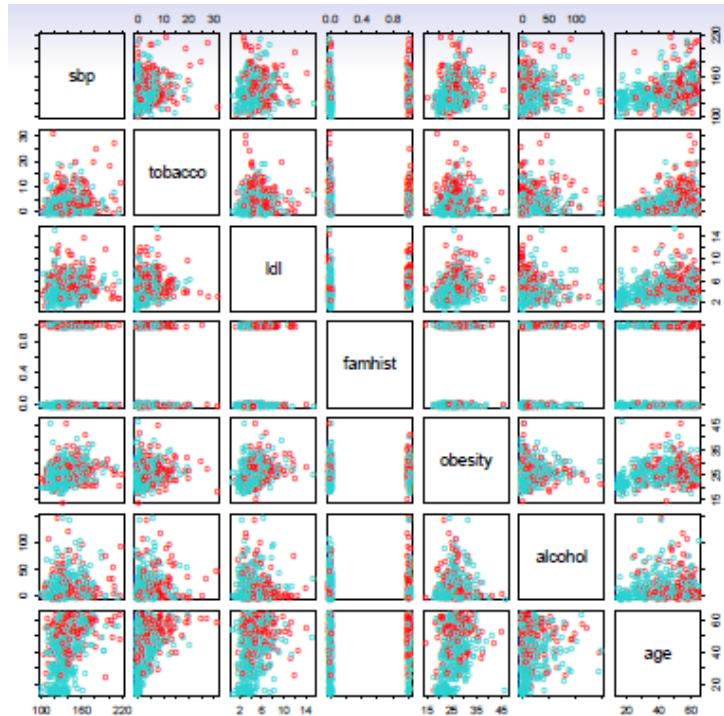
- Establish the relationship between salary and demographic variables in population survey data



Income survey data for males from the central Atlantic region of the USA in 2009

CLASSIFICATION EXAMPLE

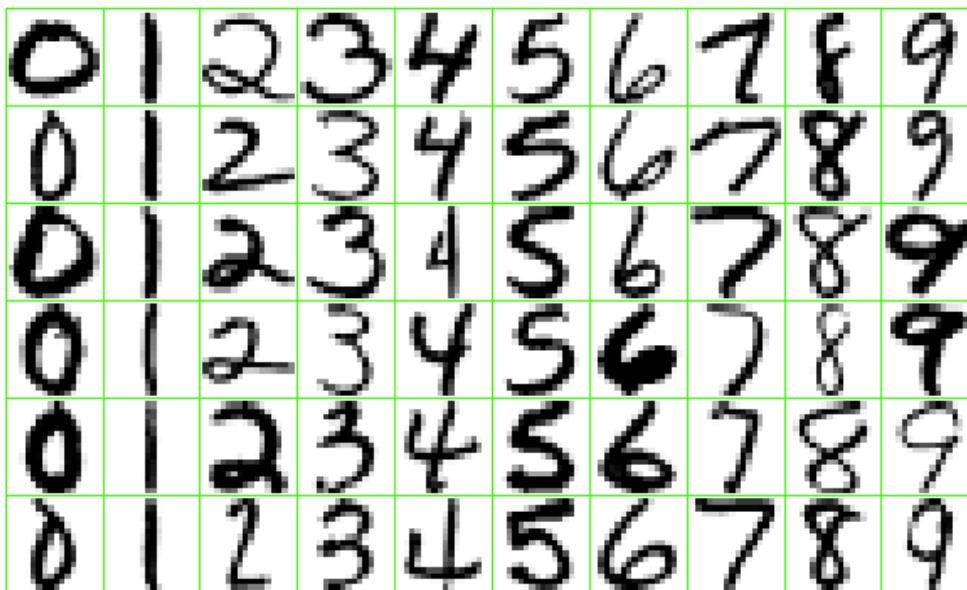
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements



Case-control sample of men from South Africa
Red = heart disease
Blue = no heart disease

CLASSIFICATION EXAMPLE

- Identify the numbers in a handwritten zip code



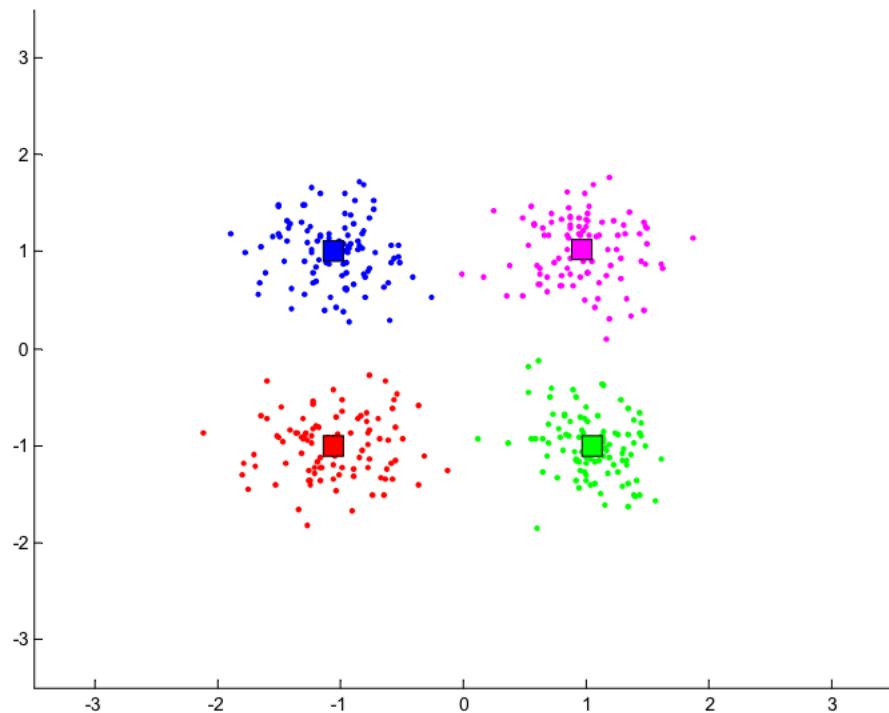
Source: <https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf>

III. UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

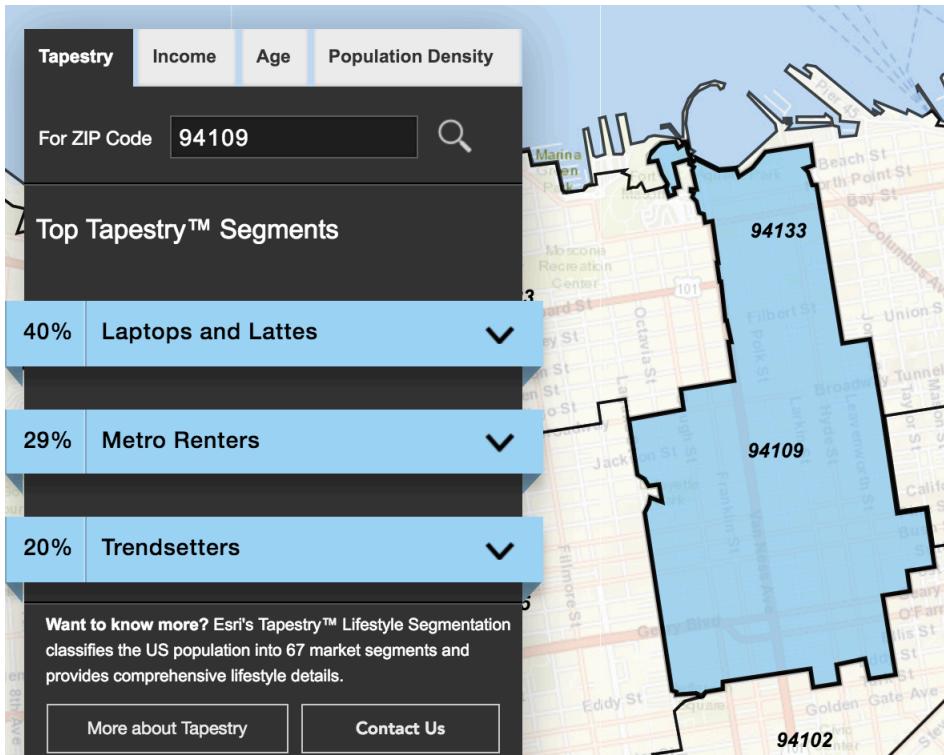
- No response variable y , just a set of predictors X
- Objective is more open:
 - Find groups of observations that behave similarly
 - Find predictors that behave similarly
 - Find combinations of features that explain the variation in the data
- Difficult to evaluate how well you are doing
- Data is easier to obtain for unsupervised learning since it can be “unlabeled” (i.e., it hasn’t been labeled with a response)
- Sometimes useful as a preprocessing step for supervised learning
- Common techniques: clustering, principal components analysis

CLUSTERING EXAMPLE



CLUSTERING EXAMPLE

- Classify US residential neighborhoods into 67 unique segments based on demographic and socioeconomic characteristics



Example of cluster: **Laptops and Lattes**

- We're affluent, well-educated singles and partner couples who love life in the big city
- Regular expenses include nice clothes, traveling, and treating ourselves to lattes at Starbucks or treatments at spas. Laptops, cell phones, and iPads are always on so we can stay connected.
- Leisure time is filled with visiting art galleries and museums; attending the theater, opera, and going to bars and clubs.

Source: <http://www.esri.com/landing-pages/tapestry/>

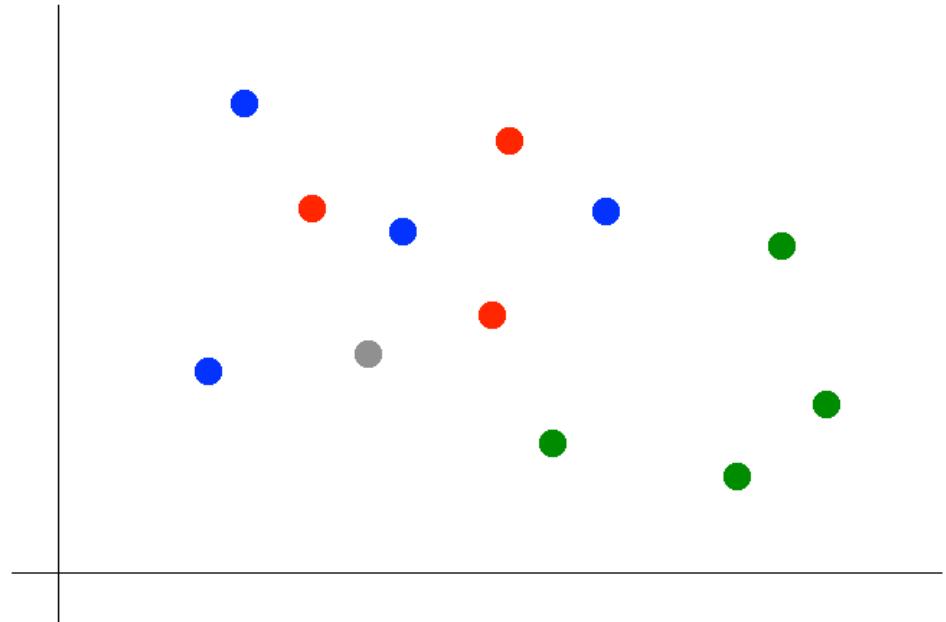
IV. CLASSIFICATION WITH K-NEAREST NEIGHBORS

CLASSIFICATION WITH KNN

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

QUESTION:

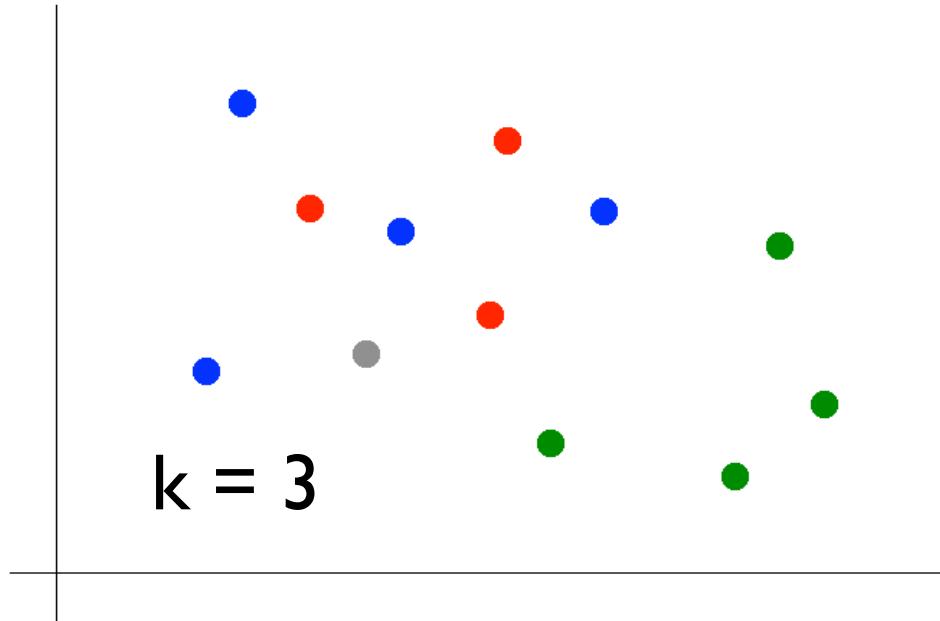
What are the predictors?
What is the response?



CLASSIFICATION WITH KNN

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

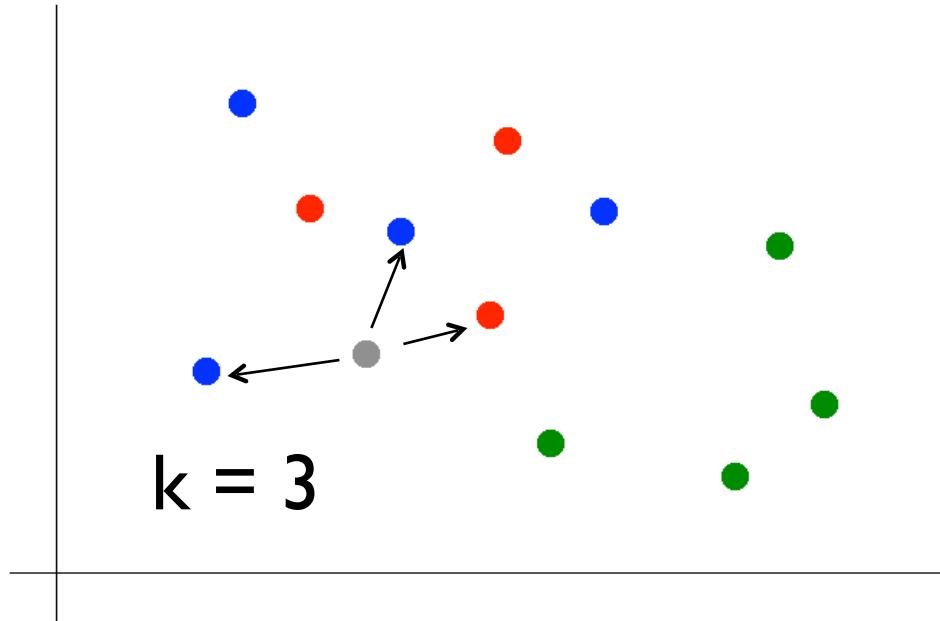
- 1) Pick a value for k.



CLASSIFICATION WITH KNN

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

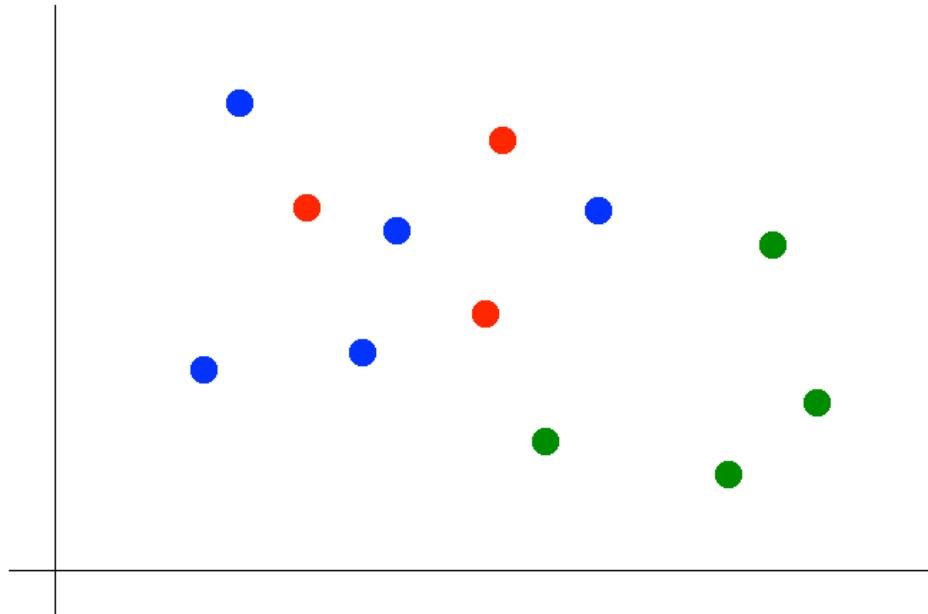
- 1) Pick a value for k.
- 2) Find colors of k nearest neighbors.



CLASSIFICATION WITH KNN

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

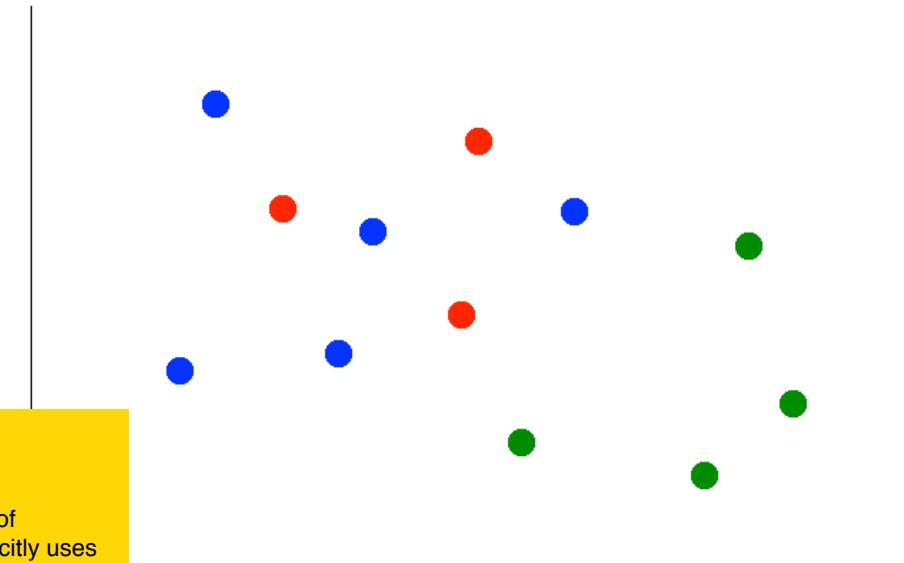
- 1) Pick a value for k.
- 2) Find colors of k nearest neighbors.
- 3) Assign the most common color to the gray dot.



CLASSIFICATION WITH KNN

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

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



NOTE:

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

CLASSIFICATION WITH KNN

Advantages of KNN:

- Simple to understand and explain
- Model training phase is fast
- Non-parametric (does not presume a “form” of the “decision boundary”)
 - Can be used to estimate many shapes of decision boundaries

Disadvantages of KNN:

- Prediction phase can be slow when number of observations or number of features is large
- Sensitive to irrelevant features (feature selection is important)

DATA SCIENCE
