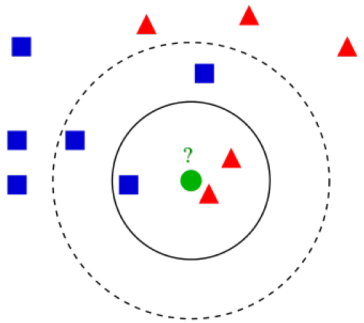


Machine Learning

KNN Classification

Jay Urbain, PhD

Topics

- What is machine learning?
- Supervised learning
- Unsupervised learning
- Classification with K-Nearest Neighbors

Data Science Process

- Ask questions – hypothesize
- Data collection
- Data exploration
- **Data modeling**
- Evaluation
- Visualization of results

What is machine learning?

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

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.

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” it’s own definition (model) of a face.

EXAMPLE: FACIAL RECOGNITION: MACHINE LEARNING

Training set

Face

Not Face

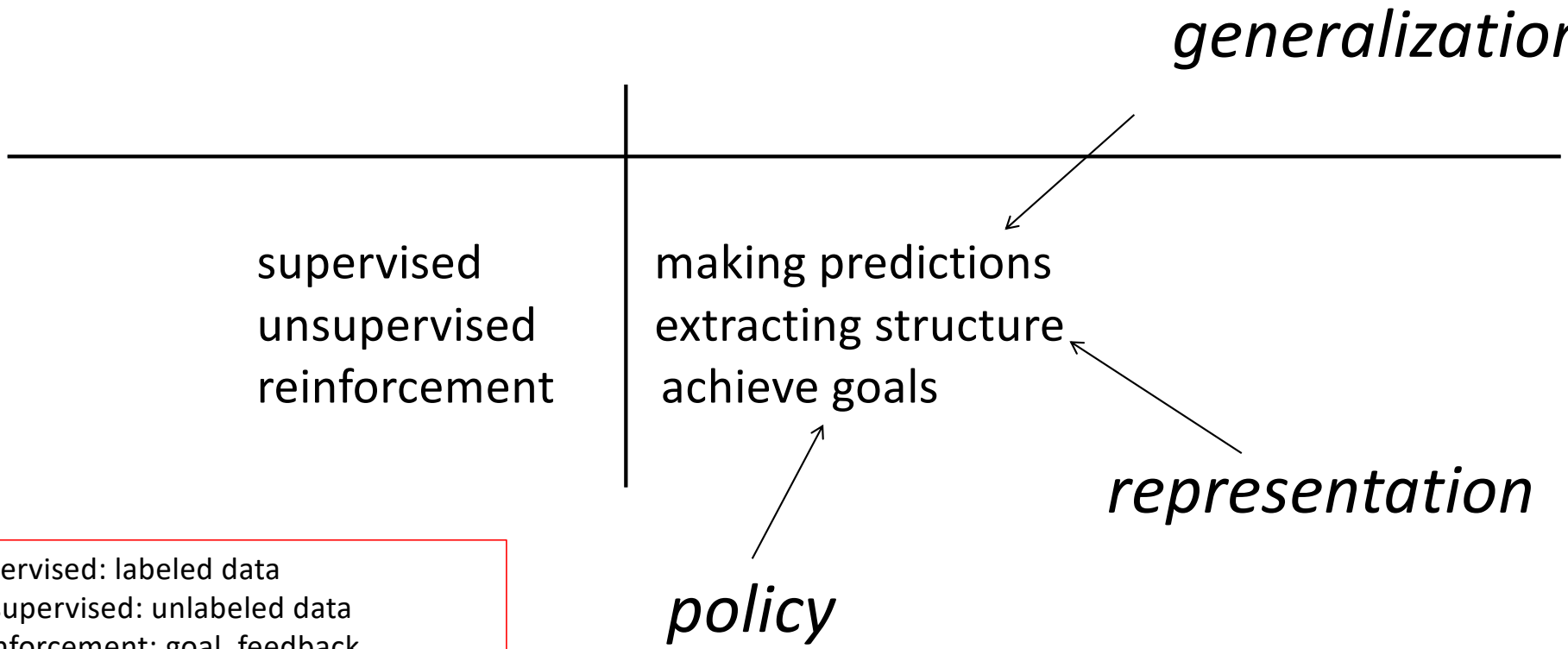
Face

Test

Face?

The core of machine learning deals with *representation* and *generalization*…

Types of Machine Learning Problems



Supervised: labeled data
Unsupervised: unlabeled data
Reinforcement: goal, feedback

Supervised learning

- Vector (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 is composed of “observations” (predictors and the associated response)
  - Also known as samples, examples, instances, records

Example: Predicting neonatal infection – binary classification

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

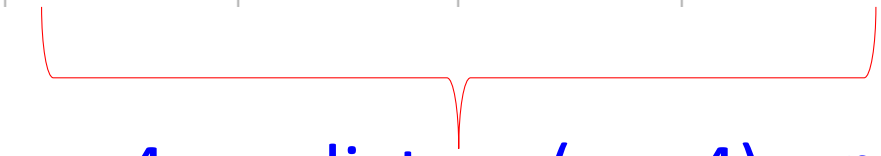
Sample response: Did the child develop an infection? True/False

Supervised learning – classification

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. 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*



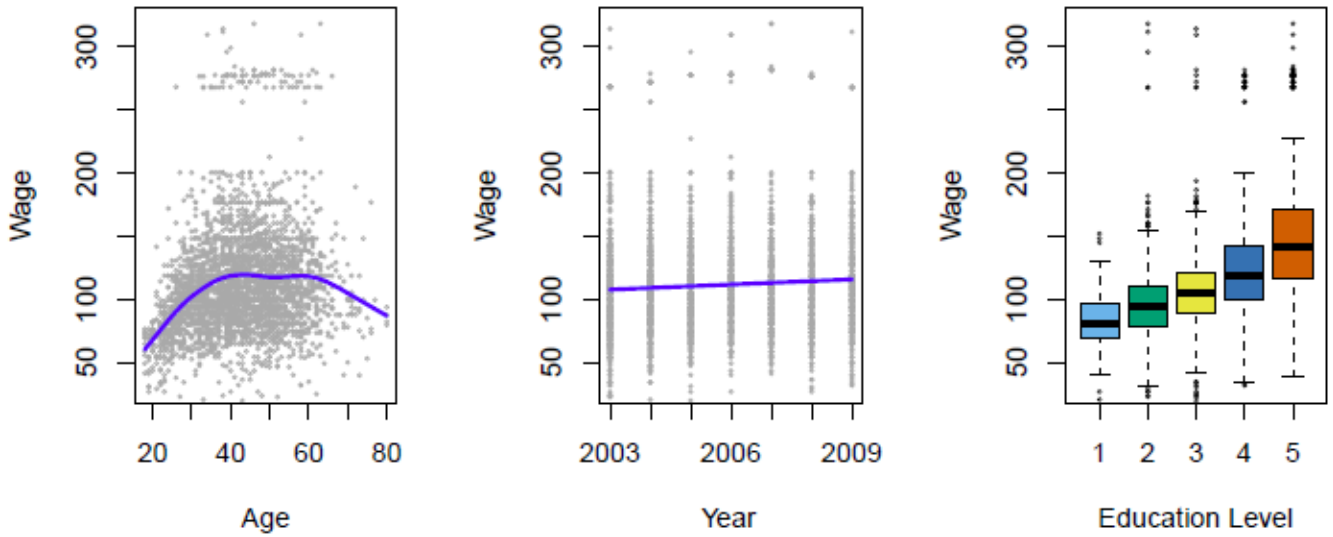
4 predictors (p = 4) response

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 to
  - Assess the quality of our predictions

Supervised learning: regression

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

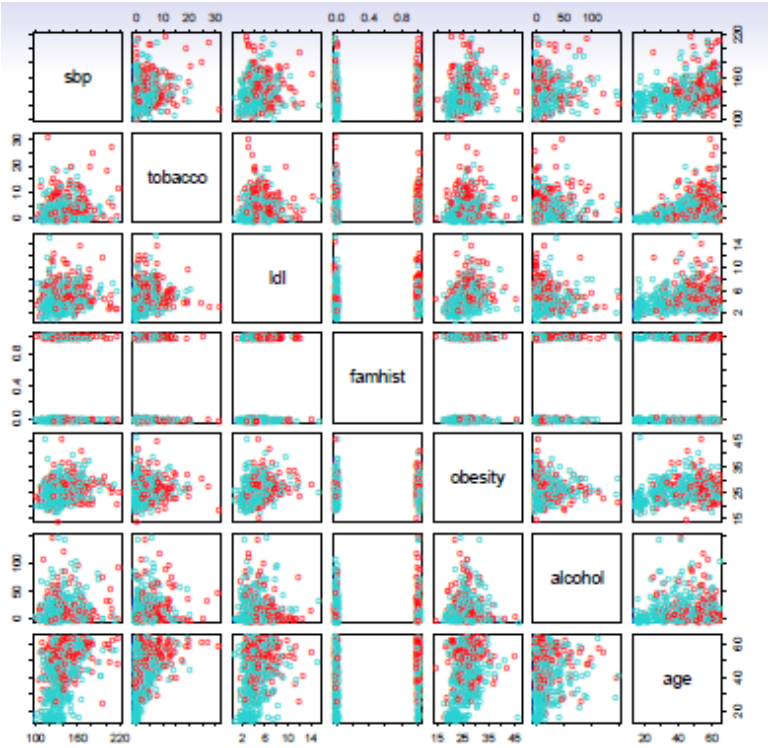


Income survey data for males from the central Atlantic region of the USA in 2009

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

Supervised learning: classification

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

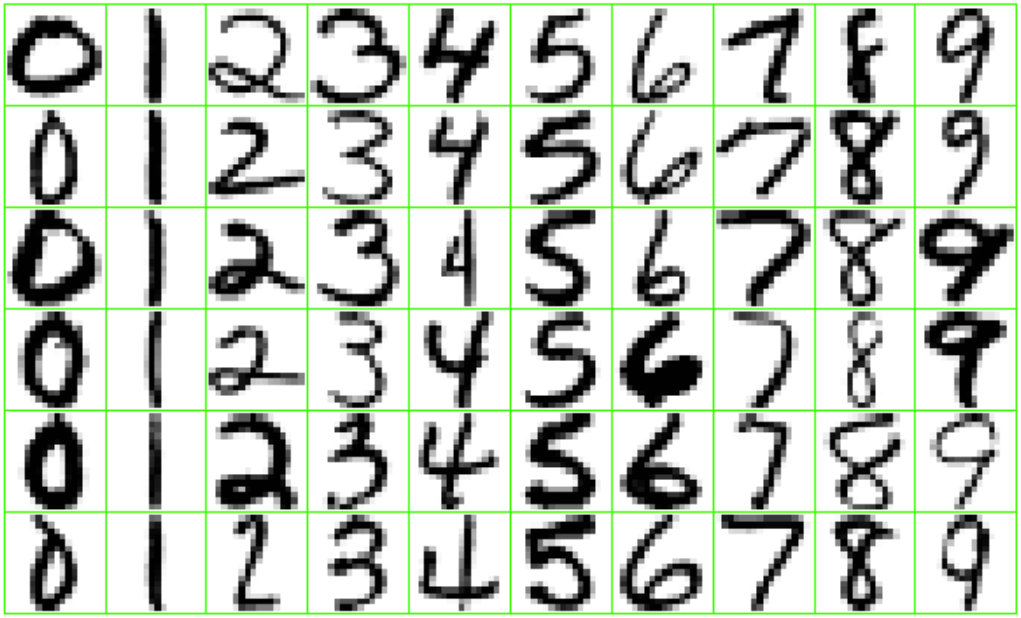


Case-control sample of men from South Africa
Red = heart disease
Blue = no heart disease

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

Supervised learning: classification example

- Identify the numbers in a handwritten zip code



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

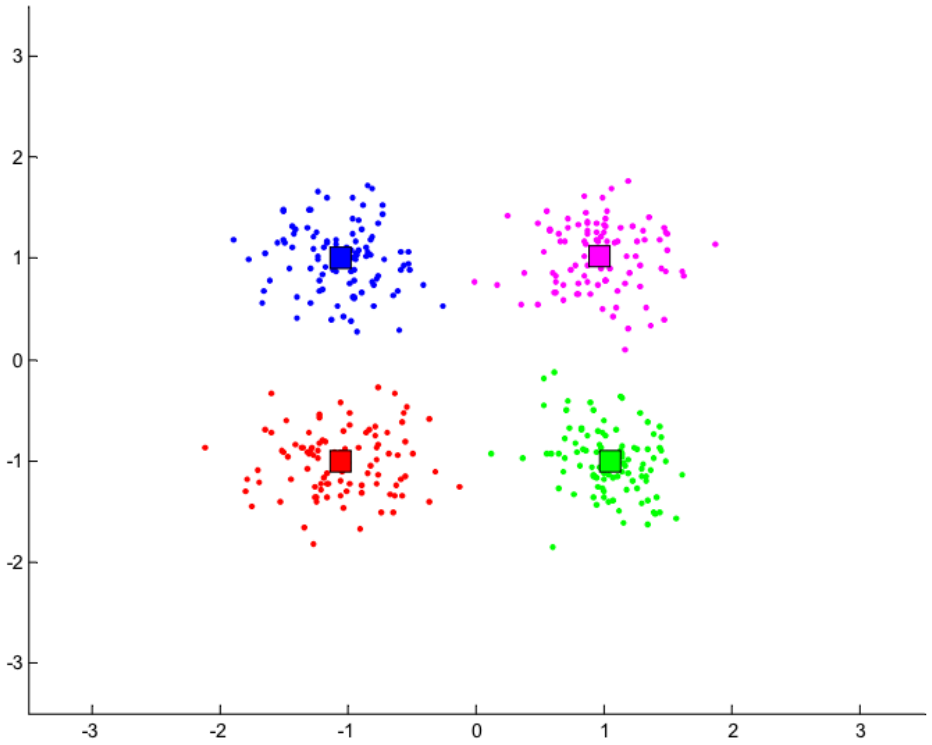
Unsupervised learning

- 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, PCA, LDA, embeddings.

Supervised vs. unsupervised learning

	continuous	categorical
supervised	regression	classification
unsupervised	dimensionality/ reduction	clustering

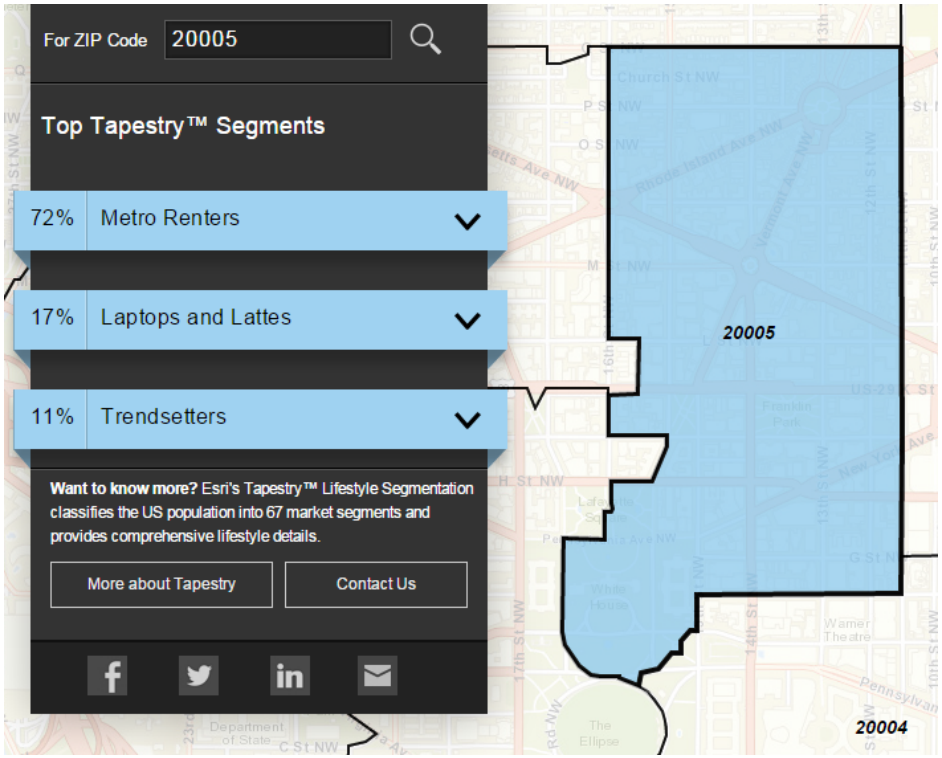
Clustering example



Source: http://people.cs.pitt.edu/~milos/courses/cs2750-Spring03/lectures/class17.pdf

Clustering example

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



Example of cluster: **Metro Renters:**

- Young, mobile, educated, or still in school
- Live alone or with a roommate
- Works long hours
- Buys groceries at Whole Foods and Trader Joe's
- Shops at Banana Republic, Nordstrom, and Gap
- Loves yoga, go skiing, and attend Pilates sessions.

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

Classification problem

Q: How does a classification problem work?
A: Data in, predicted labels out.

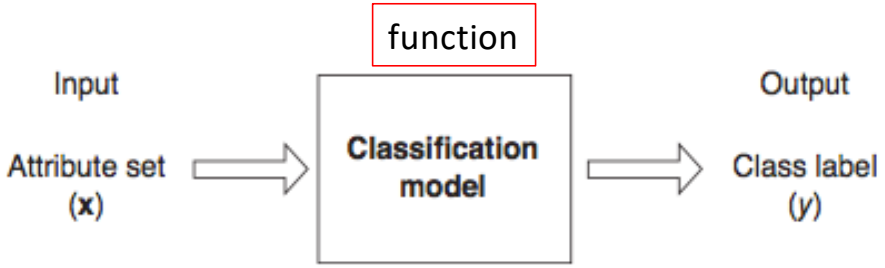


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

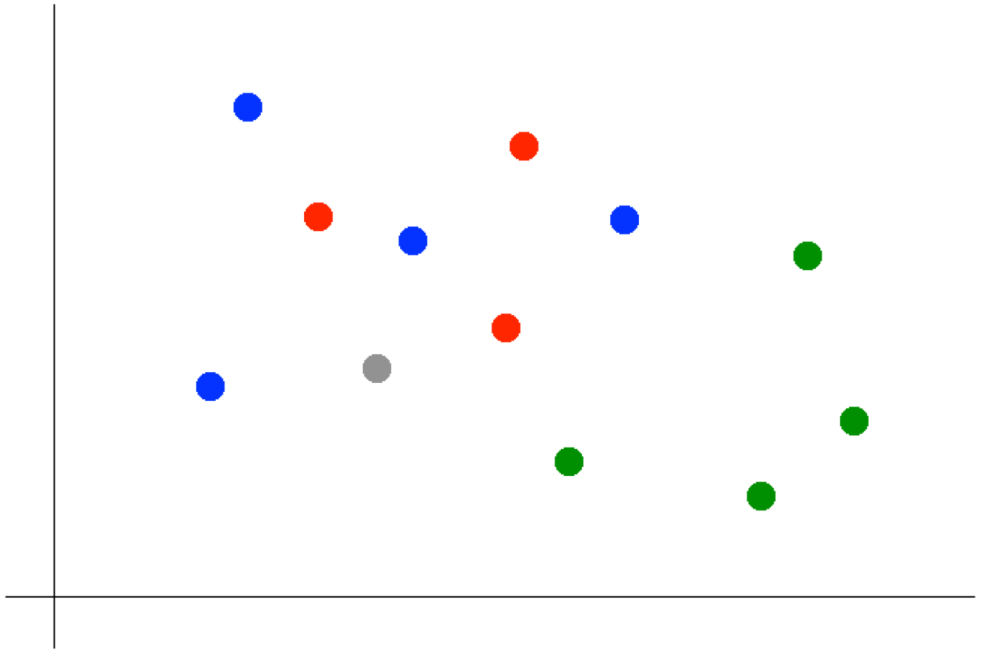
Source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

Classification with KNN

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

What are the predictors?

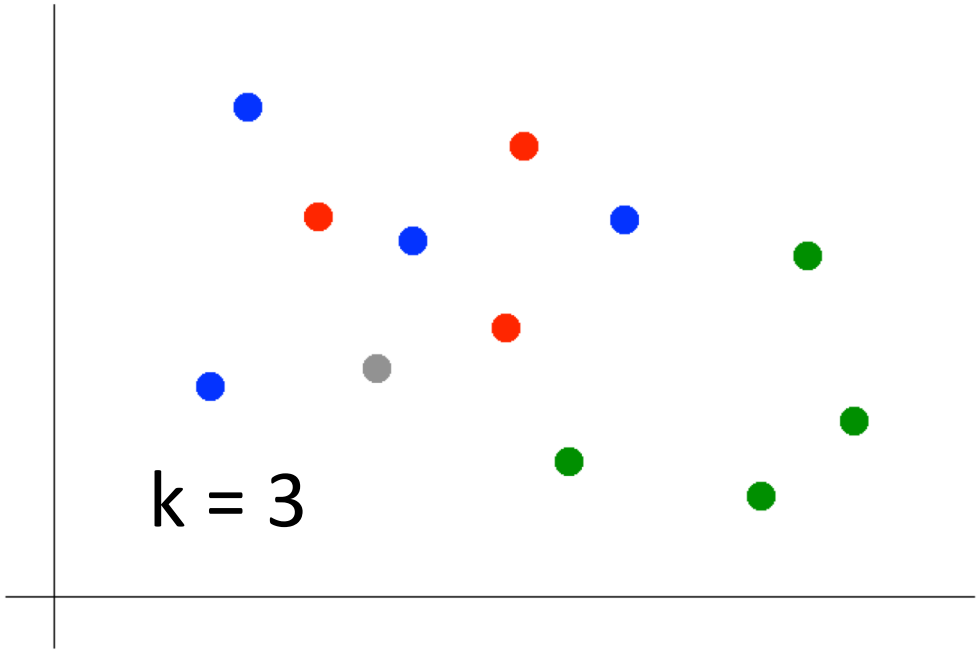
What is the response?



Classification with KNN

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

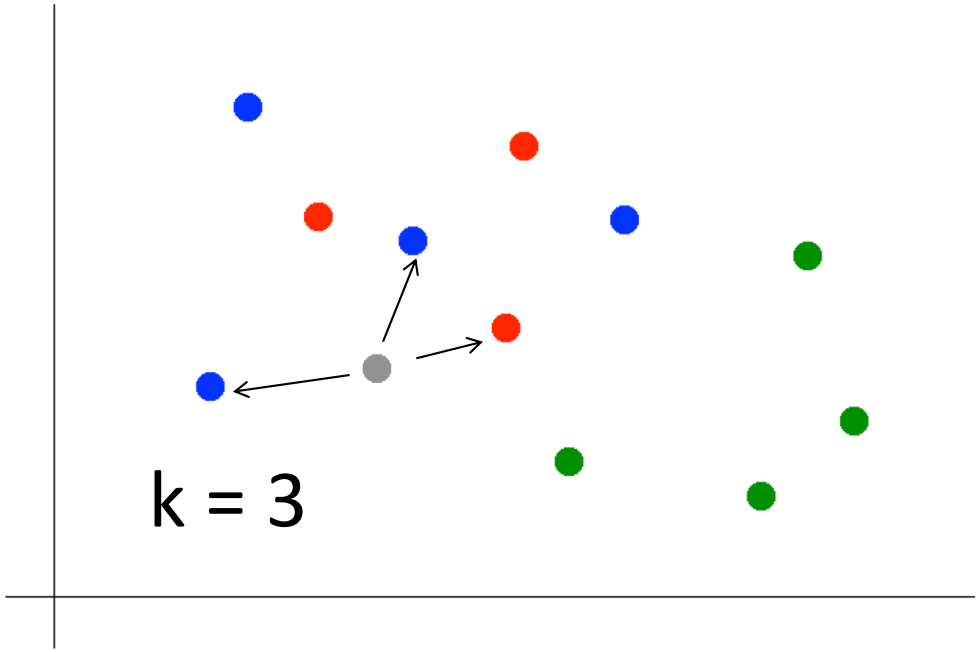
1) Pick a value for **k**.



Classification with KNN

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

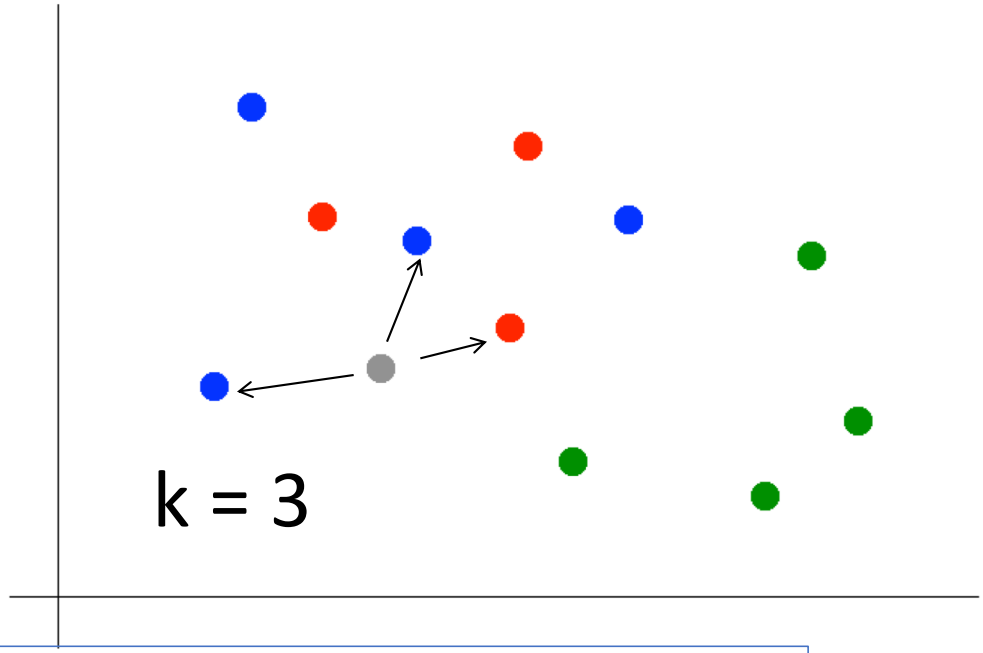
1) Pick a value for **k**.
2) Find colors (classifications) of **k** nearest neighbors.



Classification with KNN

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

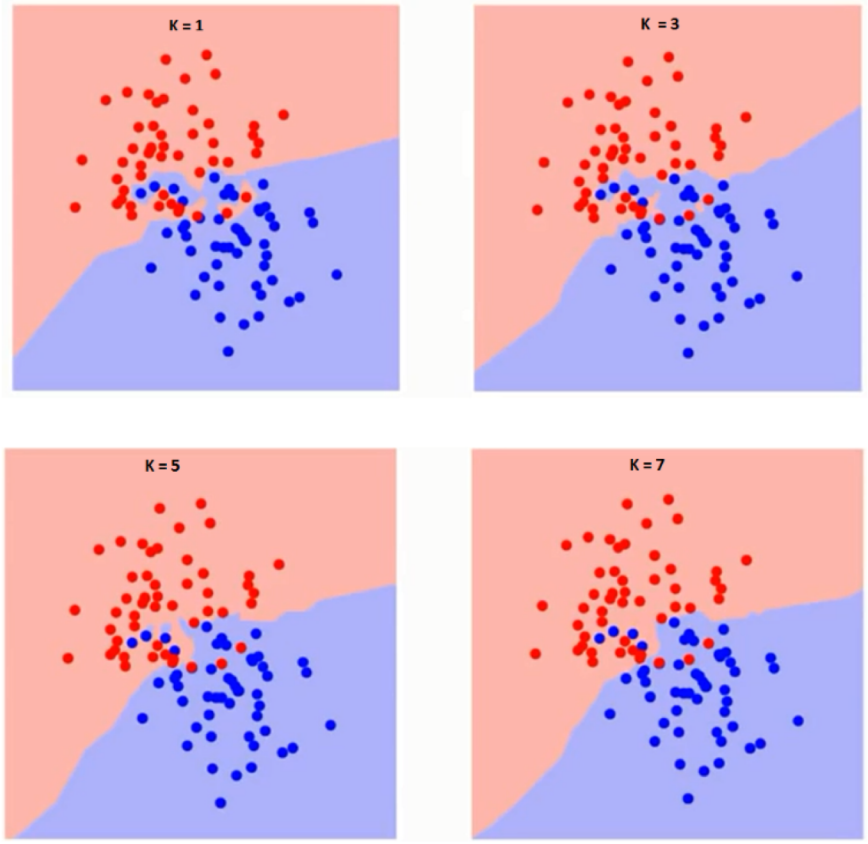
1) Pick a value for k.
2) Find colors (classifications) of **k** nearest neighbors.
3) Assign the most common (mode) color to the gray dot.



Note: nearest is defined by a distance function, e.g., Euclidean distance based on predictors

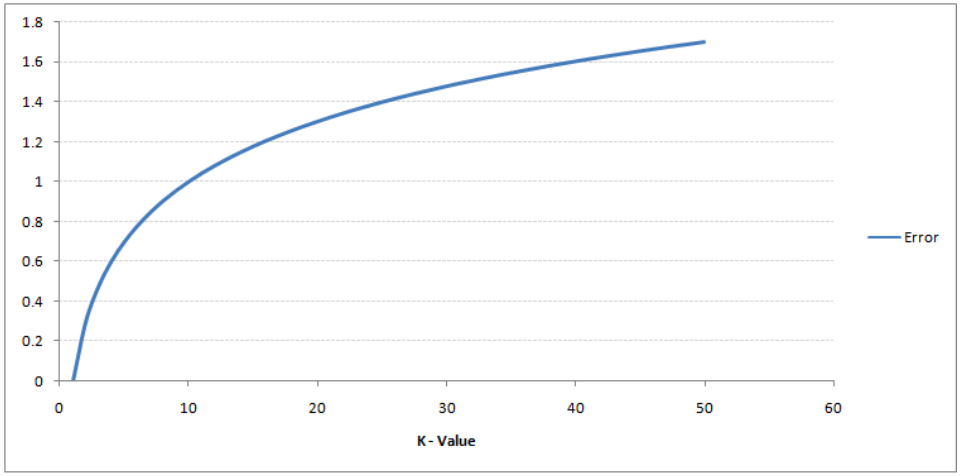
Picking a value for K

- K value determines decision boundary.
- Decision boundary becomes smoother as k is increased. Less likely to overfit.
- Increasing K to infinity will set all values to either red or blue.

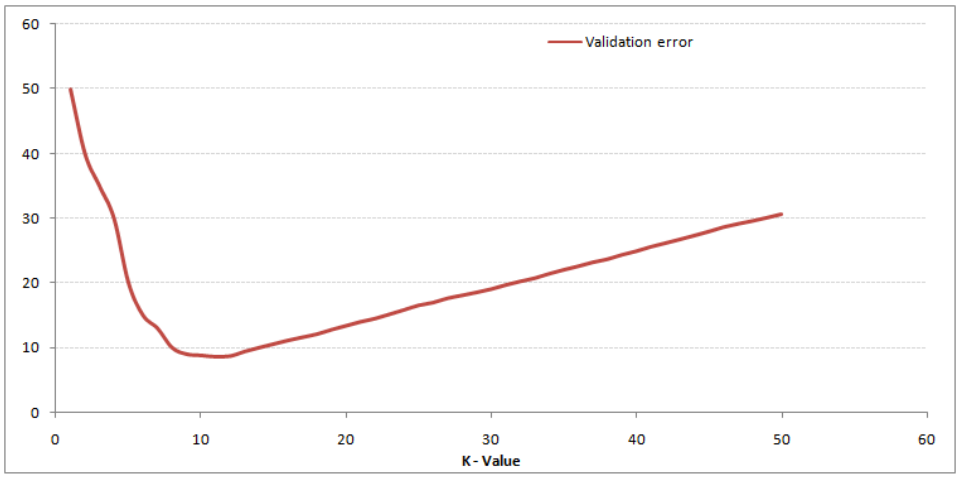


Training vs. Validation

- Blue – training error
- Red – validation error



- Note:
- At k=1 you are overfitting the boundary
- K eventually decreases until you get a minima.



Pseudo Code of KNN

- Load the data
- Initialize the value of k
- For getting the predicted class, iterate from 1 to total number of training data points
  - Calculate the distance between test data and each row of training data. E.g., Euclidean distance, Chebyshev, cosine, etc.
  - Sort the calculated distances in *ascending* order based on distance values
  - Get top k rows from the sorted array
  - Get the most frequent class of these rows
  - Return the predicted class

KNN with scikit-learn

```
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=3)
neigh.fit(data.iloc[:,0:4], data['Name'])

# Predicted class
print(neigh.predict(test))

-> ['Iris-virginica']

# 3 nearest neighbors
print(neigh.kneighbors(test)[1])
-> [[141 139 120]]
```

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”)

Disadvantages of KNN:

- Prediction phase can be slow when **n** is large
- Sensitive to irrelevant features