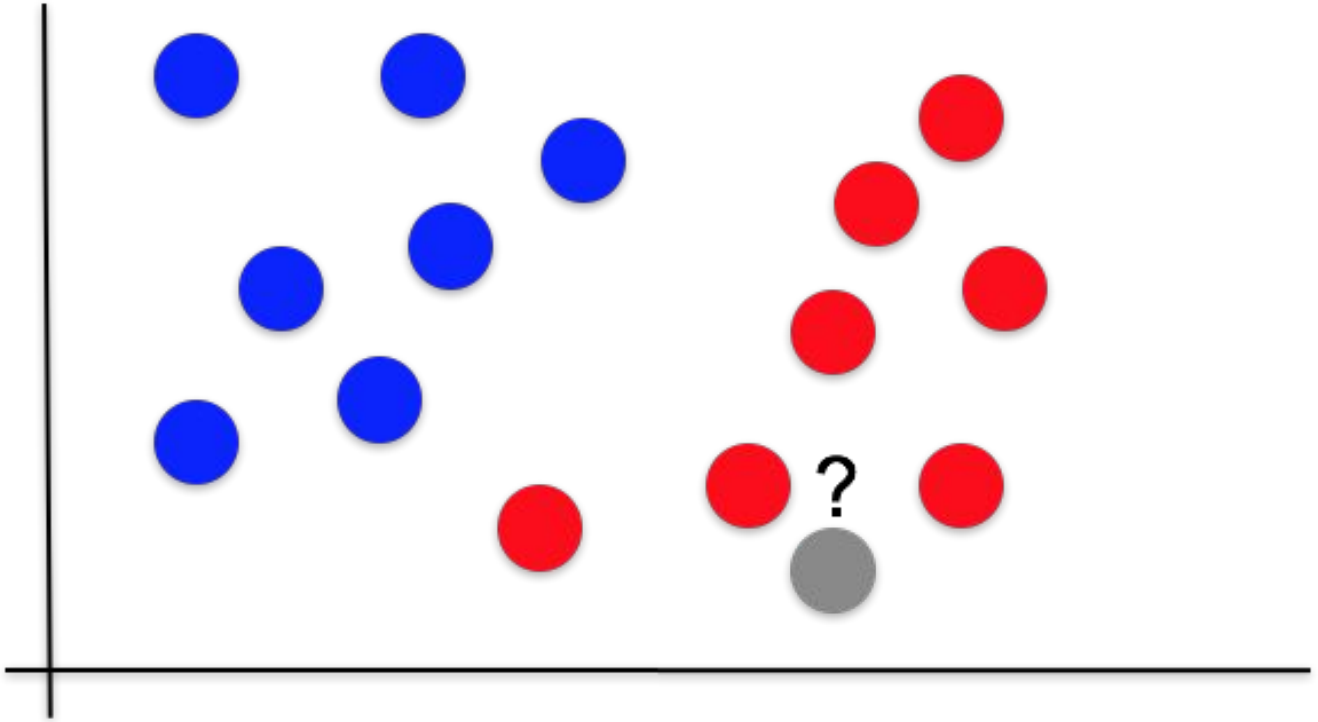
UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

k-Nearest Neighbors

Machine Learning

Graphical Overview — Example 1



Algorithmic Overview

- Inputs:
  - Training feature vectors (train_x) and training labels (train_y)
  - Neighbourhood hyperparameter (k)
  - Testing feature vectors (test_x)
- Outputs: predicted testing labels (hypotheses)
- Fit function is a NOOP
- Predict function:

```
foreach t ∈ test_x:
    points = sort(nearest_points(t))
    neighbours = get_top(k, points)
    hyp = majority_class(neighbours)
    hypotheses.append(hyp)
return hypotheses
```

How to Choose k?

- Small value = “noise” in training data influence outcome
- Typically:
  - Cross validation against dev set
  - Odd numbers when classes = 2
  - Start with sqrt(n)
  - Use cross-validation and find an “elbow”

Advantages & Disadvantages

- Advantages
  - Non-parametric: makes no assumptions about data
  - Insensitive to outliers
  - Easy to interpret output
- Disadvantages
  - Lazy algorithm: requires all data to make predictions — i.e. requires a lot of memory
  - Computationally expensive and has large memory footprint
  - Distance calculation is (often) not relative — i.e. differences on a small scale appear to be closer even when they’re not
- Other random stuff
  - k-NN is also a regression algorithm, not covered here
  - Subject to curse of dimensionality