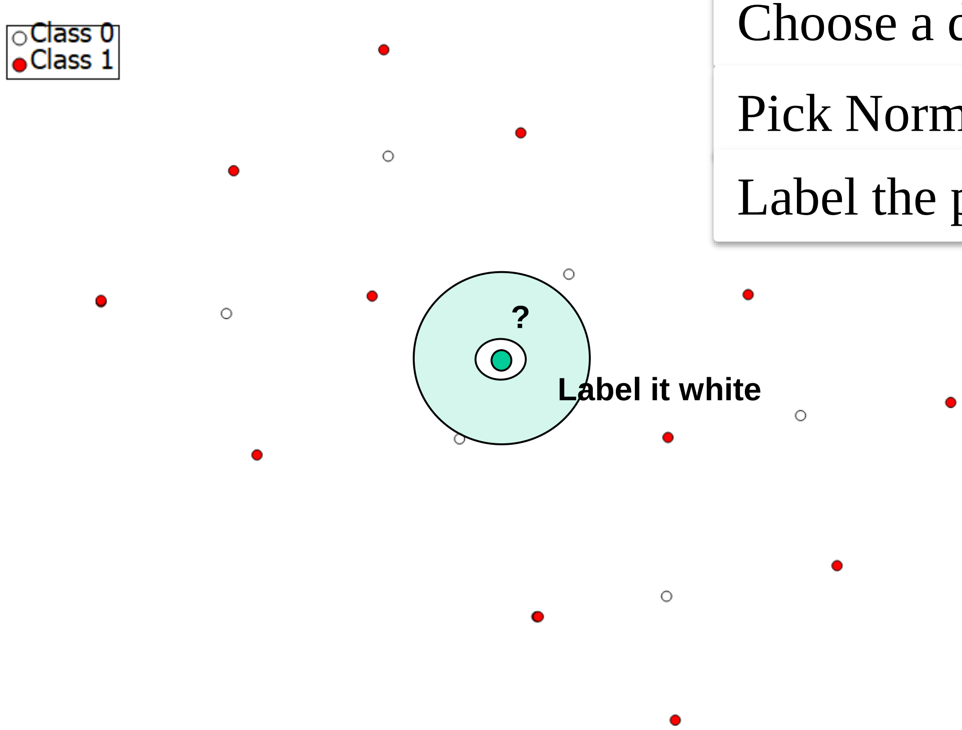
APPLIED MACHINE LEARNING

Classification with K-nearest Neighbors (KNN)

Nearest Neighbor Classifier

- One of the simplest of all machine learning classifiers
- Simple idea: label a new point with the same label as that of the closest known point



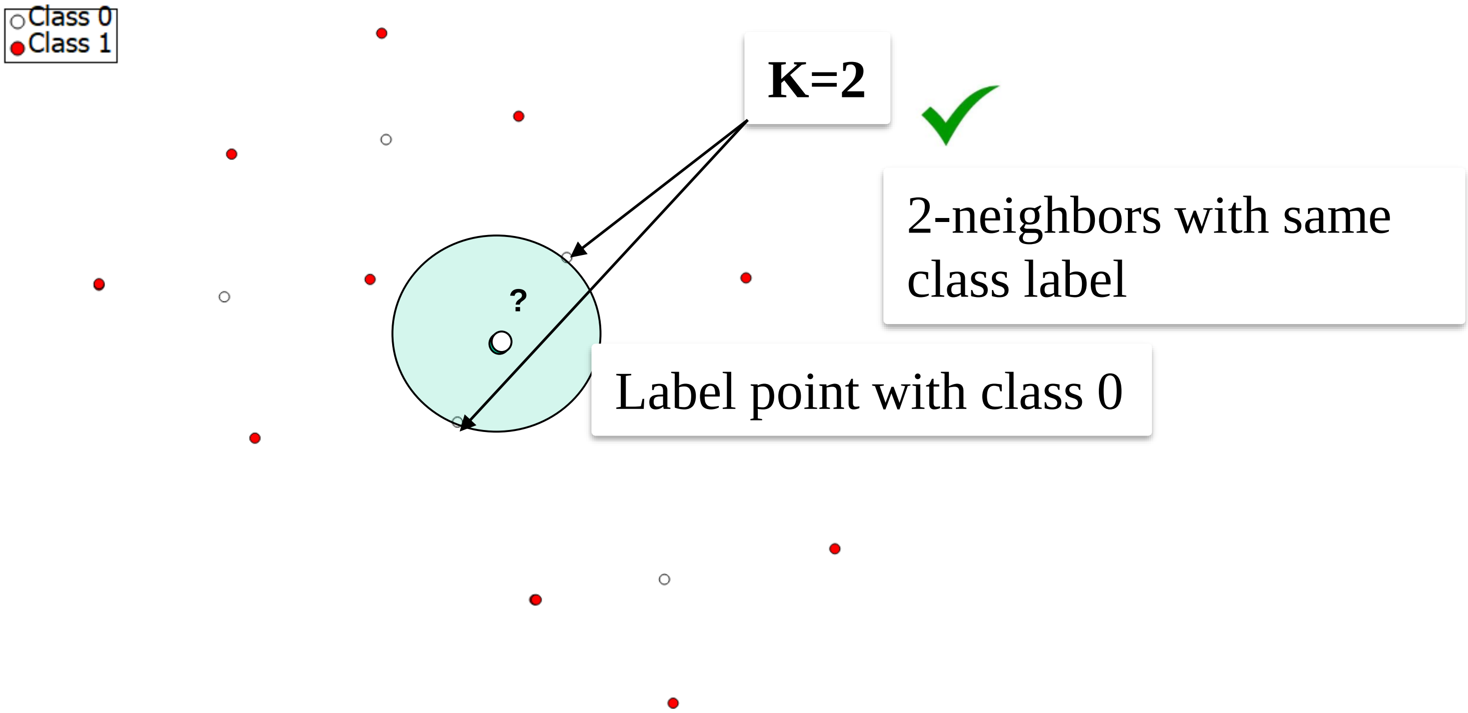
Choose a distance measure

Pick Norm-2

Label the point with class 0

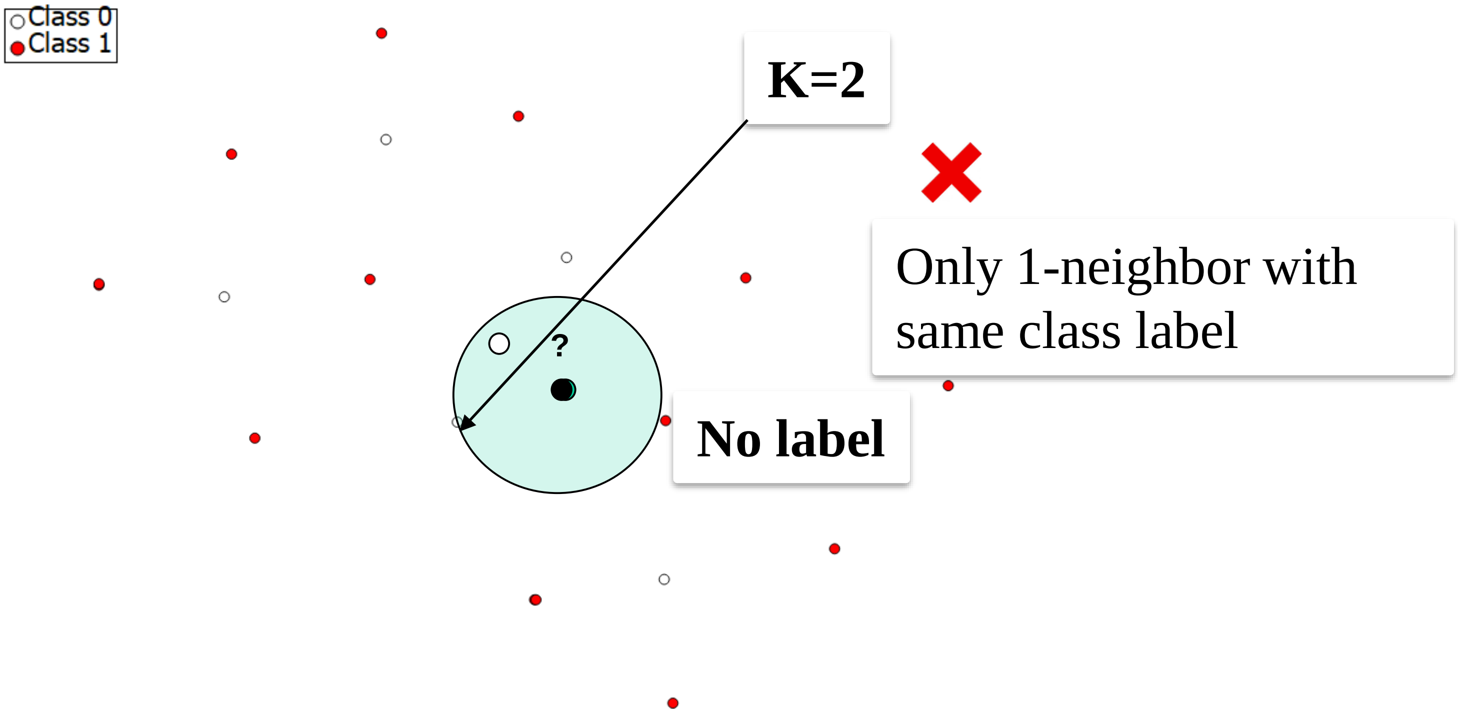
Nearest Neighbors Classifier

- Label a new point with the same label as that of the closest known group of data points
- K=number of neighbors



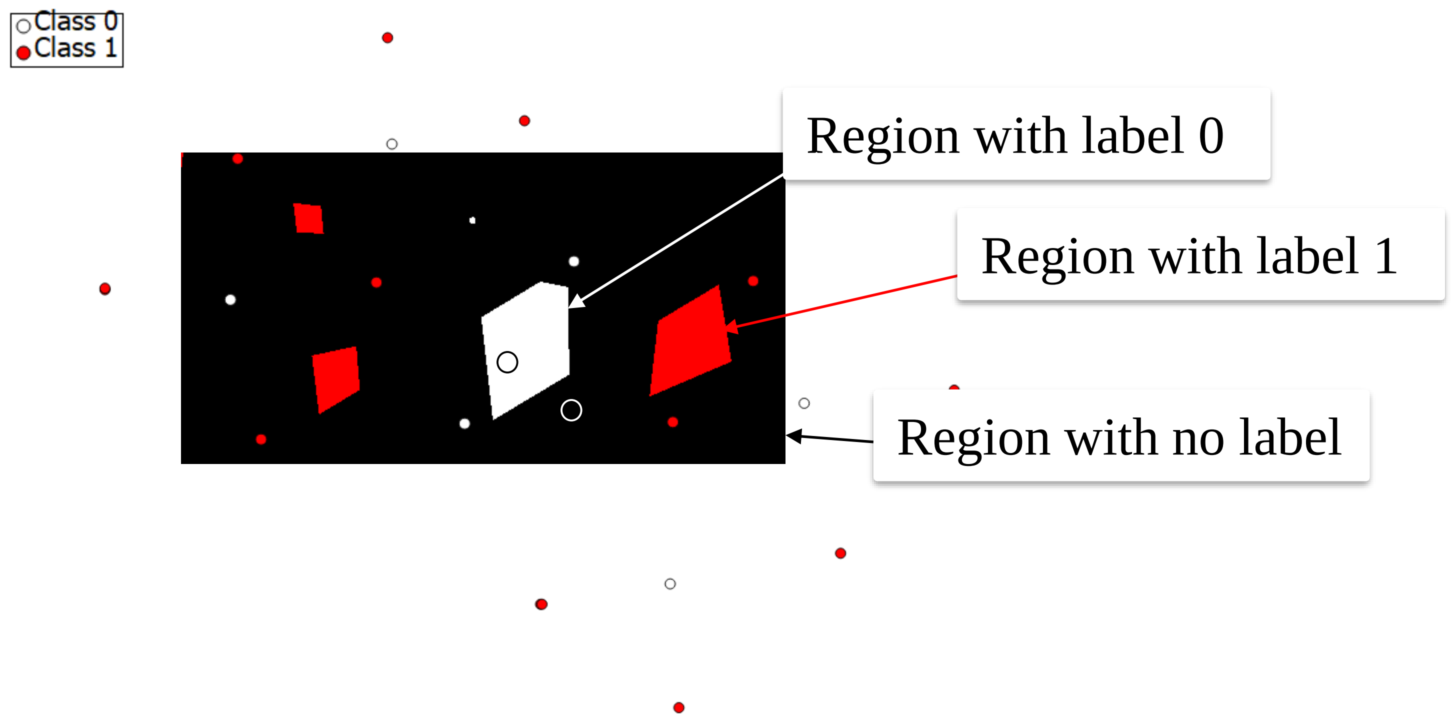
Nearest Neighbors Classifier

- Label a new point with the same label as that of the closest known group of data points
- K=number of neighbors



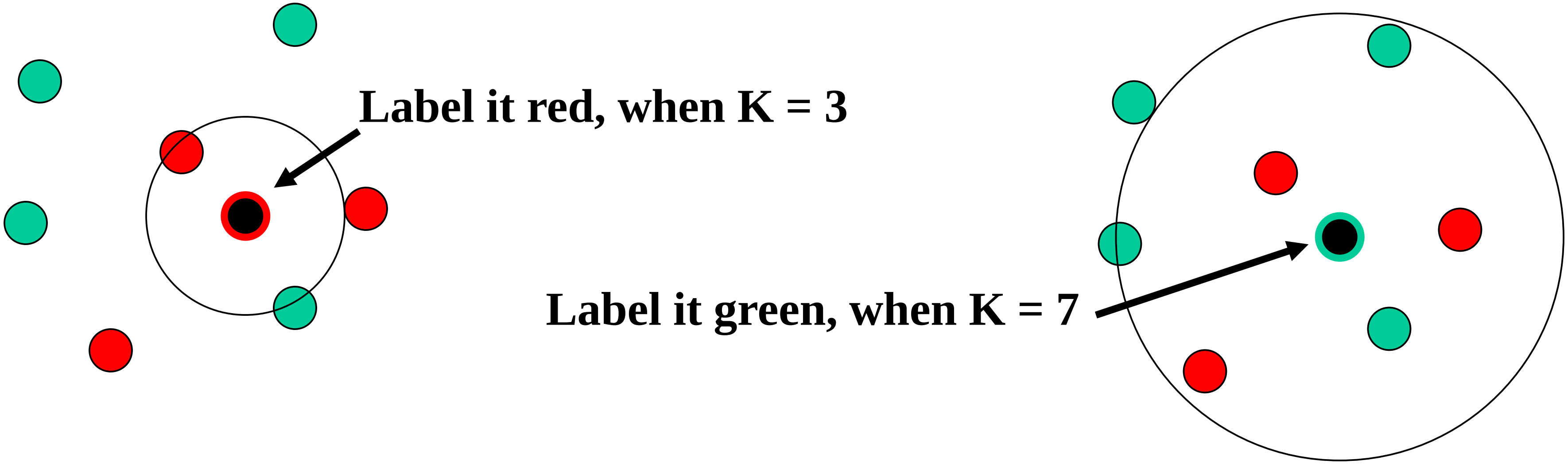
Nearest Neighbors Classifier

- Label a new point with the same label as that of the closest known group of data points
- K=number of neighbors



Nearest Neighbors Classifier

- Probabilistic approach to smooth away noise in the labels
- A new point is now assigned the most frequent label of its *K* nearest neighbors



Classification with K- Nearest Neighbors (K-NN)

Probability to be classified with label of class c :

$$p(y=c \mid x, K) = \frac{N_{c,K}}{K}$$

K: : Number of nearest neighbors

$N_{c,K}$: Number of samples of class c included in K nearest neighbors

Classification with k- Nearest Neighbors (K-NN) Summary

Algorithm:

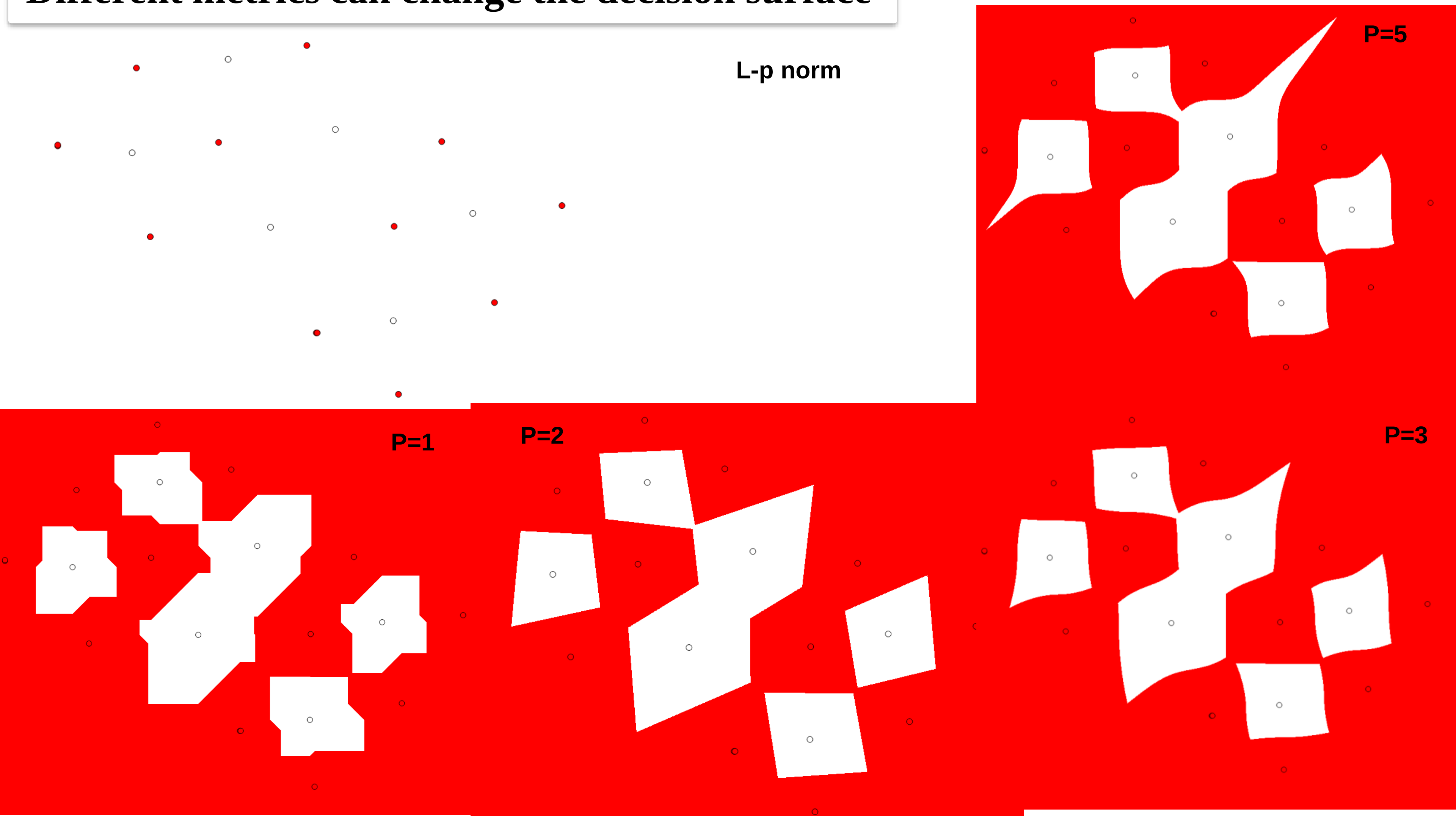
- Pick hyperparameters: K, Distance metric
- Training: No training! Save simply dataset
- Testing: x sample to be classified
  1. Calculate distance between x and all the other samples
  2. Pick the K samples closest to x
  3. $N_{c,K}$ → Count number of samples of each class *c* in K
  4. Get probability of each class:

     $$p(y=c\,|\,x,K)=\frac{N_{c,K}}{K}$$

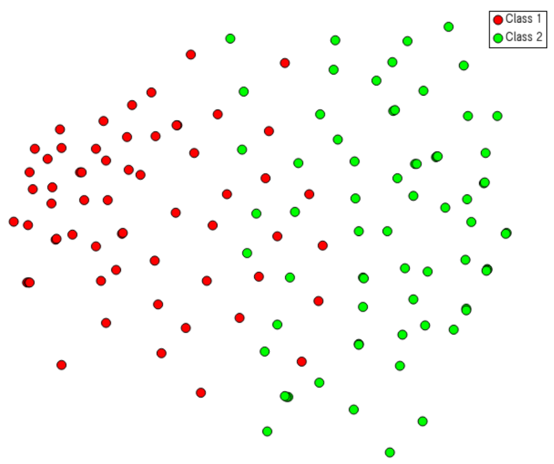
  5. Assign label of most likely class

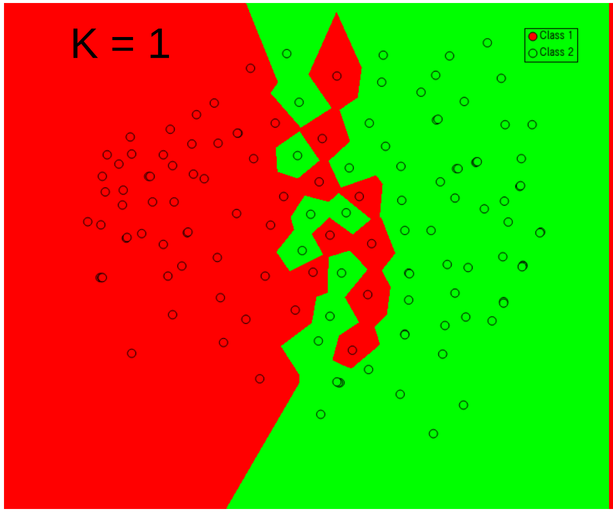
Classification with k- Nearest Neighbors (K-NN)
Impact of metric

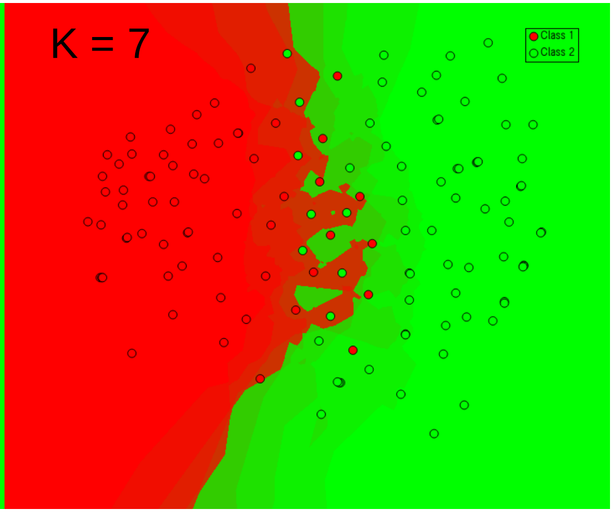
Different metrics can change the decision surface

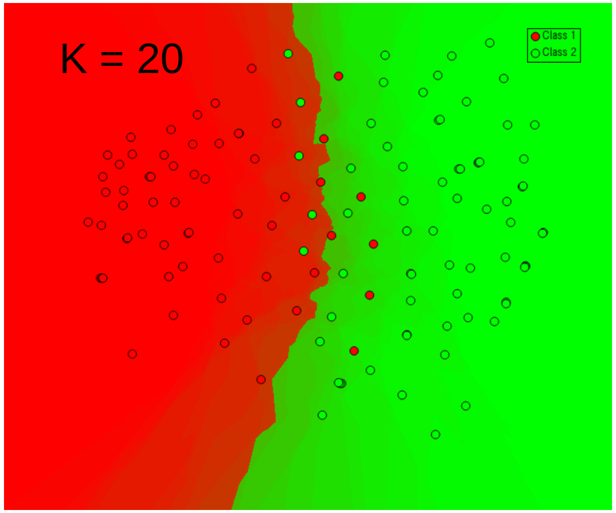


Classification with k- Nearest Neighbors (K-NN)
Impact of K selection









- Large K:
  - ➔ Makes KNN less sensitive to noise
- Small K:
  - ➔ Allows capturing finer structure of space
- Pick K not too large, but not too small (depends on data)

Classification with k- Nearest Neighbors (K-NN) Summary

Advantages:

- Simplicity
- No assumptions regarding the distribution of data
- **Does not need a model for data distribution**

Disadvantages:

- Curse of dimensionality (*stores the entire dataset)*
- High computational complexity *(pass through the entire dataset for each classification; prohibitive for large data sets.)*
- Large Impact of choice of K and of metric
- **Does not learn a model**

Possible Improvements:

- *Weighting* examples from the neighborhood
- Measuring "*closeness*"
- Finding "close" examples in a large training set *quickly*