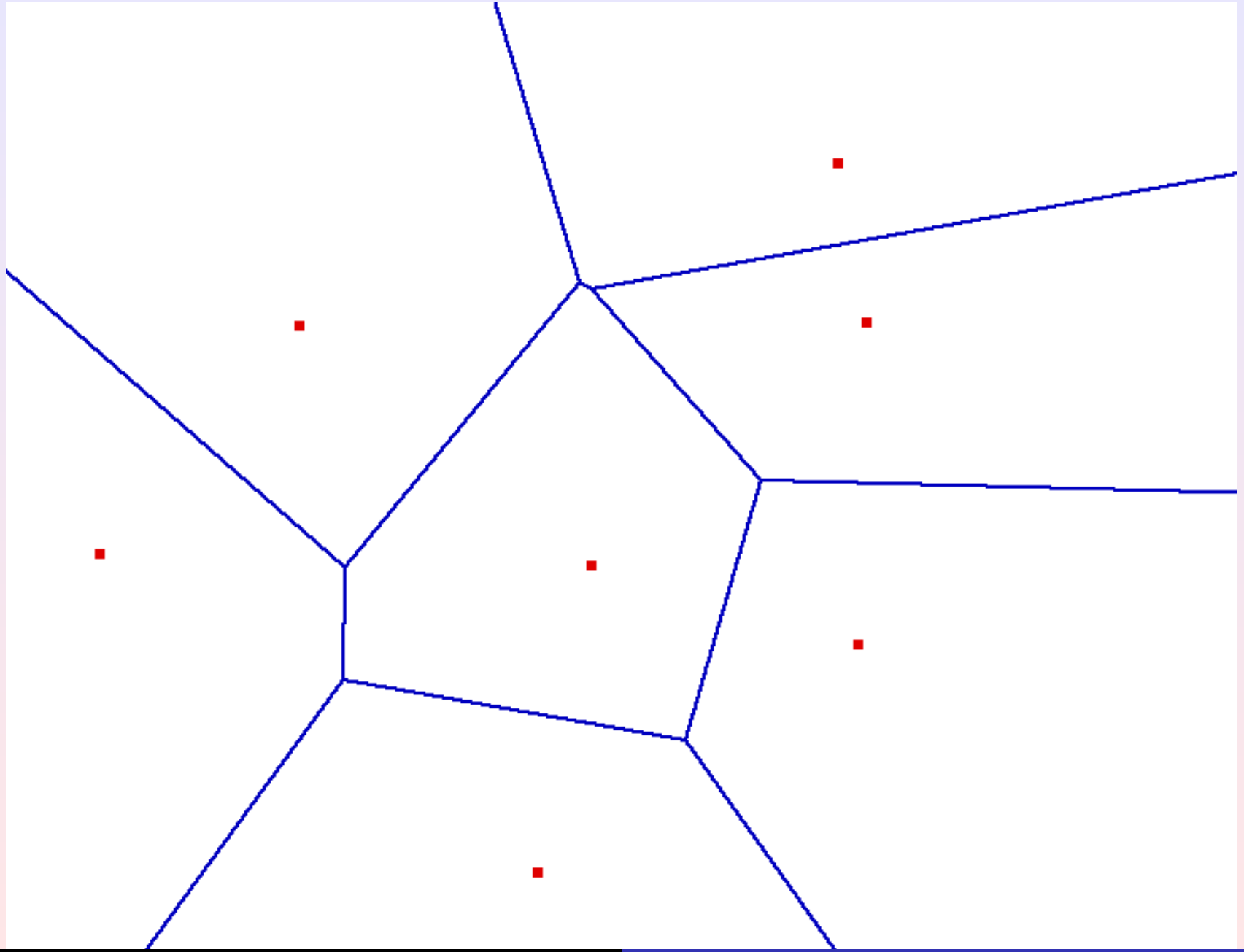
k-Nearest Neighbour Learning

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

1-Nearest Neighbour classification



Measuring the distance between instances

Metric or *distance* definition

Given a set $\mathcal{X}$, a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ is a *metric* for $\mathcal{X}$ if for any $x, y, z \in \mathcal{X}$ the following properties are satisfied:

reflexivity $d(x, y) = 0 \quad iff \quad x = y$

symmetry $d(x, y) = d(y, x)$

triangle inequality $d(x, y) + d(y, z) \geq d(x, z)$

E.g. Euclidean distance in $\mathbb{R}^n$

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

k-Nearest Neighbour classification

The algorithm

for all test examples x **do**
  for all training examples (x_i, y_i) **do**
    compute distance $d(x, x_i)$
  end for
  select the k-nearest neighbours of x
  return class of x as majority class among neighbours:

$$argmax_y \sum_{i=1}^{k} \delta(y, y_i)$$

end for

Note

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

k-Nearest Neighbour regression

The algorithm

for all test examples x **do**
  for all training examples (x_i, y_i) **do**
    compute distance $d(x, x_i)$
  end for
  select the k-nearest neighbours of x
  return the average output value among neighbours:

$$\frac{1}{k} \sum_{i=1}^{k} y_i$$

end for

Characteristics of k-nearest neighbour learning

instance-based learning the model used for prediction is calibrated for the test example to be processed

lazy learning computation is mostly deferred to the classification phase

local learner assumes prediction should be mainly influenced by nearby instances

uniform feature weighting all features are uniformly weighted in computing distances

Distance-weighted k-nearest neighbour

Classification

$$argmax_y \sum_{i=1}^{k} w_i \delta(y, y_i)$$

Regression

$$\frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}$$

where: $w_i = \frac{1}{d(x,x_i)}$