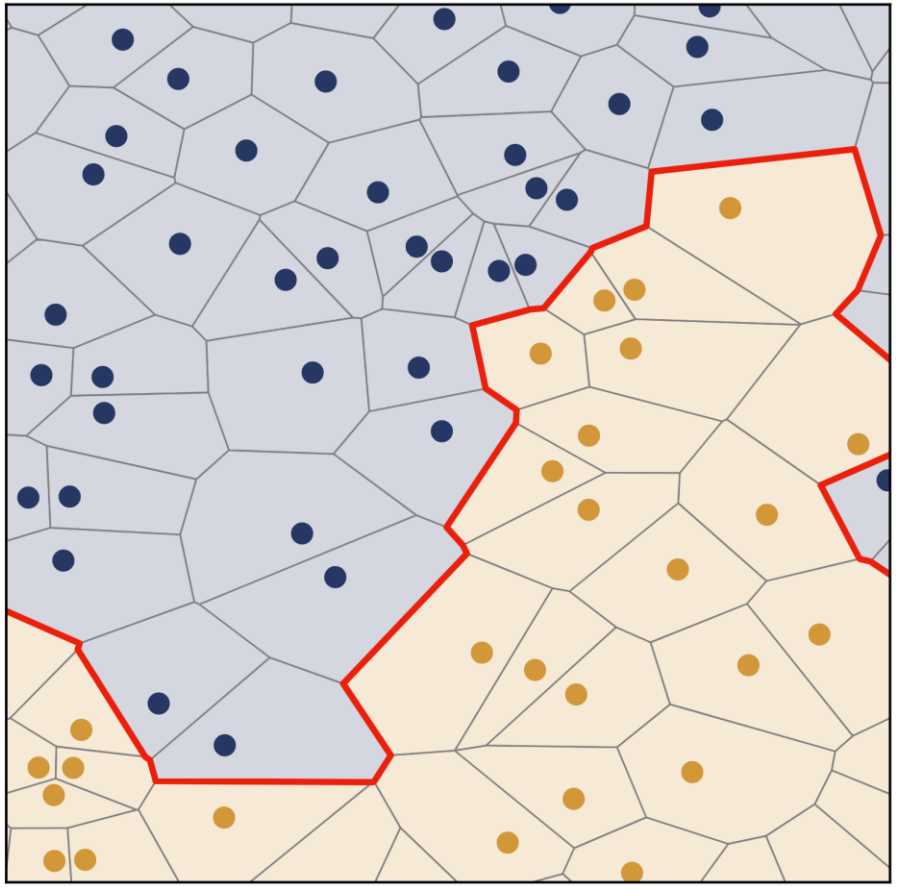
k-Nearest Neighbor

Nearest Neighbor

Idea
- Find most similar object and take its label

Requirements
- Similarity measure between objects

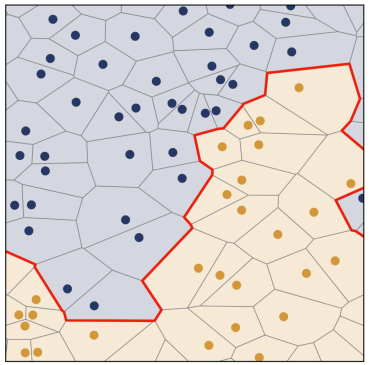
Alternatives
- Find k most similar objects and take their majority label
- Weight those similar objects by their similarity
- Transform coordinates such that classification is easier (Neighborhood component analysis)
- Transform metric (Large margin nearest neighbor)

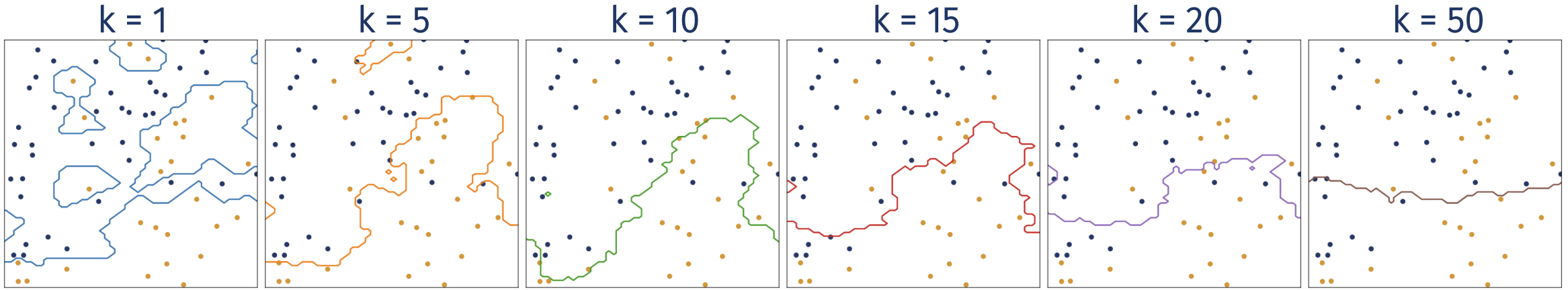


k-Nearest Neighbor

Idea

- Noisy data creates artefacts
- Find k most similar objects and take their majority label
- Small k: prone to overfit
- Large k: underfit





When to use

Pros
- No training except hyperparameter optimisation
- Simple method
- Need cheap model

Cons
- High dimensions: all points are far away from each other
- Large k needed due to noise may underfit
- Hard to query efficiently

How to fix
Project dimensions
Exploit intrinsic dimension
Change representation
Exploit static training data: index
Use approximate nearest neighbors

Summary k-Nearest Neighbor

- Pick majority vote under k closest data points
- Sensitive to features / transformation of features
- Risk of underfitting/overfitting: needs crossvalidation (coming later)
- Can be a cheap model even for millions of data points
- Inefficient in high dimensions