Introduction to Machine Learning Applications

Spring 2023

K-nearest-neighbors (k-NN)

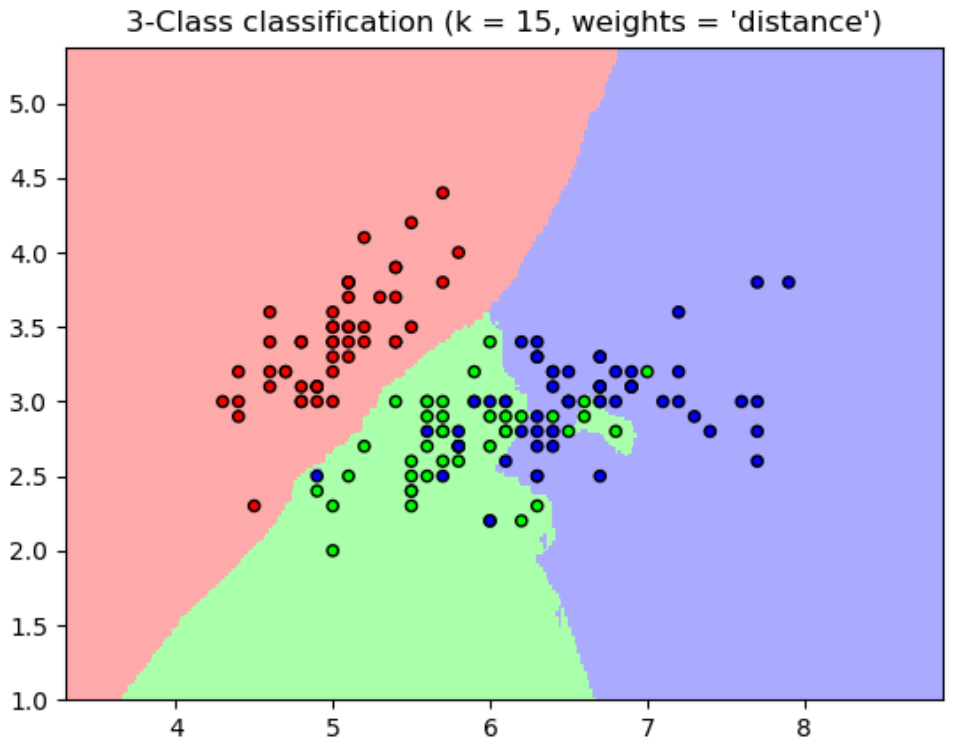
Minor Gordon

gordom6@rpi.edu

Rensselaer

Nearest Neighbor Classification

- Imagine data projected in a n-dimensional space, where n is the number of features
- Classification can be based on K neighbors or density



k-NN variations

- Best choice of *k* depends upon the data
  - Hyperparameter optimization
- Skewed class distribution causes issues for majority voting
  - Weight the classification, accounting for distance from the distance point to *k* nearest neighbors

k-NN distances

- Distance metrics: Euclidean, Jaccard coefficient (binary vectors), Hamming distance, …
- May need feature engineering and dimensionality reduction to make Euclidean distance more useful
- Naïve KNN computers distances from the test example to all stored examples
  - Nearest neighbor search algorithms

k-NN in Scikit-Learn

- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

- https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py

Challenge

Review the impact of changing some of the hyperparameters to the k-NN model.