Supervised learning and K nearest neighbors classification

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 25, Fall 2018

Wednesday April 10
Problems with KNN (1/2)

• 1) Computationally (and memory) intense!
  • We store ALL inputs in memory! Very bad for large datasets.
  • We need to compute the distance from all pairs of exemplars $x_i$ to each test point $y_j$.

• Fixes
  • Do K-means clustering beforehand on each training class. Make the K centroids our class exemplars.
  • Example: Each class has 50 points. K-means with $K = 10$ reduces this to 10 points per class.
  • Do PCA first to reduce dimensionality.
Problems with KNN (2/2)

• 2) Results are training sample dependent. Different training set → different classification results.

• 3) How do we pick $K$?
  • Can't analyze test data to inform $K$ — that's cheating!
  • Use a **Train** → **Validate** → **Test** workflow.
  • Validation is a "pseudo test" — withhold a subset of the input data from training memorization.
  • See how different $K$ valids affect classification on the withheld validation set.
  • Retrain (memorize) entire training set, test "for real" with the best $K$ value.