Supervised learning and K nearest neighbors classification

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 25, Fall 2018

Wednesday April 10

Plan

- Supervised learning and classification problems
- K nearest neighbors (KNN) classification

Classification problems

- **Clustering:** Assign each point into a group based on intrinsic structure of data.
- **Classification:** Give our algorithm the "correct answer": the group each data point really belongs to.
  - Our algorithm learns by generalizing the answers we told it to predict answer in new contexts.
- Procedure of giving answers then predicting is called **supervised learning**.
  - **Training phase:** Give algorithm data, along with correct answers (groups).
  - **Testing phase:** Predict the group of new data (withhold correct answer).
  - **Performance:** Percent correct on predicted groups vs actual.
- In machine learning, we usually call the algorithm a **model** or **classifier**.

K nearest neighbors (KNN) classifier

- *Not related to K-means*! K "means" something else with KNN!
- **Memory-based algorithm**
  - *Training phase*: Memorize (store) the entire dataset, along with the correct group assignment (**class**).
  - Memorized data are the **class exemplars**: What a ideal member of each class "looks like".
  - *Testing phase:* Present new data, classify based on distance to training data class exemplars.