Classification
Supervised learning

· 从个 training set 中已给出 sample 的 label，学习分类，给 unseen data 分类。

· underfitting vs. overfitting

  underfitting: training error 和 test error 都高。

  overfitting: training error 很低, test error 很高。

Instance-Based Classifier

得到 1 个新 x 时，看 training set 中有无 x、有则 return 其 label。

K Nearest Neighbor Classifier

Classify an unseen record:

  1. compute ↙ 与其它 record 间 distance。

  2. 找 k nearest neighbors

  3. Aggregate the labels of ↘ 来 predict 该 record 的类。

✓ Aggregation 的方法 很重要

  ① Majority rule ② Weighted majority based on distance

· Scaling issue:

  不同 feature 的 scale 不同会导致 feature 对于 distance 的影响不同。

scale 太大 将 dominate。