Nearest-Neighbors

Statistical Learning

Master in Big Data. University of Santiago de Compostela

Manuel Mucientes

Introduction

- Model-free method: memory-based
  - Fitting is not required
- Very effective for classification
- Works reasonably well for low-dimensional regression problems
  - For high-dimensional regression the bias-variance trade-off is not so good
- Bayes classifier: gold standard
  - Real data: we do not know the conditional distribution $\Pr(Y|X)$
- K-nearest neighbors (KNN)
  - Estimates the conditional distribution $\Pr(Y|X)$
  - Classifies an observation to the class with highest estimated probability

KNN

- Given K and x_0 (test observation):
  - Identify the K training points closest to x_0
  - Estimate the conditional probability for class j:

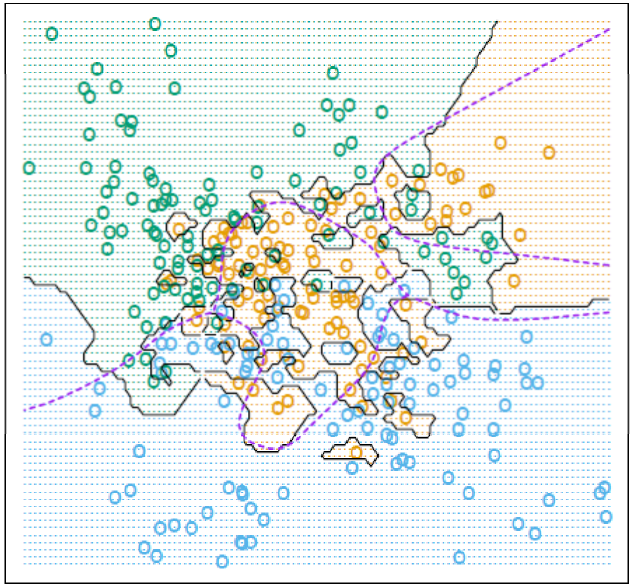
$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

  - Apply Bayes rule: classify x_0 to the class with largest probability
    - Ties are broken at random
- Closest points:
  - For real-valued features, typically the Euclidean distance in the feature space
  - First standardize each of the features: mean zero, variance 1

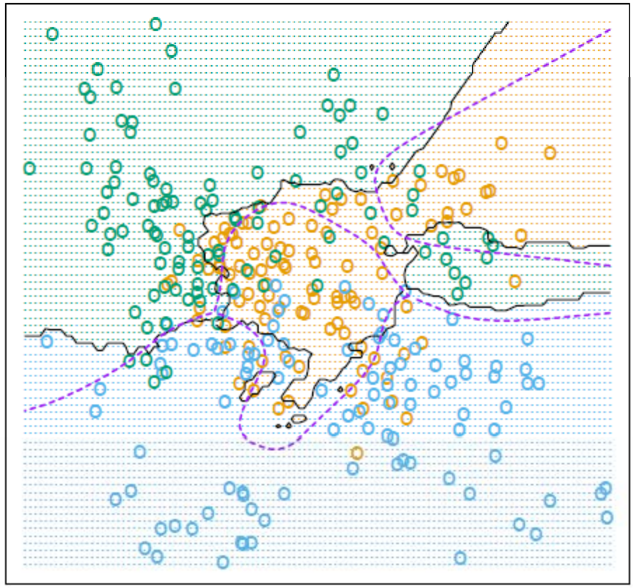
KNN (ii)

- KNN is successful with very irregular decision boundaries
- Asymptotically the error rate of 1-NN classifier is never more than twice the Bayes rate
  - Provides a rough idea of the best possible performance
  - “Asymptotic”: assumes the bias of the NN rule to be zero
    - In real problems the bias can be substantial
- Example: simulated, three classes

1-Nearest Neighbor

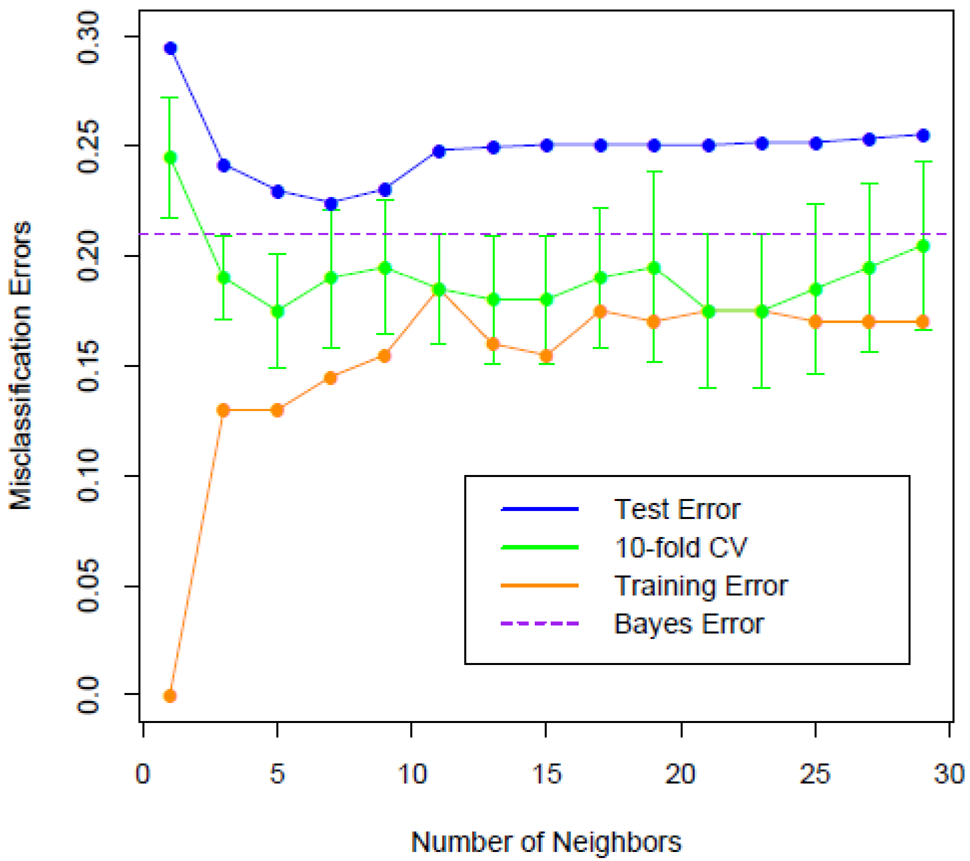


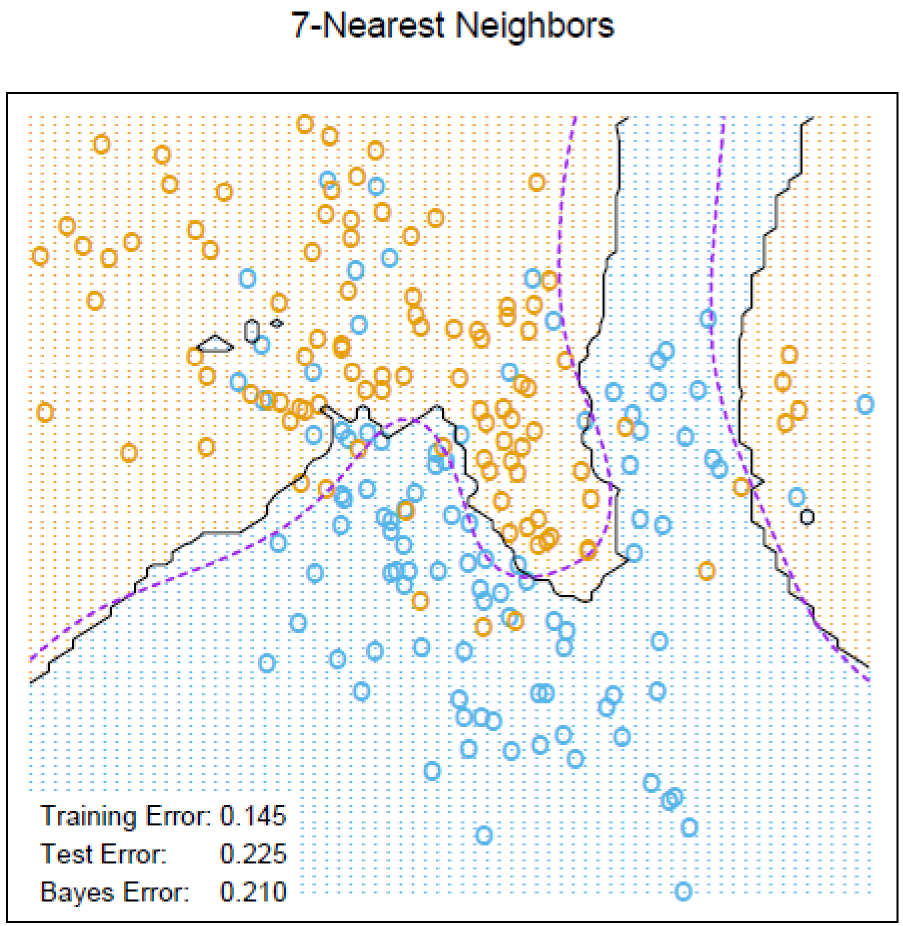
15-Nearest Neighbors



KNN (iii)

- Example: simulated, two classes



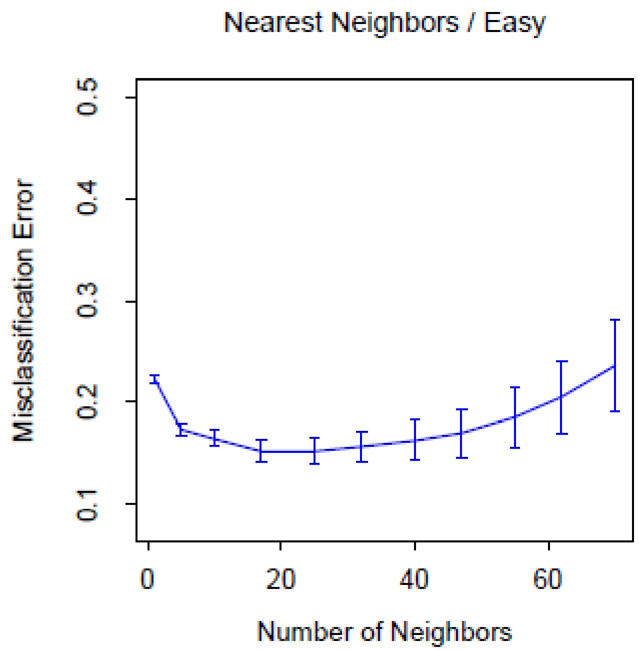


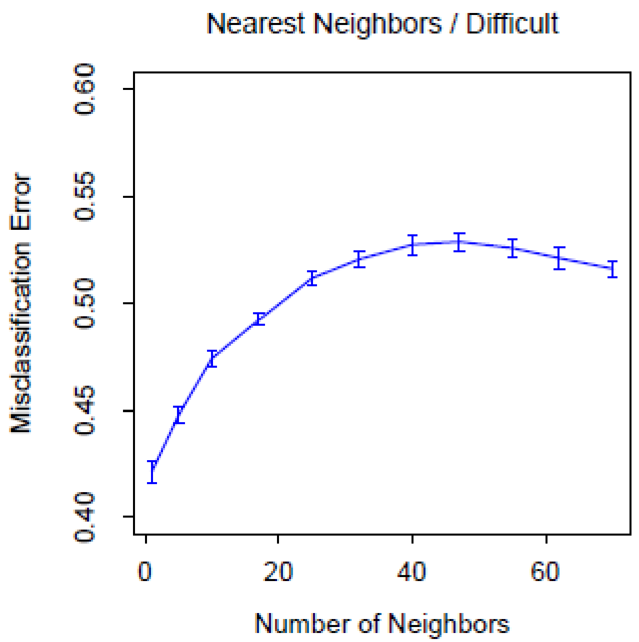
KNN (iv)

- Example: ten independent features, two classes

$$Y = I\left(X_1 > \frac{1}{2}\right); \quad \text{problem 1: “easy”},$$

$$Y = I\left(\text{sign}\left\{\prod_{j=1}^{3}\left(X_j - \frac{1}{2}\right)\right\} > 0\right); \quad \text{problem 2: “difficult.”}$$



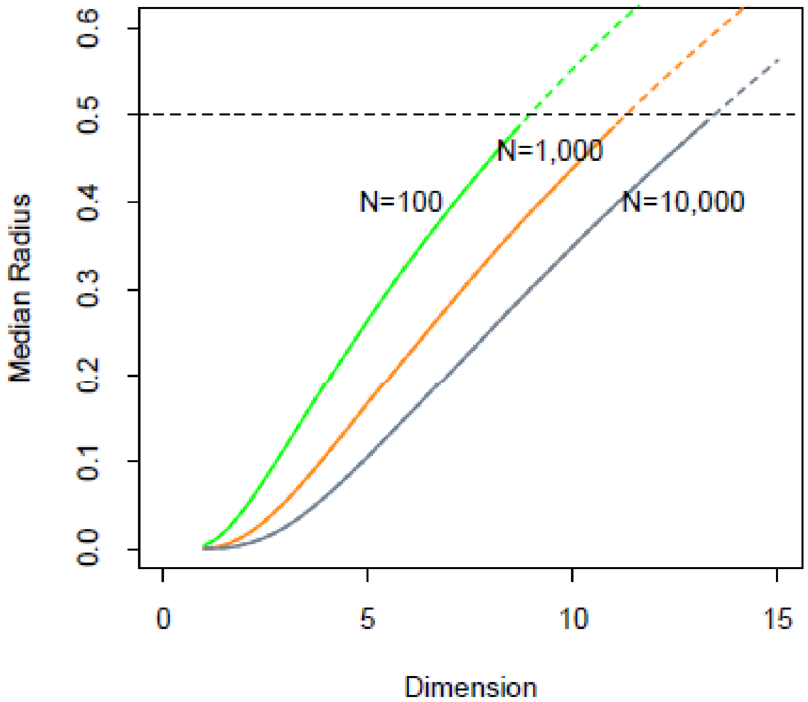


High-dimensional Feature Spaces

- The nearest neighbors can be very far away
  - Increases the bias of KNN
- Radius of 1-NN for N data points in the unit cube [-0,5, 0,5]^p

$$\text{median}(R) = v_p^{-1/p}\left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

  - The median quickly approaches 0.5 (the distance to the edge of the cube)



Computational Considerations

- Computational load:
  - Finding the neighbors
  - Storing the entire training set
- With N observations and p predictors, $N \times p$ operations for finding the neighbors
  - Fast algorithms for finding nearest-neighbors
- Reducing the storage requirements: instances selection
  - Keep the most important points: near the decision boundaries and on the correct side of those boundaries

Bibliography

- G. James, D. Witten, T. Hastie, y R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer, 2013.
  - Chapter 2, pp. 39-42

- T. Hastie, R. Tibshirani, y J. Friedman, The elements of statistical learning. Springer, 2009.
  - Chapter 13, Sec. 13.3-13.5