PS5841

Data Science in Finance & Insurance

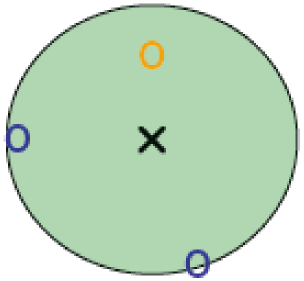
KNN

Yubo Wang

Autumn 2021

COLUMBIA UNIVERSITY

K-Nearest Neighbor Classifier



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

- Classifies $\boldsymbol{x}_0$ to the class with the highest probability

Probability

- In the KNN neighborhood of a particular test observation, $\hat{p}_{C_j}$ is the proportion of training observations that are from the j-th class C_j , where $j = 1, \dots, J$

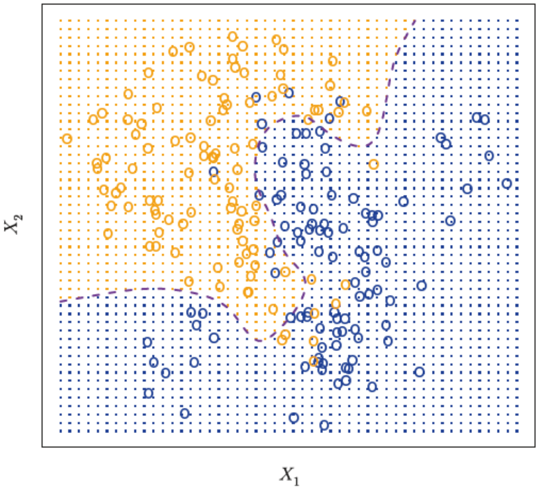
$$\hat{p}_{C_j} = \frac{n_{C_j}}{K}$$

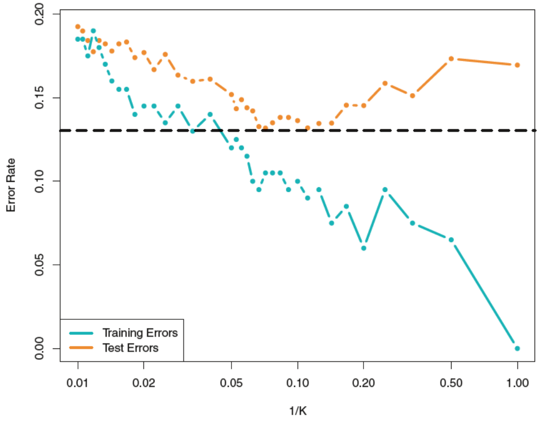
- The empirical probability of the predicted class

$$\max\left(\hat{p}_{C_j}\right)$$

K-Nearest Neighbor Classifier

- The bias-variance tradeoff tends to produce a U-shaped test error





- Can use cross validation to find the optimal K based on “accuracy”

K-Nearest Neighbor Classifier

- The choice of K is important
- The bias-variance tradeoff
  - A small K
    - More flexible decision boundary
    - Low-bias and high-variance classifier
  - A large K
    - Near-linear decision boundary
    - High-bias and low-variance classifier
- Neither gives good predictions

Gini Index

- The Gini index for the neighborhood is a measure of variance across the J classes

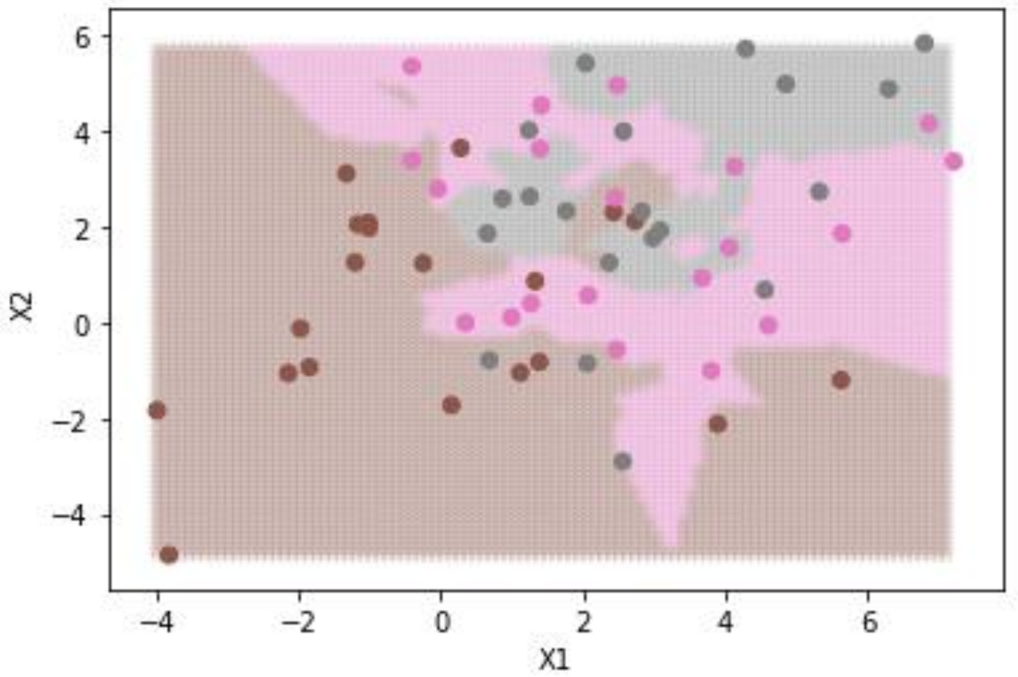
$$G = \sum_{j=1}^{J} \hat{p}_{C_j}\left(1 - \hat{p}_{C_j}\right)$$

- G will take on a small value if the neighborhood contains predominantly observations from a single class
- Example: for a 2-class response

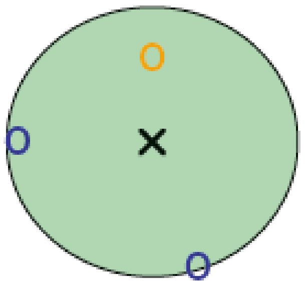
$$G = \hat{p}_{C_1}\left(1 - \hat{p}_{C_1}\right) + \hat{p}_{C_2}\left(1 - \hat{p}_{C_2}\right)$$

Decision Boundary

- 3-class response with 2D features



K-Nearest Neighbor Regression



$$\hat{y} = E(Y|X = \boldsymbol{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} y_i$$

- Predicted response is the mean response in the neighborhood

That was