Contents

1 K-nearest neighbours

1.1 Introduction

- In theory we would always like to predict the class Y using the Bayes classifier.
- But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.
- Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods.
- **K-nearest neighbors** (KNN) classifier is one of the many methods that try to estimate the conditional distributionof Y given X, and then classify a given observation to the class with highest *estimated* probability.

1.2 Parametric vs nonparametric methods

- KNN method is one of the simplest and best-known non-parametric methods.
- **Parametric methods** assume a functional form for $f(X)$ (e.g., linear regression, logistic regression).
  - often easy to fit (need estimate only a small number of parameters); can have a simple interpretation; inference can be easily performed.
  - But, they make strong assumptions about the form of $f(X)$: if this is far from the truth, and prediction accuracy is our goal, what does this involve for method performance?
- **Nonparametric methods** do not explicitly assume a parametric form for $f(X)$, and thereby are more flexible. But, we know that an excessive flexibility can involve . . . . . .

- In general, parametric methods will be preferred when
  - interpretability is the primary purpose
  - there is a small number of observations per predictor (the so-called *curse of dimensionality* problem).

1.3 KNN classifier

1. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$.
2. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

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

3. Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

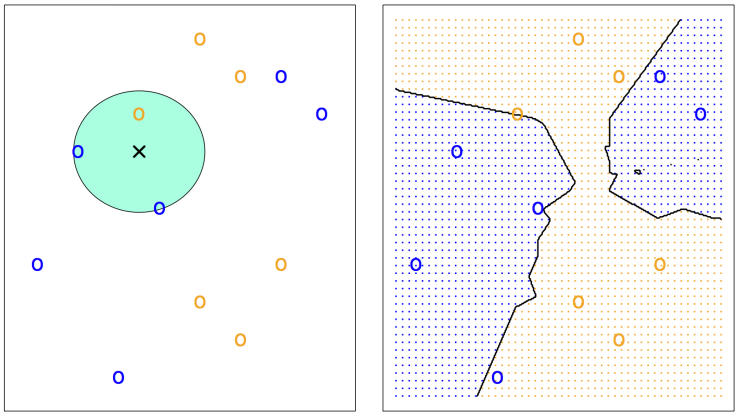

Figure 1: The KNN approach, using K=3.

- A small training data set consisting of six blue and six orange observations (see Figure 1).
- Our goal is to make a prediction for the point labeled by the black cross. Suppose that we choose $K = 3$.
- What class will KNN predict for the blackcross?
- At right, the KNN decision boundary.

- Despite the fact that it is a very simple approach, KNN can often produce classifiers that are close to the optimal Bayes classifier.

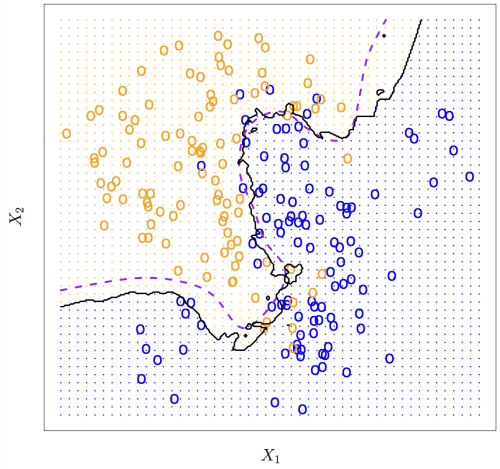

Figure 2: KNN ($K = 10$) and Bayes decision boundaries at comparison.

- A simulated data set consisting of 100 observations in each of two groups (see Figure 2).
- The test error rate using KNN is 0.1363, the Bayes error rate of 0.1304.

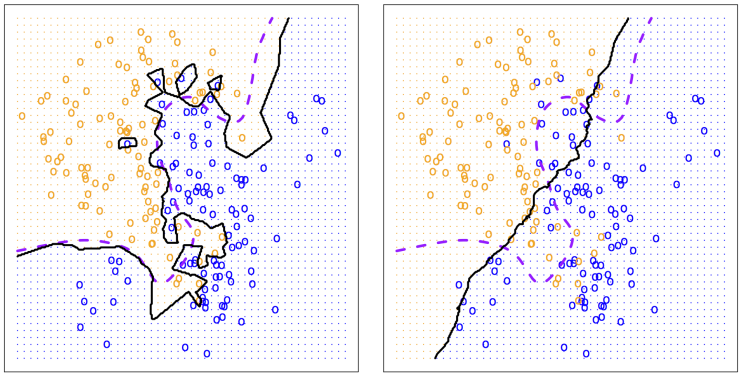

Figure 3: KNN with $K = 1$ compared to KNN with $K = 100$.

1.4 Which K?

- When $K = 1$, the decision boundary is overly flexible. This corresponds to a . . . -bias but . . . -variance classifier (see Figure 3).
- When $K = 100$, the decision boundary is not enough flexible. This corresponds to a . . . -variance but . . . -bias classifier.
- Test error rates of 0.1695 and 0.1925, respectively.
- What is the training error for $K = 1$?

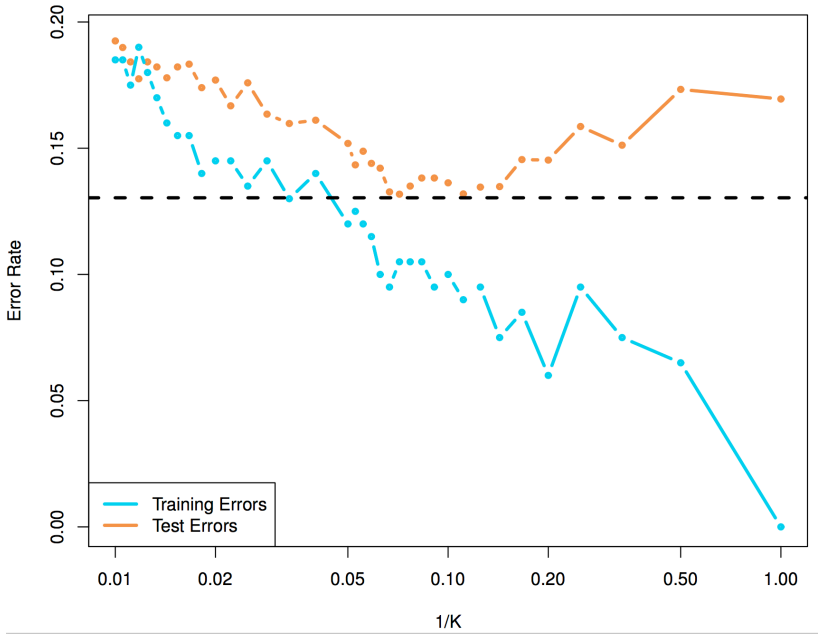

Figure 4: KNN test and training errors as a function of $1/K$.

- Training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) (see Figure 4).
- The black dashed line indicates the Bayes error rate.

- In both the regression and classification settings, choosing the optimal level of *flexibility* is critical to the success of any statistical learning method.
- It amounts to a *bias-variance tradeoff* (U-shape in the test error)
  - interpretability-accuracy tradeoff
  - under-fit versus over-fit tradeoff (when the fit is just right?)
  - parsimony versus black-box tradeoff (simple vs all of the variables).

1.5 Nonparametric methods

- A nonparametric method does not explicitly assume a parametric form for $f(X)$ (flexible).
- In general, the nonparametric approach will outperform the parametric one if the parametric form that has been selected deviates from the true form of f.
- Nevertheless, when dimension p increases parametric methods tend to outperform nonparametric approaches.
- The reason is the **curse of dimensionality**: in higher dimensions there is a reduction in effective sample size.
  - Spreading n observations over quite large p dimensions results in a phenomenon in which a given observation has no nearby neighbors.
  - That is, the K nearest neighbors tend to be far away from x in high dimensions, leading to a very poor prediction of $f(x)$.

1.5.1 Curse of dimensionality

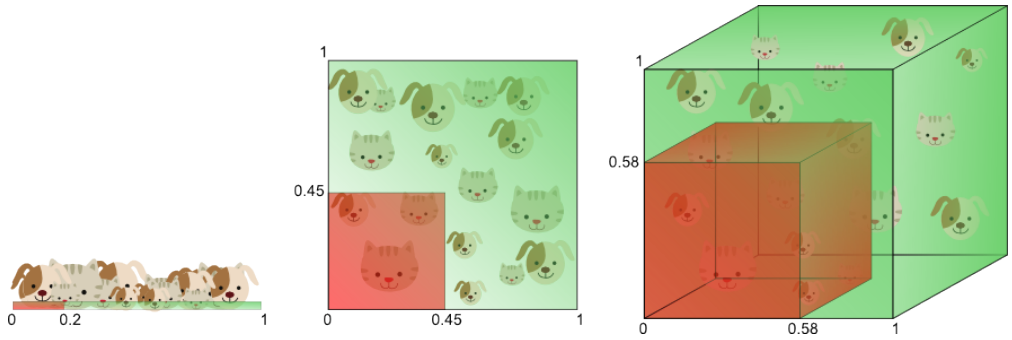

Figure 5: A 20% neighborhood for $p = 1, 2, 3$.

If we want our neighbourhood of training data to cover 20% of (see Figure 5):

- 1D feature space, the amount of training data needed is 20% of the feature range
- 2D feature space, the amount needed is ... 45% in each dimension
- 3D feature space, the amount needed is ... 58% in each dimension

---

- Nearest neighbor averaging can be pretty good for *small* p and *largish neighbourhood* $\mathcal{N}$.
- Nearest neighbor methods can be losing when p is large (curse of dimensionality).
  - We need to get a reasonable fraction of the n values of y_i to average to bring the variance down, e.g. 10%.
  - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging (see Figure 6)

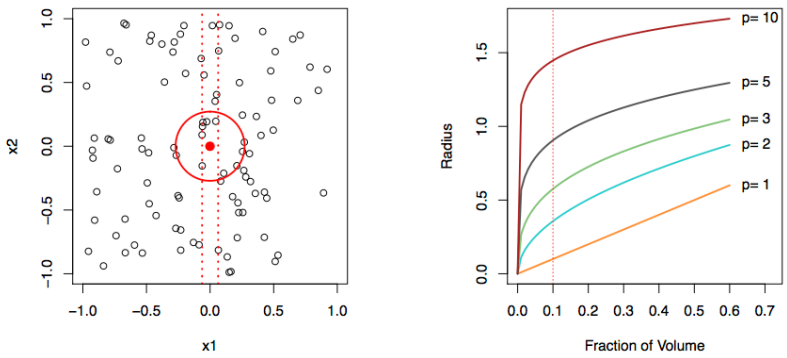

Figure 6: A 10% neighborhood for $p = 2$ (left); the increase in radius as p gets larger (right).

1.6 Comparison of Classification Methods

- Logistic regression and LDA methods are closely connected.
  - Both produce *linear* decision boundaries.
  - They differ in the estimation method
  - LDA outperforms logistic regression when the assumptions of normality and homoscedasticity approximately hold, and viceversa.
- QDA serves as a *compromise* between the KNN method and the LDA and logistic regression approaches.
  - It assumes a quadratic decision boundary, then it can accurately model a wider range of problems than can the linear methods.

- KNN is a completely non-parametric approach: *no assumptions are made about the shape of the decision boundary.*
  - We expect it dominates LDA and logistic regression when the decision boundary is highly non-linear. Though, the level of smoothness must be carefully chosen.
  - We need p not overly large.
  - Serves just a prediction purpose.