Elements of Statistical Learning

Chapter 2
Overview of Supervised Learning

Content: 2.1 - 2.5

Exercises: 2.1 - 2.3

2.3 APPROACHES TO PREDICTION: NEAREST NEIGHBOUR & LEAST SQUARES

Nearest Neighbour

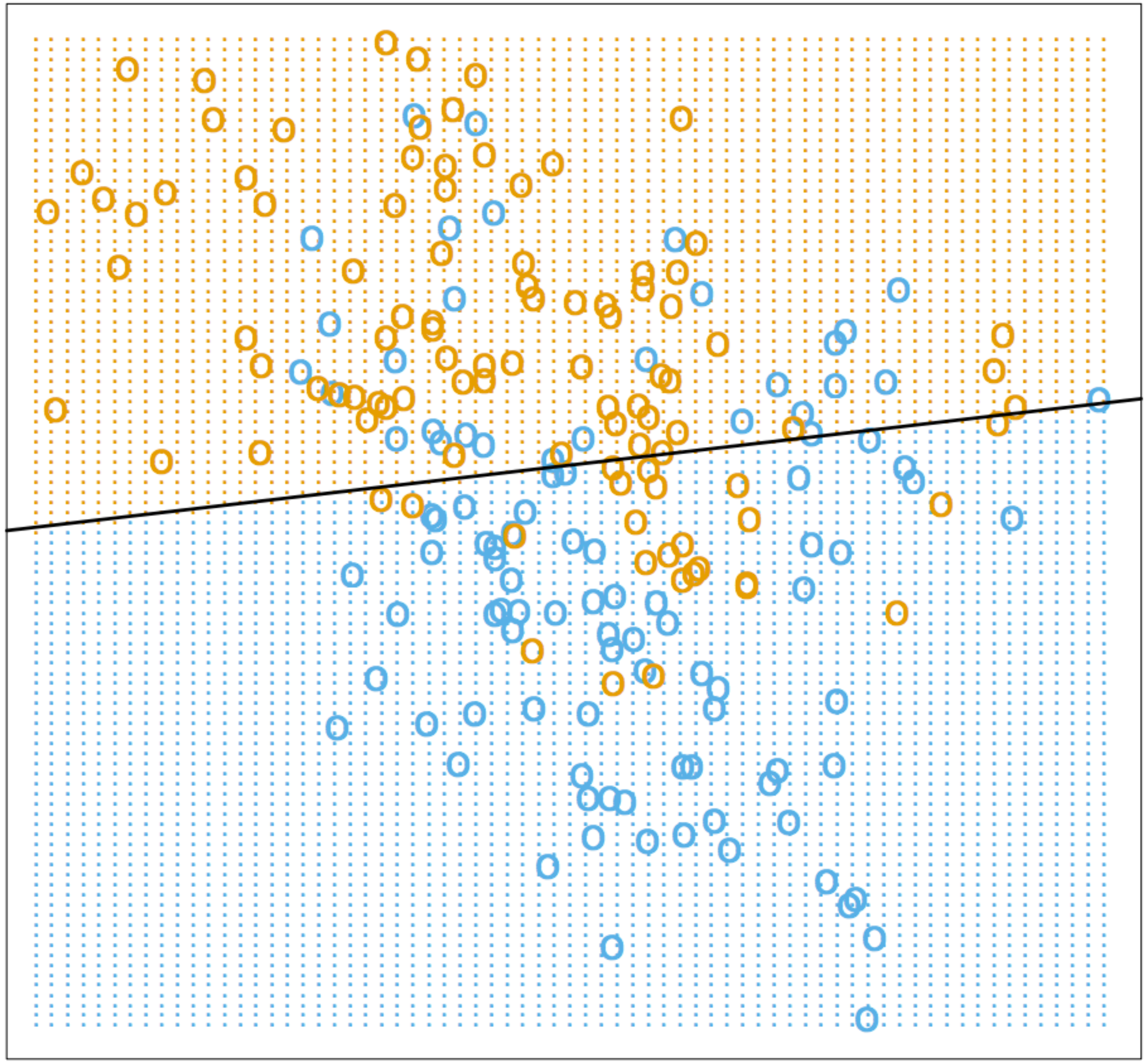
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Linear Model & Least Squares

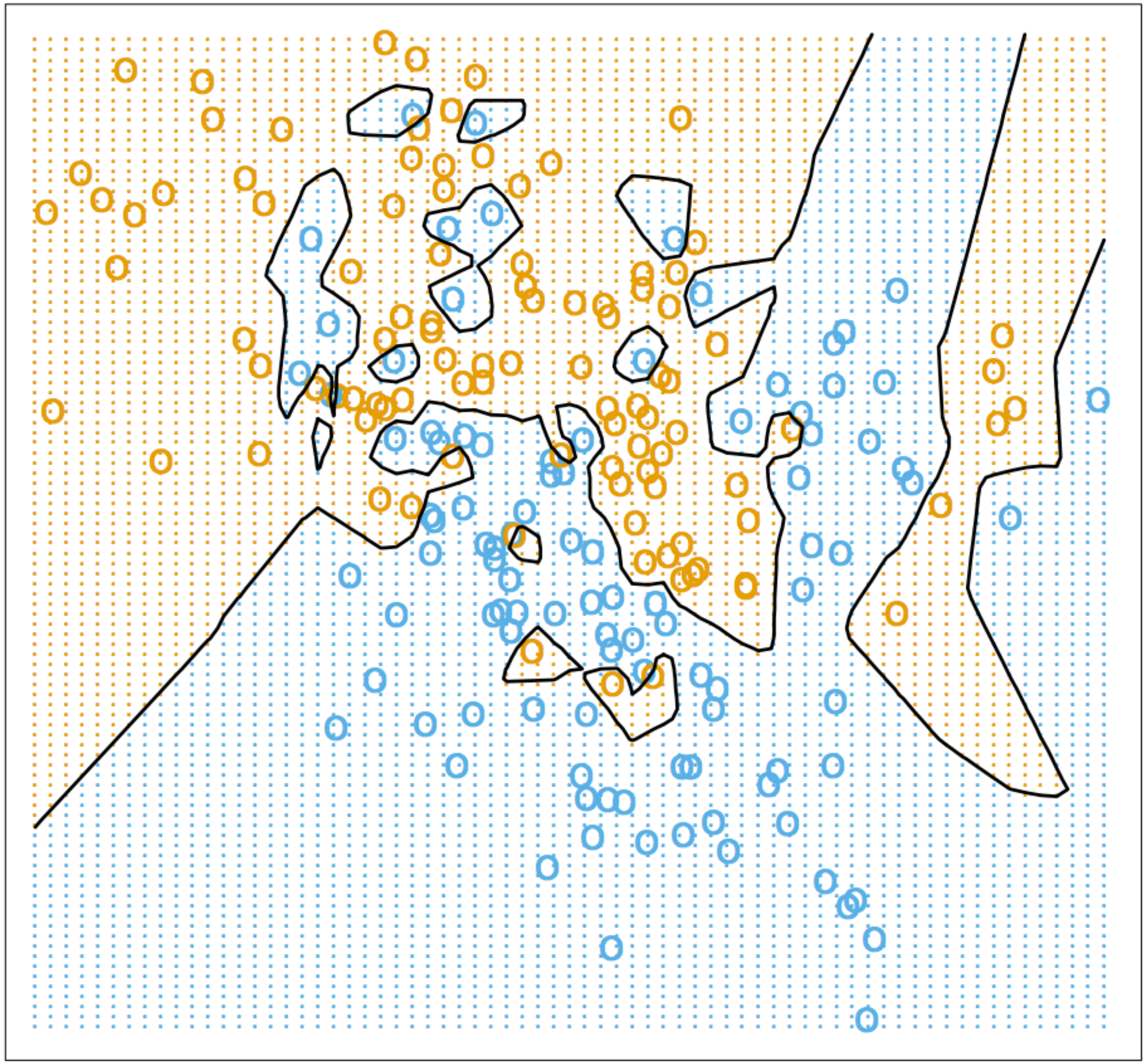
$$\hat{Y} = X^T \hat{\beta}$$

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

2.3 APPROACHES TO PREDICTION: NEAREST NEIGHBOUR & LEAST SQUARES



2.3 APPROACHES TO PREDICTION: NEAREST NEIGHBOUR & LEAST SQUARES



2.3 APPROACHES TO PREDICTION: NEAREST NEIGHBOUR & LEAST SQUARES

Scenario 1:

The training data in each class were generated from bivariate Gaussian distributions with uncorrelated components and different means

High bias, low variance

Scenario 2:

The training data in each class came from a mixture of 10 low variance Gaussian distributions, with individual means themselves distributed as Gaussian

Low bias, high variance

2.4 STATISTICAL DECISION THEORY

$$L(Y, f(X)) = (Y - f(X))^2$$

$$EPE(f) = E(Y - f(X))^2$$

• • •

2.4 STATISTICAL DECISION THEORY

$$EPE = E[L(G, \hat{G}(X))]$$

$$\hat{G}(x) = G_k \text{ if } \max_{g \in G} P(g \mid X = x)$$

Bayes-optimal decision boundary

2.5 LOCAL METHODS IN HIGH DIMENSIONS

So for a reasonably large set of training data why not just use Nearest Neighbour as an approximation of $E(Y|X=x)$?

2.5 LOCAL METHODS IN HIGH DIMENSIONS

So for a reasonably large set of training data why not just use Nearest Neighbour as an approximation of E(Y|X=x)?

The curse of dimensionality!

Consider nearest neighbour in a unit hypercube of uniformly distributed inputs -

Capturing a fraction r of the unit volume in p dimensions requires edge length:

$$e_p(r) = r^{1/p}$$

All sample points are close to the edge of the sample -

Median distance from the origin to the nearest data point is given by:

$$d(p, N) = (1 - \frac{1}{2}^{1/N})^{1/p}$$

The sampling density is proportional to $N^{1/p}$

E.g. if 100 samples is dense for a given problem in 1-d then 100^10 samples is required to match that density in 10-d

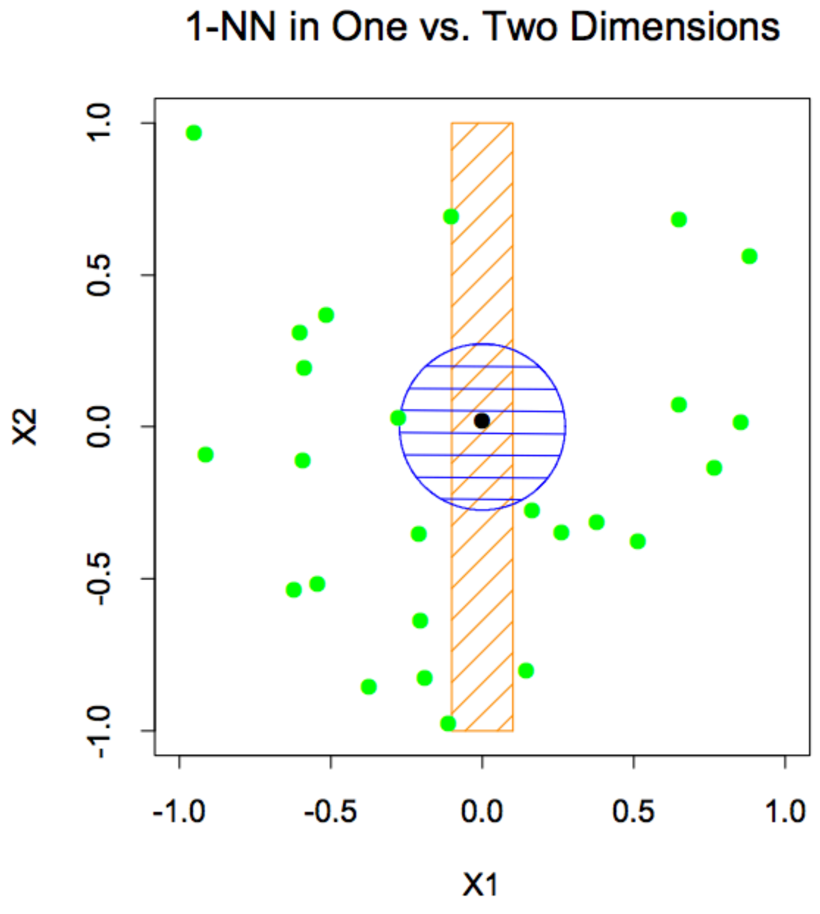
2.5 LOCAL METHODS IN HIGH DIMENSIONS

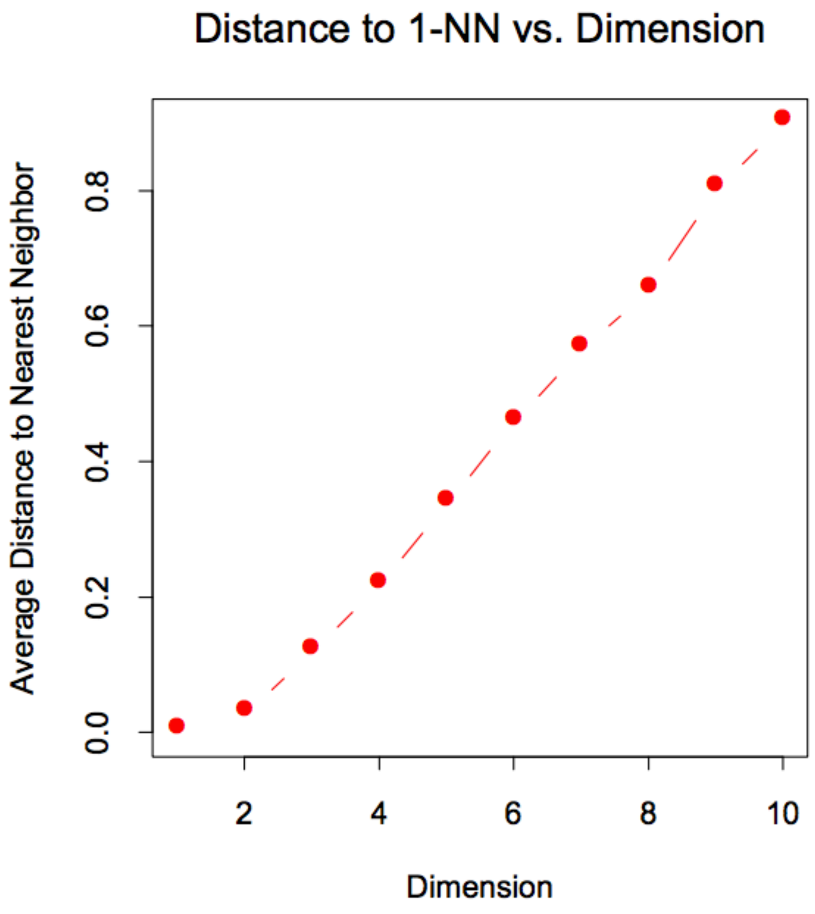
Example:

$$Y = f(X) = e^{-8||X||^2}$$

2.5 LOCAL METHODS IN HIGH DIMENSIONS

Example:





2.5 LOCAL METHODS IN HIGH DIMENSIONS

Example:

$$MSE(x_0) = E_T[f(x_0) - \hat{y}_0]^2$$

$$MSE(x_0) = Var_T(\hat{y}_0) + Bias^2(\hat{y}_0)$$

EXERCISES