Classification and Regression trees (CART)

Recall: Kernel smoothing

Given: TS $(x_1, y_1) \dots (x_n, y_n) \sim iid\ (X, Y)$

$X \in \mathbb{R}$

Goal: Estimate $E(Y|x) = f(x)$

Kernel smoother:

$K(x)$ positive symmetric Kernel

(Think Gaussian)

$K_\lambda(x) = K\left(\frac{x}{\lambda}\right)$ large $\lambda \to$ wide Kernel

Kernel estimate

$$\hat{f}_\lambda(x) = \sum_i K_\lambda(x - x_i)\, y_i \Big/ \sum_i K_\lambda(x - x_i) \leftarrow$$

Distance-weighted response average

Estimate λ by cross-validation

Various extensions: local polynomials; robust smoothers; adaptive λ

⋮

Generalizes to multi-dim. (in principle)

However

The curse of dimensionality gets in the way

Curse of dimensionality

- High-dim data are always sparse

  Given $(\underline{x}_1, y_1) \dots (\underline{x}_n, y_n)$ $\quad \underline{x}_i \sim U[0,1]^{10}$.

  We want to estimate $E(Y|\underline{x})$ using a box Kernel with side length 0.1

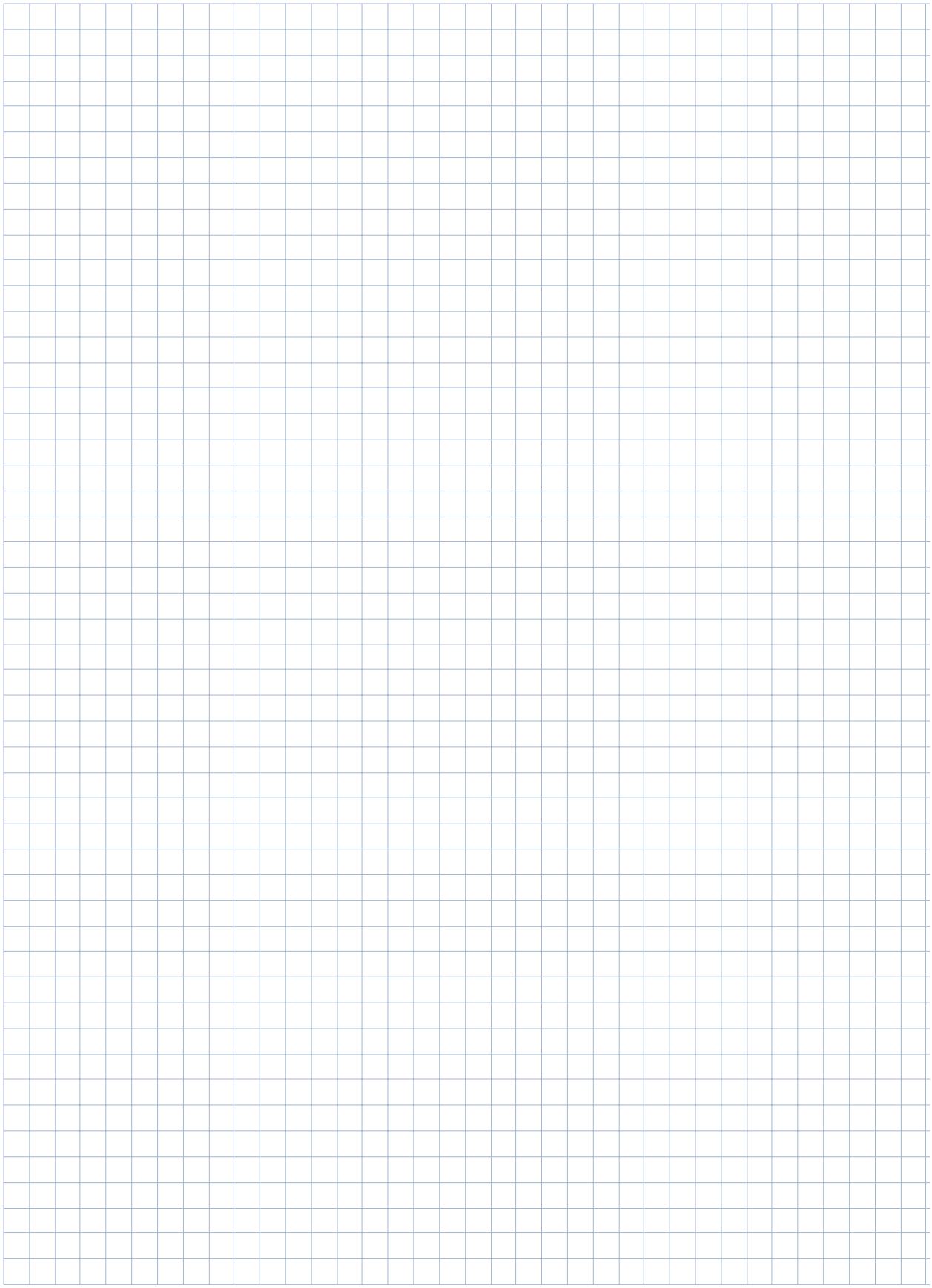
  Volume of Kernel is $0.1^{10} = 10^{-11}$

  The expected # of training obs with $\underline{x}_i$ in box is $n * 0.1^{10}$. Say $n = 1000$

  Chances are box is empty.

- Suppose you choose sidelength of the box to make box contain 10% of the data

  $\Rightarrow$ width of box has to be 0.8



Curse of dimensionality:

- High-d data are always sparse

Ex. Given $(x_1, y_1) \dots (x_n, y_n)$ $x_i \sim U[0,1]^{10}$

- Want to estimate $E(Y|x)$ using box kernel with side length 0.1 $\Rightarrow$ Vol of kernel $= (0.1)^{10}$

Expected # of training obs in box $= n \cdot 0.1^{10}$

Box will almost certainly be empty unless n is enormous

- Want to choose side length of box so that volume $= 0.1 \Rightarrow$ width of box $= 0.8$

- High d space has strange properties

Consider regular grid

In 2-d every grid cell has $3^2 - 1$ neighboring cells

In 10-d every grid cell has $3^{10} - 1$ neighbors - 59,000 neighbors.

- Need to decide how to scale the predictors (how many apples make an orange)?

$d^2(x_1, x_2) = \sum_j w_j (x_{1j} - x_{2j})^2$ how to choose $w_1 \dots w_p$

Good choise depends on nature of

response variation.

Consider $p = 2$, two scenarios

$E(Y|\underline{x}) = f(x_1)$ Best choice $w_2 = 0$

$E(Y|\underline{x}) = f(x_2)$ $w_1 = 0$

CART basic idea: Local averaging but choose neighborhood over which to average based on nature of response variation.

Recursive partitioning CART

Given: TS $(\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)$

$\underline{x} \in \mathbb{R}^p \quad y \in \mathbb{R}^1$

Goal: Estimate $E(y \mid \underline{x})$

"understand" dependence of y on $\underline{x}$.

Idea:

- Estimate $E(y \mid \underline{x})$ by local averaging
- choose neighborhood adaptively, based on nature of response variation

Form of model

$$\hat{f}(\underline{x}) = \sum_{i=1}^{K} c_i \, I(\underline{x} \in N_i)$$

piecewise constant over hyper-rectangles $N_1 \ldots N_K$ with $\cup N_i = \mathbb{R}^p$

Obvious: For given $N_1 \ldots N_K$

$$c_i = \text{mean}(y_j \mid \underline{x}_j \in N_i)$$

This choice minimizes resubstitution error

Key question How to determine K and $N_1 \dots N_K$

Special case: $K = 2$

$N_0 = \mathbb{R}^p$

Split into two boxes N_1, N_2 is characterized by split coordinate i and split point s.

$$N_1(i,s) = \{x \mid x_i < s\}$$
$$N_2(i,s) = \{x \mid x_i \geq s\}$$

$$S_1(i,s) = \{j \mid x_j \in N_1(i,s)\}$$
$$S_2(i,s) = \{j \mid x_j \in N_2(i,s)\}$$

$$c_1(i,s) = \text{mean}(y_j \mid j \in S_1(i,s))$$
$$c_2(i,s) = \text{---}$$

$$rss(i,s) = (n_1 - 1)\, \text{var}(y_j \mid j \in S_1(i,s)) + (n_2 - 1)\, \text{var}(y_j \mid j \in S_2(i,s))$$

Note: rss only changes when S_1 and S_2 change

For each i we have to try at most $n-1$ split points.

Means and variances can be updated.

Can evaluate rss in $O(np)$ operations

Solves the problem for $K=2$

How about $K=3$

$(n-1)p$ possibilities for 1st split

$(n-2)p$ possibilities for 2nd split, given 1st split

total $= (n-1)(n-2)p^2$

General K

$(n-1)(n-2)\cdots(n-K+1)p^{K-1}$ possible split

Defies the possibilities of today's computers.

Idea: Apply splitting procedure recursively