**Nonparametric Regression**

- Supervised
  - Regression
    - Parametric
    - Nonparametric
  - Classification
- Unsupervised

- ML

- K-nearest neighbors
- Decision trees
Assume

\[ y = B_0 + B_1 x_1 + B_2 x_2 + \varepsilon \]

Assume

\[ Y = B_0 + B_1 x_1 + \varepsilon \]
\[ \mathbb{E}[Y \mid X = x] \]

\[ \hat{\mathbb{E}}[Y \mid X = x] = \text{Ave} \left( \{ y_i : \text{where } x_i = x \} \right) \]

\[ \hat{\mathbb{E}}[Y \mid X = x] = \text{Ave} \left( \{ y_i : \text{where } x_i \text{ is close to } x \} \right) \]
**k-Nearest Neighbors**

To estimate $\mu(x) = E[y \mid X = x]$, use:

$$\hat{\mu}_k(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x,d)} y_i$$

$k$ observations with $x_i$ nearest to $x$. 

---

$K$-Nearest Neighbors
**Tuning Parameters**

\[ Y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + \varepsilon \]

MODEL PARAMETERS \( \downarrow \)

LEARNED FROM DATA

**K in KNN**

TUNING PARAMETER \( \downarrow \)

DEFINES HOW TO LEARN FROM DATA

**K in KNN**  

LEARNED FROM DATA
Other KNN Notes

- "Fast" to train, "slow" to predict
- Which features should be used?
- Categorical features? Dummy encoding
- How to calculate distance? You pick!
- Feature scaling? !!!
IDEA: FIND NEIGHBORHOODS, PREDICT AVERAGE OF Y; IN NEIGHBORHOODS
**Decision Trees**

\[ \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

**Find "split" that minimizes**

\[ \sum_{i \in N_L} (y_i - \hat{\mu}_N)^2 + \sum_{i \in N_R} (y_i - \hat{\mu}_N)^2 \]
Recursive Partitioning

\[ \sum_{i \in N_c} (y_i - \hat{\mu}_c)^2 + \sum_{i \in N_R} (y_i - \hat{\mu}_R)^2 \]

\[ \sum_{i \in N_{R_1}} (y_i - \hat{\mu}_{R_1})^2 + \sum_{i \in N_{R_2}} (y_i - \hat{\mu}_{R_2})^2 \]
Recursive Partitioning

\[ \text{SSE} = \sum_{j=1}^{J} \sum_{i \in N_j} (y_i - \hat{y}_j)^2 \]

\[ R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \]

# Neighborhoods
How to stop?

rpart:: rpart in R

\[ \text{minsplit} \]

Only consider split in neighborhood if it has at least this many observations.

\[ \text{minsplit} = 2 \Rightarrow \text{can always split} \]

\[ \text{cp} \]

Only accept a split if it increases \( R^2 \) by this amount or more.

\[ \text{cp} = 0 \Rightarrow \text{any split will be accepted} \]

"Complexity parameter"