LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Database Systems Group • Prof. Dr. Thomas Seidl

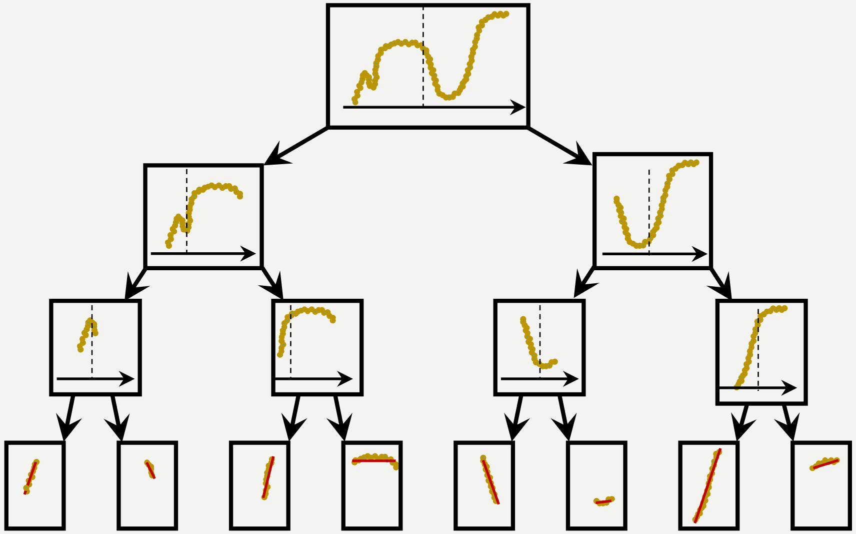
# Exercise 12: Numerical Prediction

Knowledge Discovery in Databases I

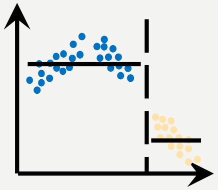
SS 2016

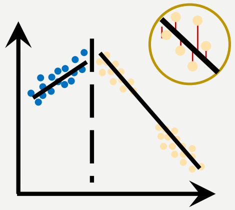
# Recap: Regression Tree Learning

- Recall:
  - Regression tree learning:
    - Given: Set of observations $T$
    - Find a split of $T$ in $T_1$ and $T_2$ with minimal summed impurity
    - If the stopping criterion is not reached: Repeat for $T_1$ and $T_2$
    - If the stopping criterion is reached: Undo the split



# Recap: Regression Tree Learning





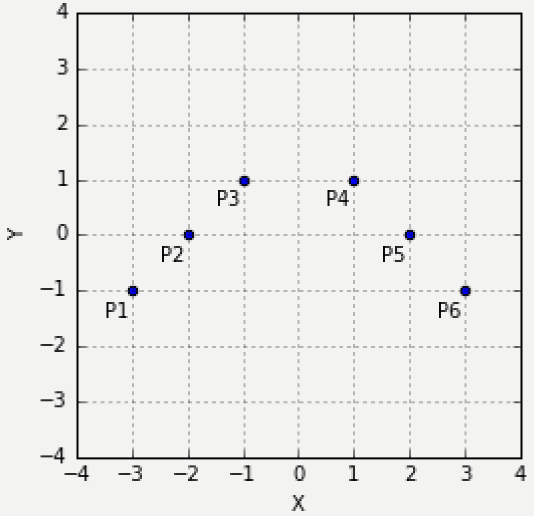
- Impurity measures:
  - Variance of the output: $imp(T) = \frac{1}{|T|}\sum_{(x,y)\in T}(y-\bar{y})^2$
    - Works well if constant models are learned …
  - Variance of the residuals: $imp(T) = \frac{1}{|T|}\sum_{(x,y)\in T}(y-f(x))^2$
    - Works better, as we will see …
  - Note: If the regression function $f$ is constant, both impurity measures yield the same value (since the optimal constant regression function will always be $f(x) = \bar{y}$)

- Stopping criterion:
  - If the relative impurity ratio induced by a split is higher than a given threshold, then the split is not significant (avoid overfitting):

$$\tau = \frac{imp(T_1) + imp(T_2)}{imp(T)} > \tau_0$$

# Exercise 12-1: Regression Trees

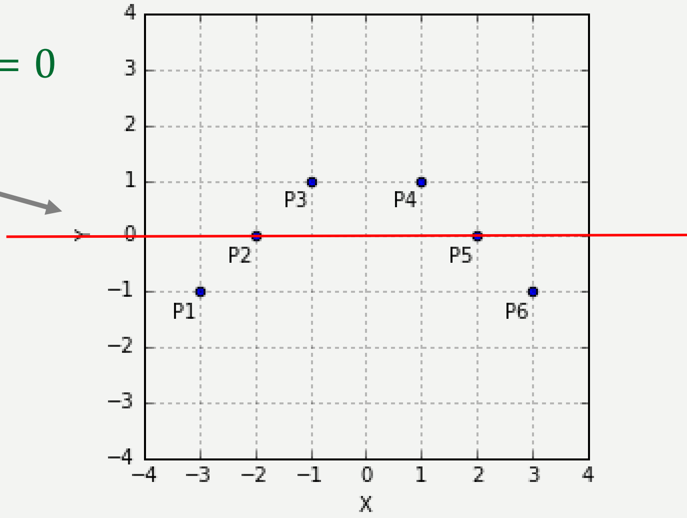
Consider the following dataset:



Search for the first best split. If the decision is obvious, you don't have to compute all possible splits. Then decide whether the split is significant or not by using the impurity ratio with $\tau_o = 0.5$.

# Exercise 12-1: Regression Trees

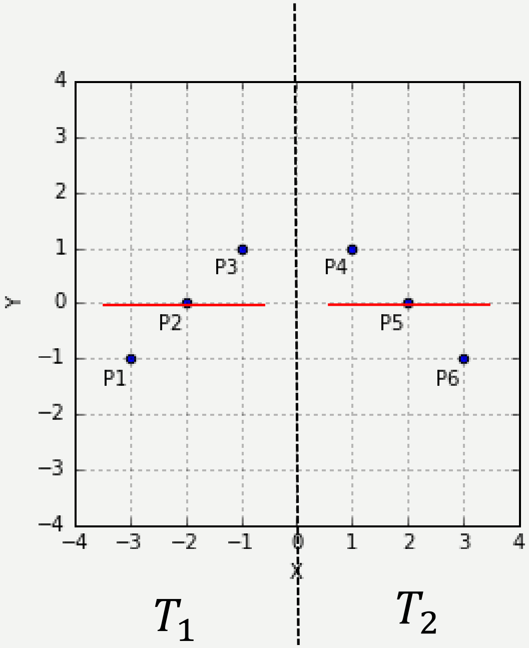
- Initially: $T = \{P_1, \dots, P_6\}$
- Calculate the impurity of $T$:
  - For the variance of the residuals, we need the optimal regression line minimizing the SSE
  - Closed form solution: $\beta_1 = \frac{Cov(x,y)}{Var(x)}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$
  - Calculate:
    - $\bar{x} = 0, \quad \bar{y} = 0$
    - $Cov(x,y) = \sum_{(x,y) \in T} (x - \bar{x})(y - \bar{y}) = 0$
    - Thus: $\beta_1 = 0, \quad \beta_0 = 0$
  - Since the optimal regression line is constant, both impurity measures will yield the same value
  - $imp(T) = \frac{1}{6}(1 + 0 + 1 + 1 + 0 + 1) = \frac{2}{3}$



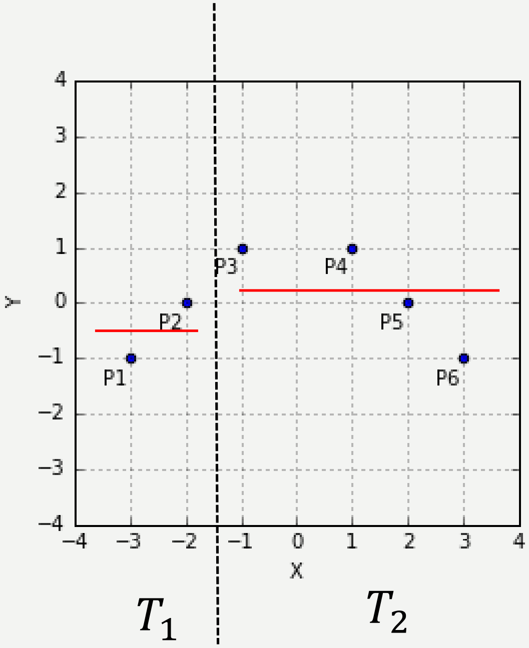
# Exercise 12-1 (a): Regression Trees

## a) Use as impurity measure the variance of the output.

- Try all possible splits, find the one with smallest summed impurity
- Split in half:
  - $\bar{y}_{T_1} = \bar{y}_{T_2} = 0$
  - $imp(T_1) = imp(T_2) = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$
  - $imp(T_1) + imp(T_2) = \frac{4}{3}$

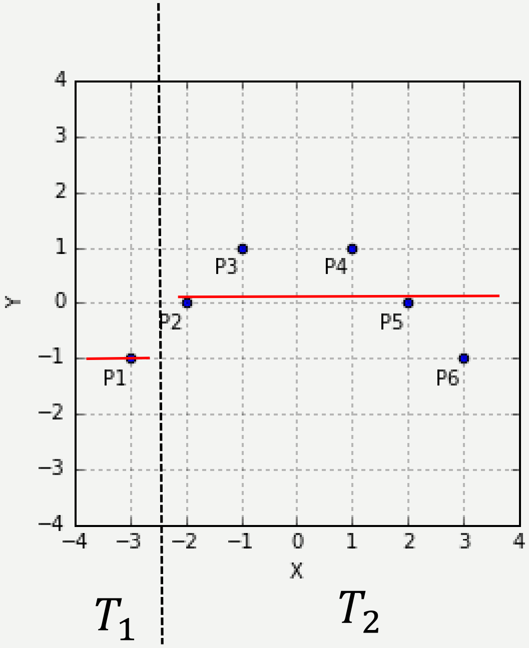


# Exercise 12-1 (a): Regression Trees



- Split 2 points:
  - $\bar{y}_{T_1} = \frac{1}{2}(0-1) = -\frac{1}{2}$
  - $imp(T_1) = \frac{1}{2}\left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = \frac{1}{4}$
  - $\bar{y}_{T_2} = \frac{1}{4}(1+1+0-1) = \frac{1}{4}$
  - $imp(T_2) = \frac{1}{4}\left(\left(1-\frac{1}{4}\right)^2 + \left(1-\frac{1}{4}\right)^2 + \left(0-\frac{1}{4}\right)^2 + \left(-1-\frac{1}{4}\right)^2\right) = \frac{11}{16}$
  - $imp(T_1) + imp(T_2) = \frac{15}{16} < \frac{4}{3}$ → better than previous split

# Exercise 12-1 (a): Regression Trees

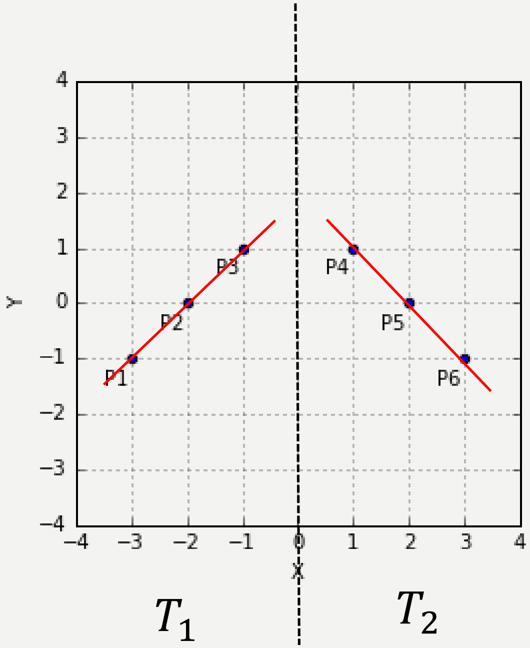


- Split 1 point:
  - $imp(T_1) = 0$ (single point)
  - $\bar{y}_{T_2} = \frac{1}{5}(0 + 1 + 1 + 0 - 1) = \frac{1}{5}$
  - $imp(T_2) = \frac{1}{5}\left(\left(0 - \frac{1}{5}\right)^2 + \left(1 - \frac{1}{5}\right)^2 + \left(1 - \frac{1}{5}\right)^2 + \left(0 - \frac{1}{5}\right)^2 + \left(-1 - \frac{1}{5}\right)^2\right) = \frac{14}{25}$
  - $imp(T_1) + imp(T_2) = \frac{14}{25} < \frac{15}{16}$ → best possible split
  - $\frac{imp(T_1)+imp(T_2)}{imp(T)} = \frac{14}{25} \cdot \frac{3}{2} = \frac{21}{25} > \tau_0$ → the split is not significant

## Exercise 12-1 (a): Regression Trees

### b) Use as impurity measure the variance of the residuals.

- Split in half:
  - The points in $T_1$ and $T_2$ are collinear
  - Thus, we can fit regression lines with no error
    - $imp(T_1) = imp(T_2) = 0$
    - $imp(T_1) + imp(T_2) = 0$
  - Since the impurity is always non-negative, there cannot exist a better split
  - $\frac{imp(T_1)+imp(T_2)}{imp(T)} = 0 < \tau_o$
  - Thus, the split is significant


4
3
2
1
0
−1
−2
−3
−4
Y
−4
−3
−2
−1
0
1
2
3
4
X
P1
P2
P3
P4
P5
P6
$T_1$
$T_2$