LECTURE 3 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

A P MOORE

END-TO-END ML

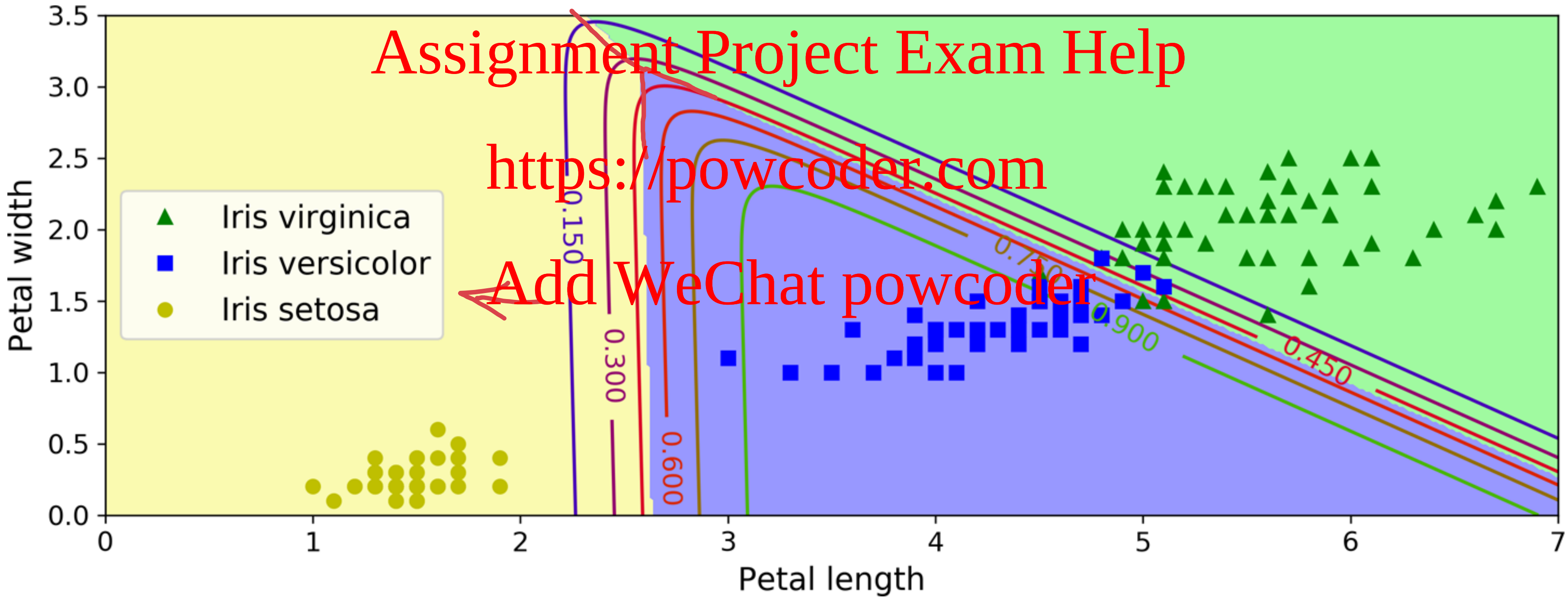
- Discover
- Explore
- Visualize

- Clean
- Sample
- Impute
- Encode
- Transform
  - Scale
- Modeling
  - Overfitting
  - Optimization
  - Model Selection
  - Regularization
  - Generalization

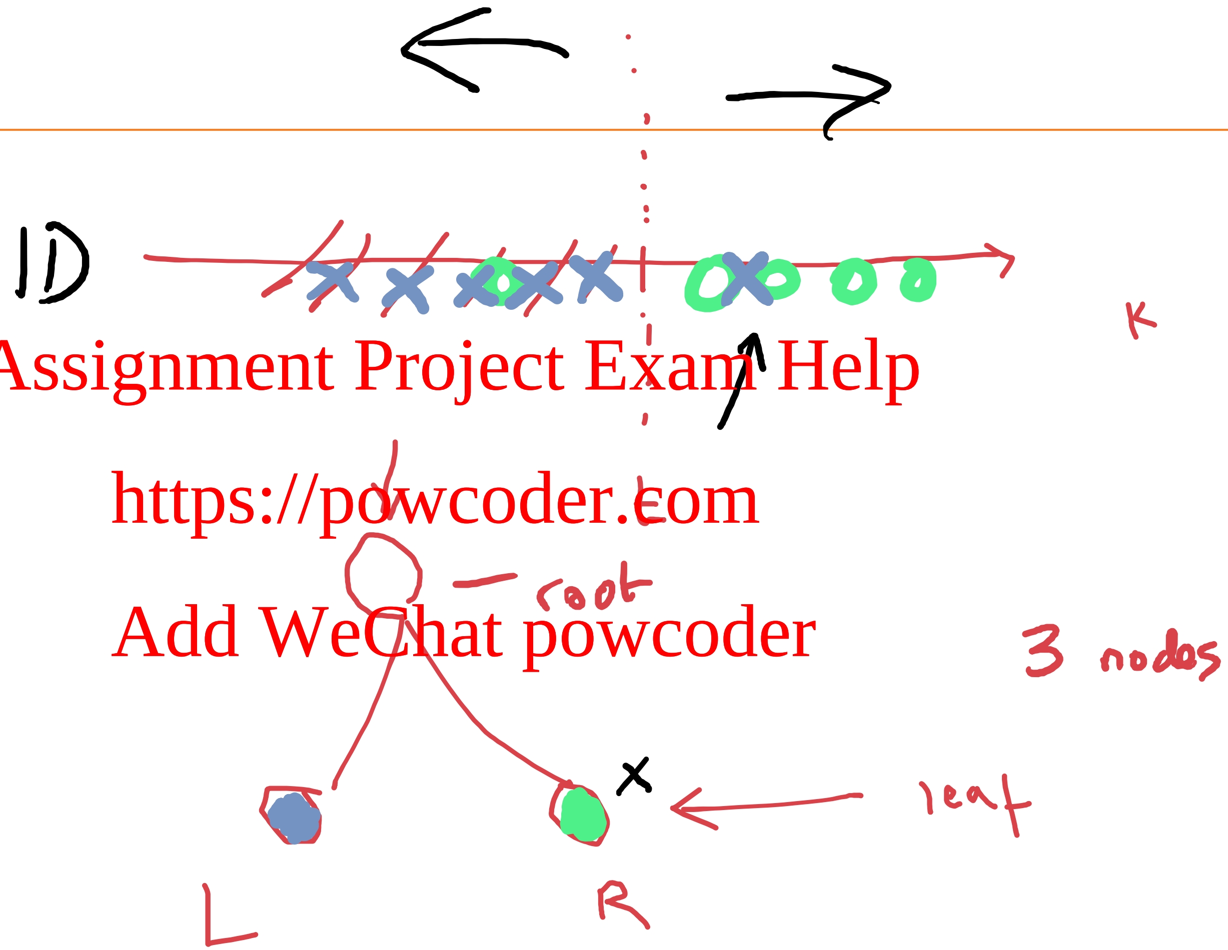
- Documentation
- Presentation
- Launch
- Monitor
- Maintain

DECISION BOUNDARY

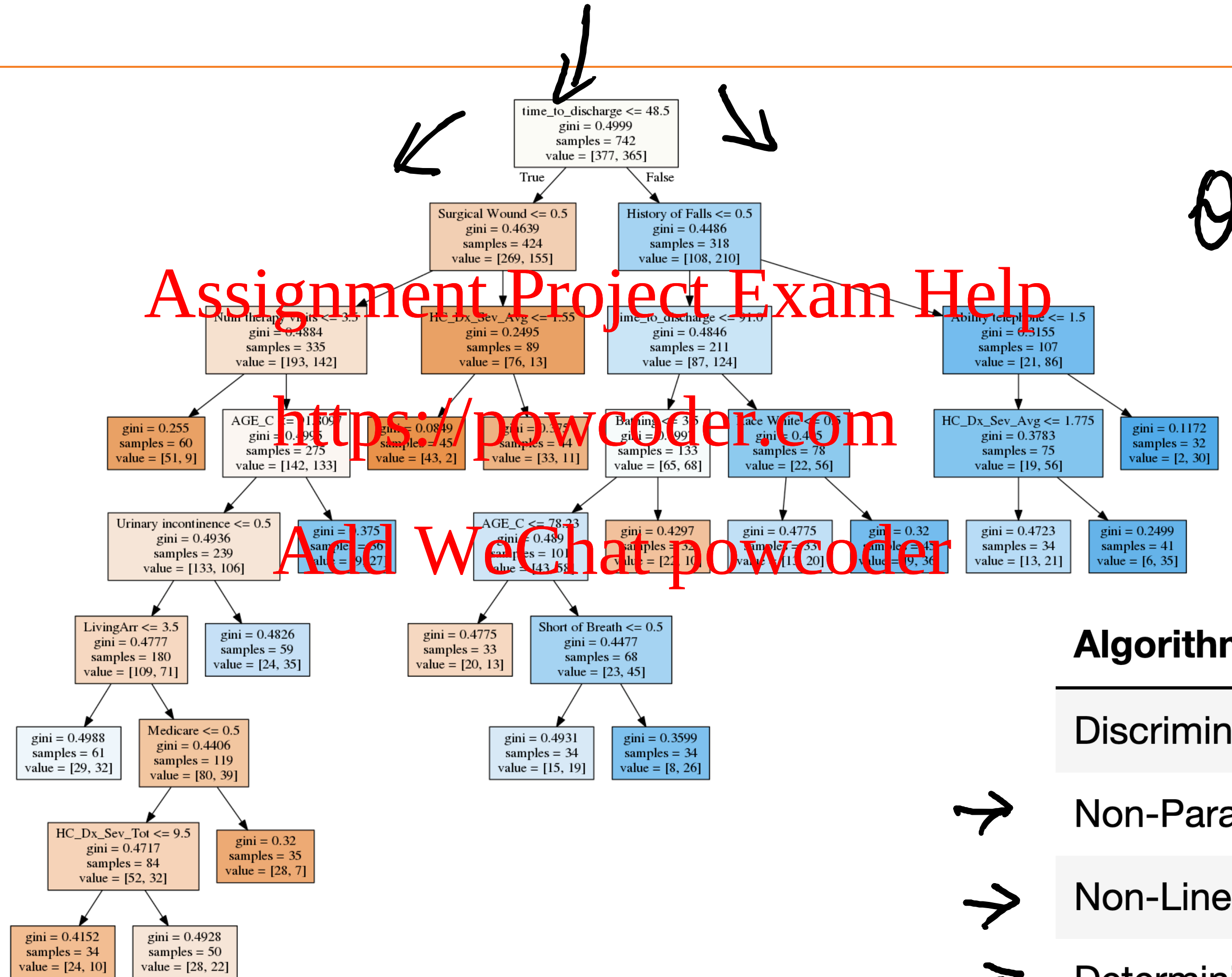
SOFTMAX REGRESSION



DECISION TREES



DECISION TREES



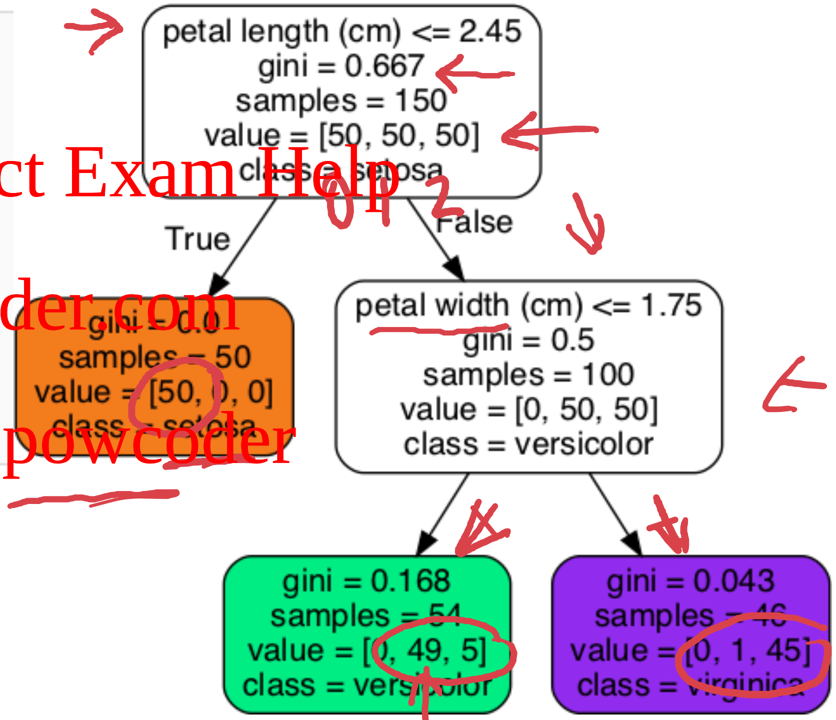
Algorithm Taxonomy
Discriminative
Non-Parametric
Non-Linear
Deterministic

CART

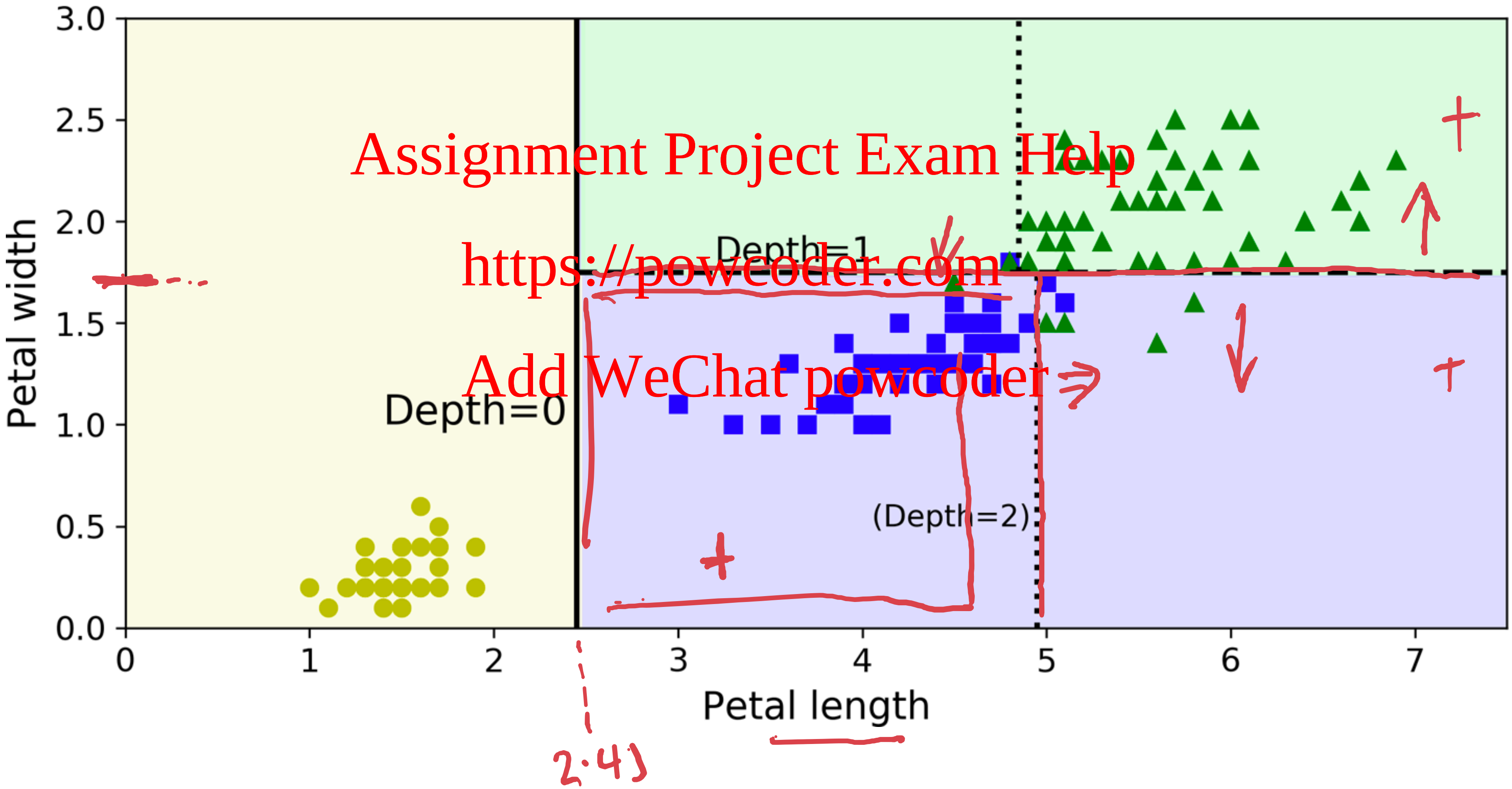
Classification Trees

DECISION TREE (IRIS DATA)

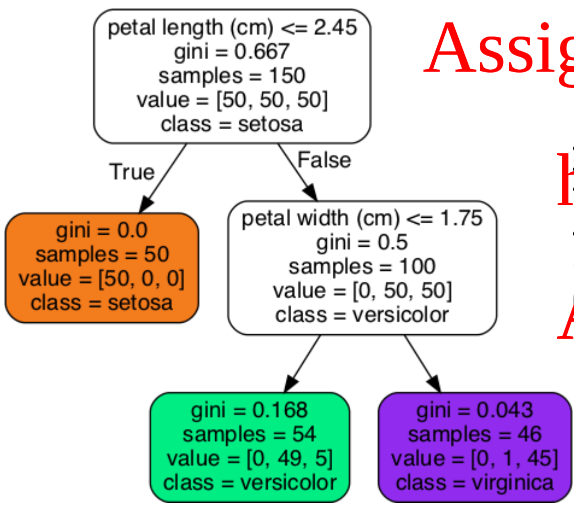

```python
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier

iris = load_iris()
X = iris.data[:, 2:] # petal length and width
y = iris.target

tree_clf = DecisionTreeClassifier(max_depth=2)
tree_clf.fit(X, y)
```

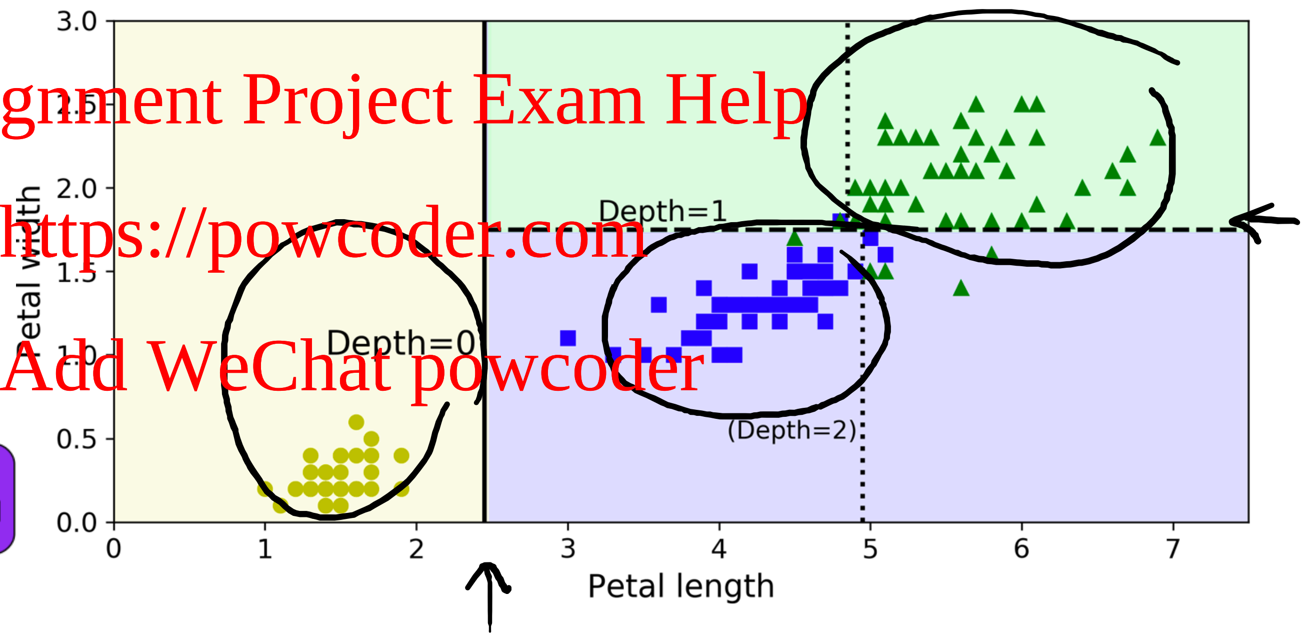


DECISION TREE BOUNDARIES



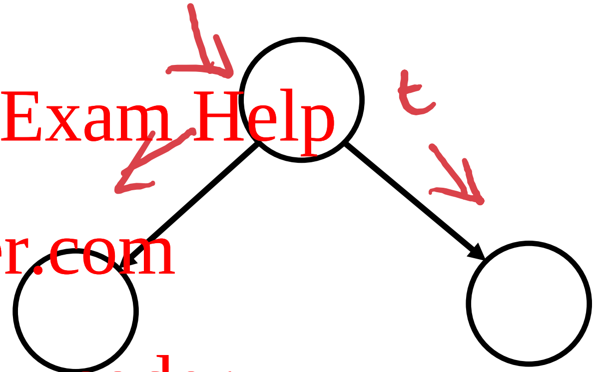
DECISION TREE BOUNDARIES





COST FUNCTIONS

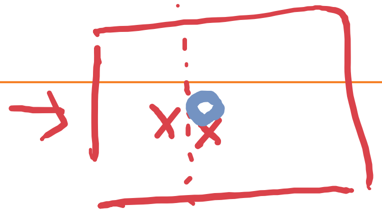
$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

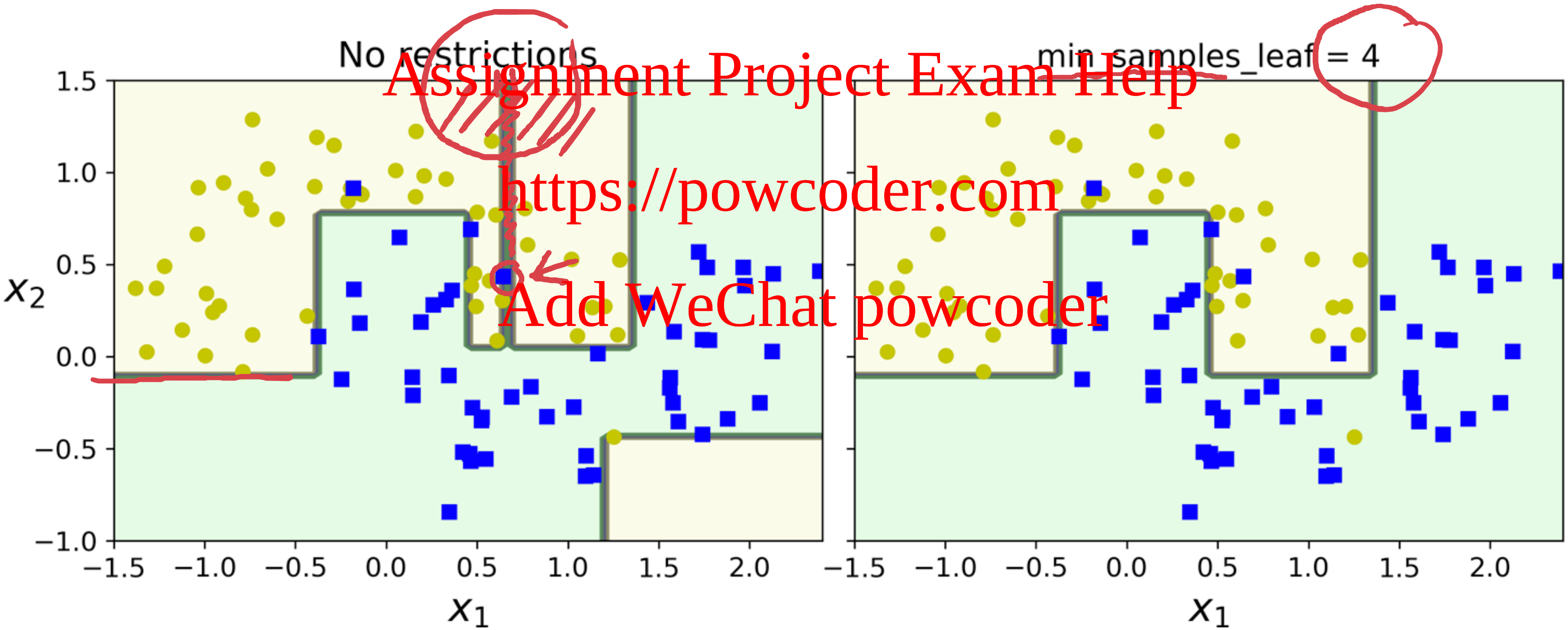


$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

where $\begin{cases} G_{\text{left/right}} \text{ measures the impurity of the left/right subset,} \\ m_{\text{left/right}} \text{ is the number of instances in the left/right subset.} \end{cases}$

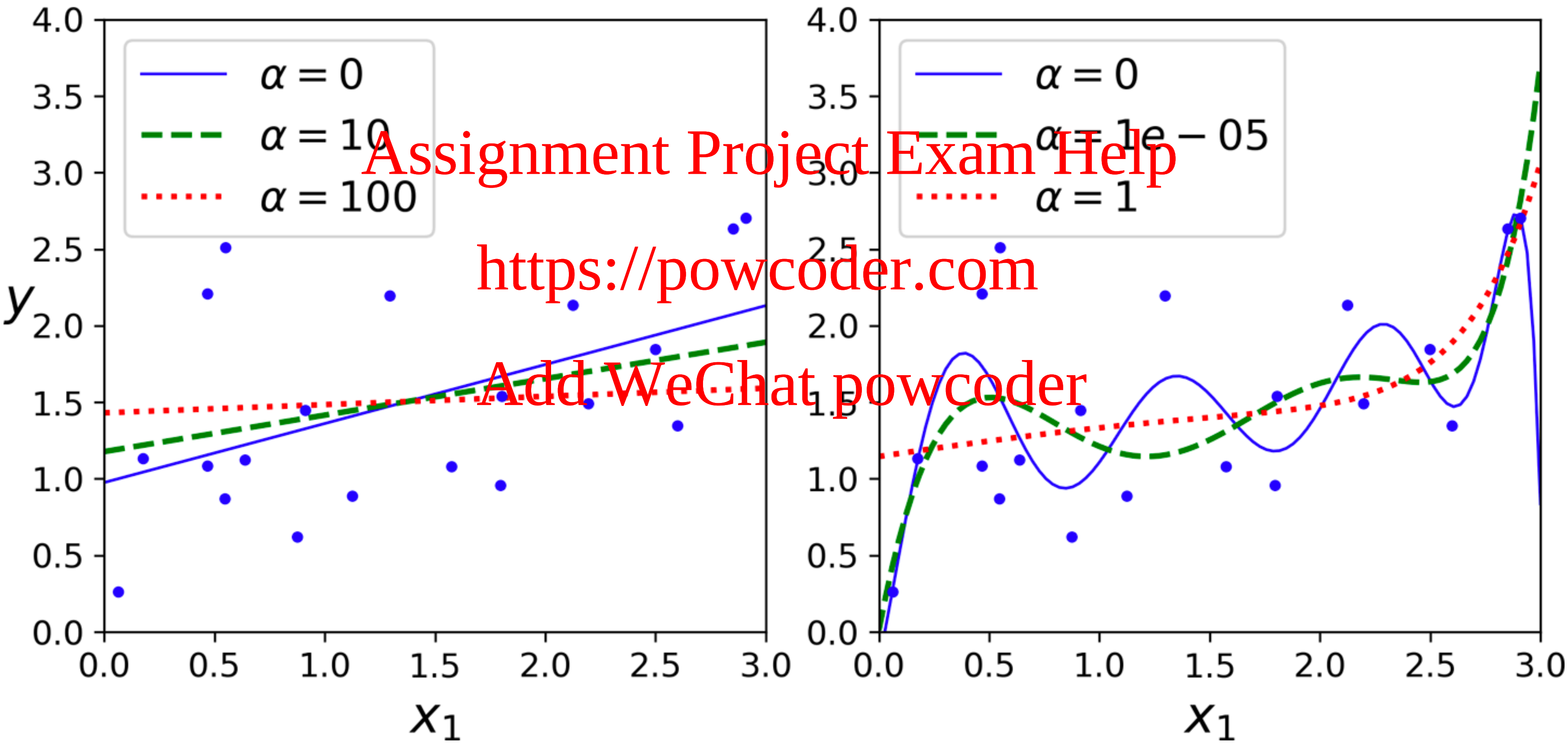
REGULARIZATION





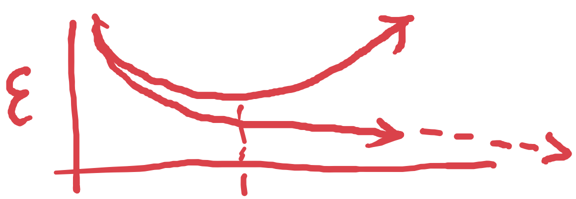
RIDGE (REGULARIZED) REGRESSION

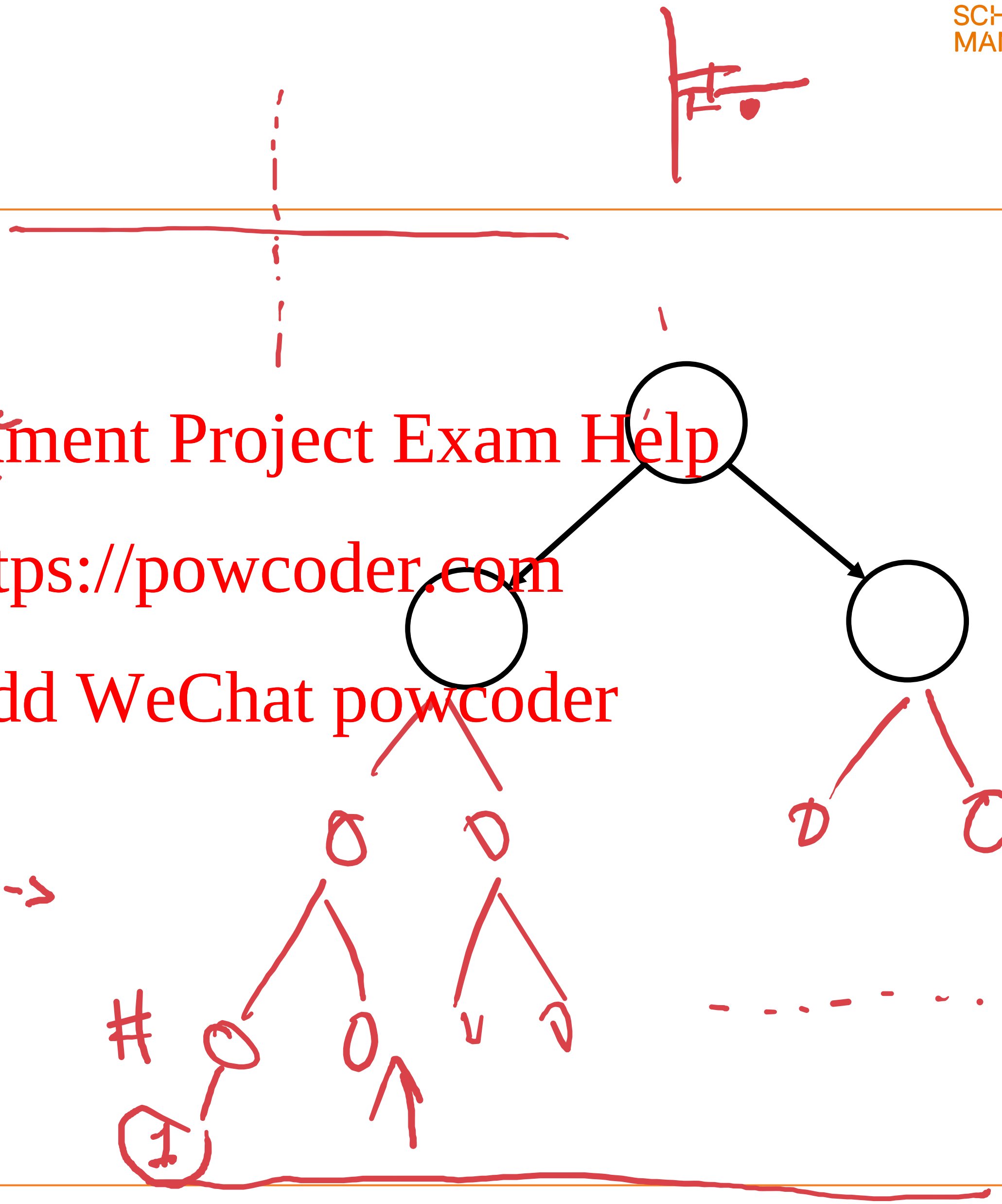
$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^{\,2}$$



REGULARIZATION

- k - features
- t_k - thresholds
- min_samples_split
- min_samples_leaf
- max_leaf_nodes
- max_features

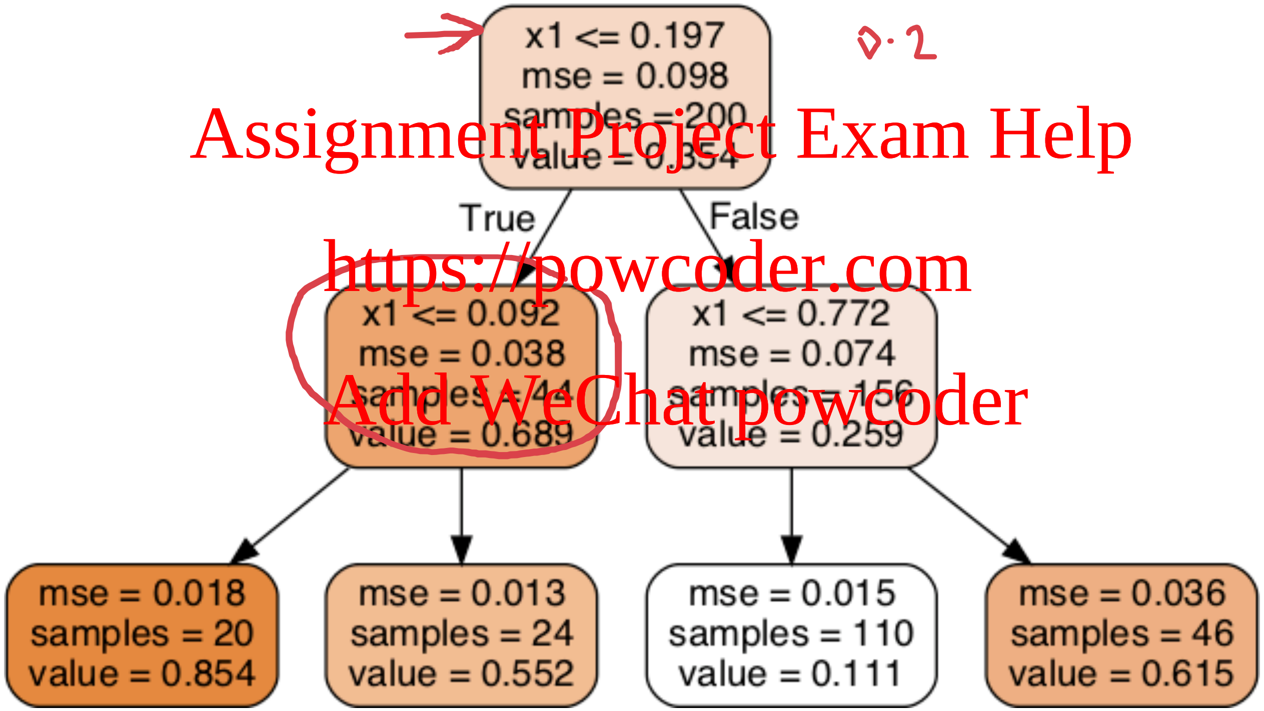




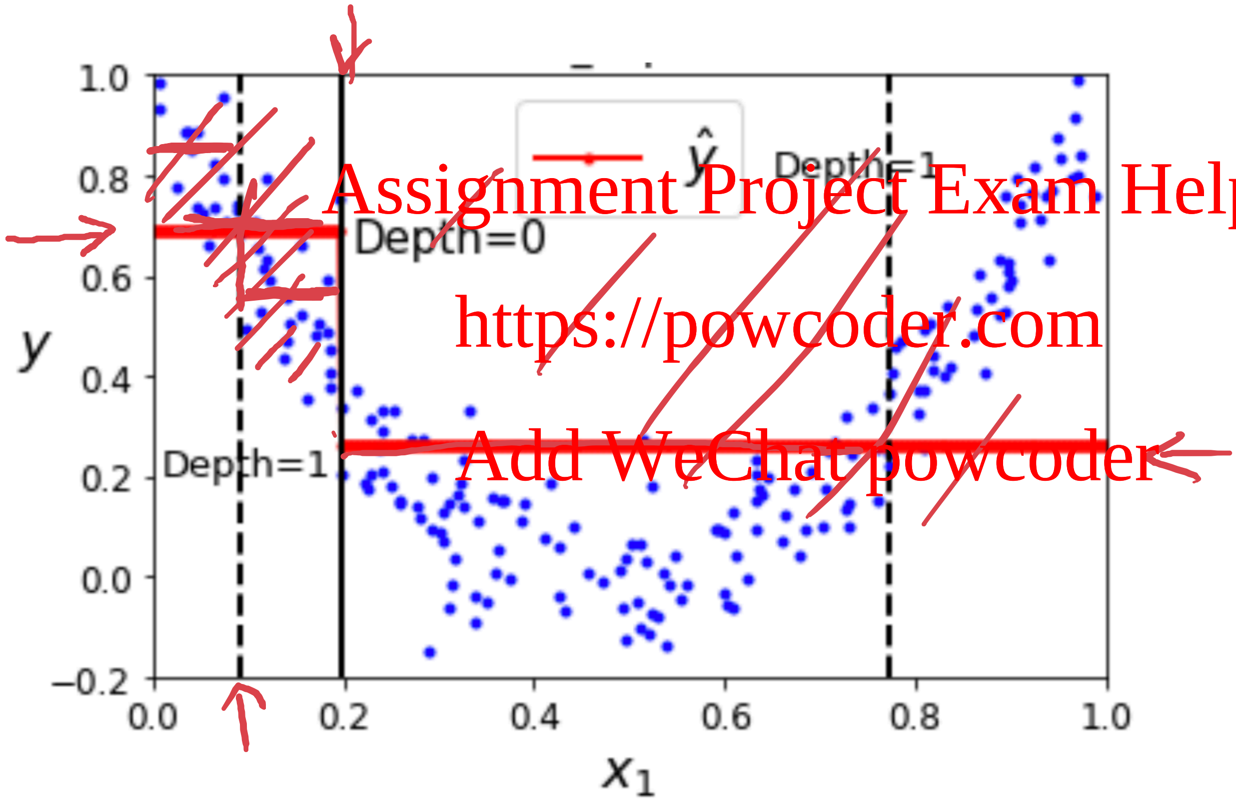
pruning

Regression Trees

REGRESSION TREES

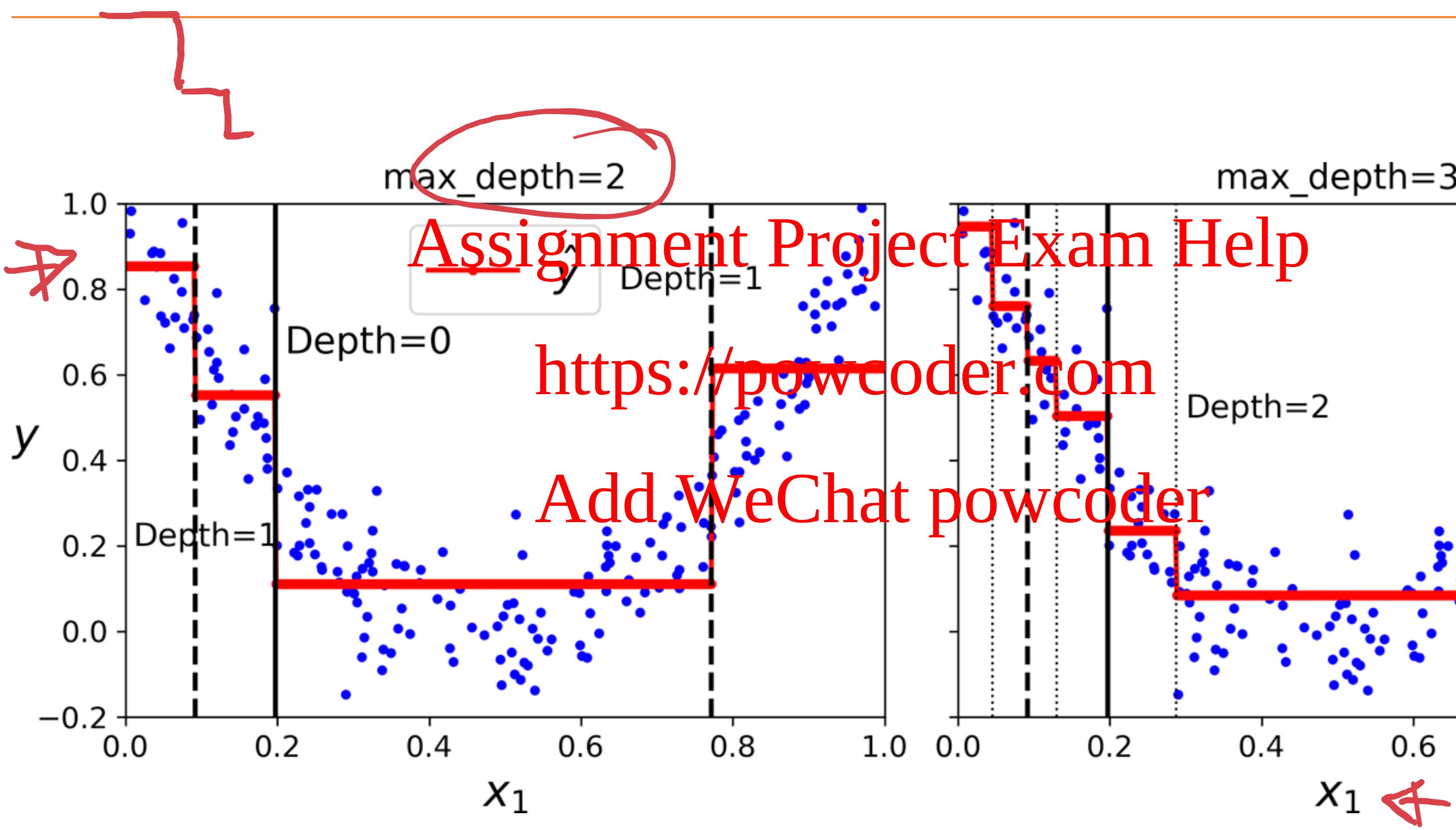


TREE REGRESSIONS



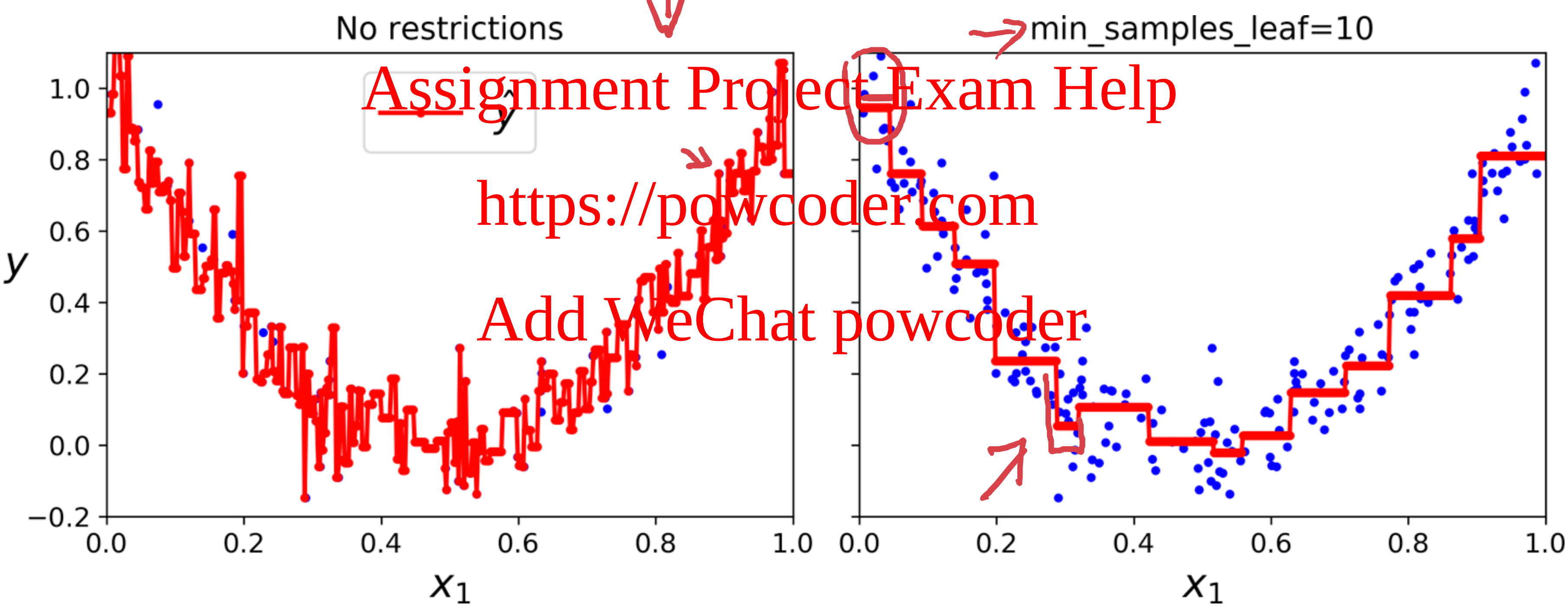
$$J(k, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right}$$

TREE REGRESSIONS



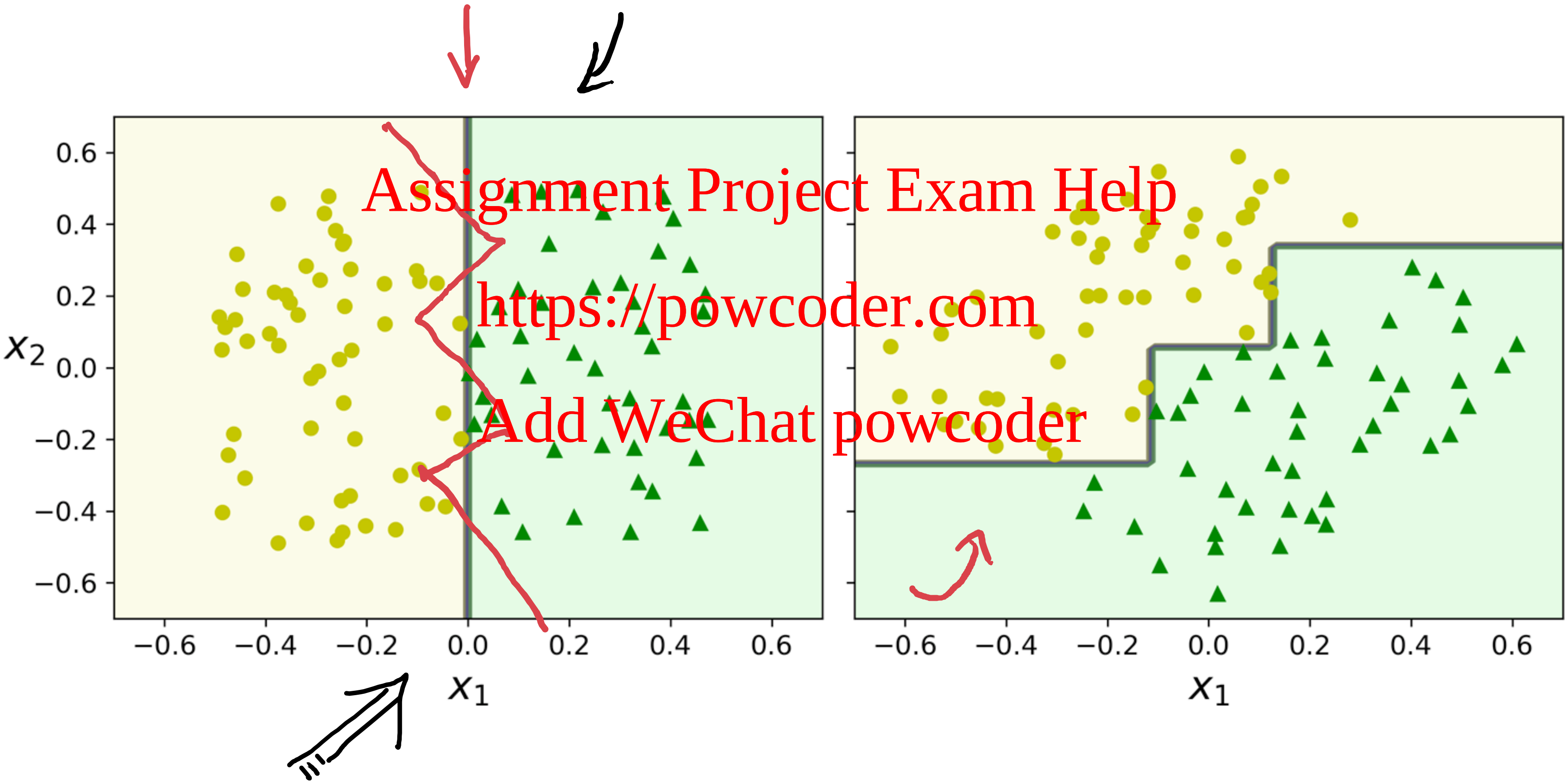
REGULARIZING A TREE REGRESSOR

UCL
SCHOOL OF
MANAGEMENT

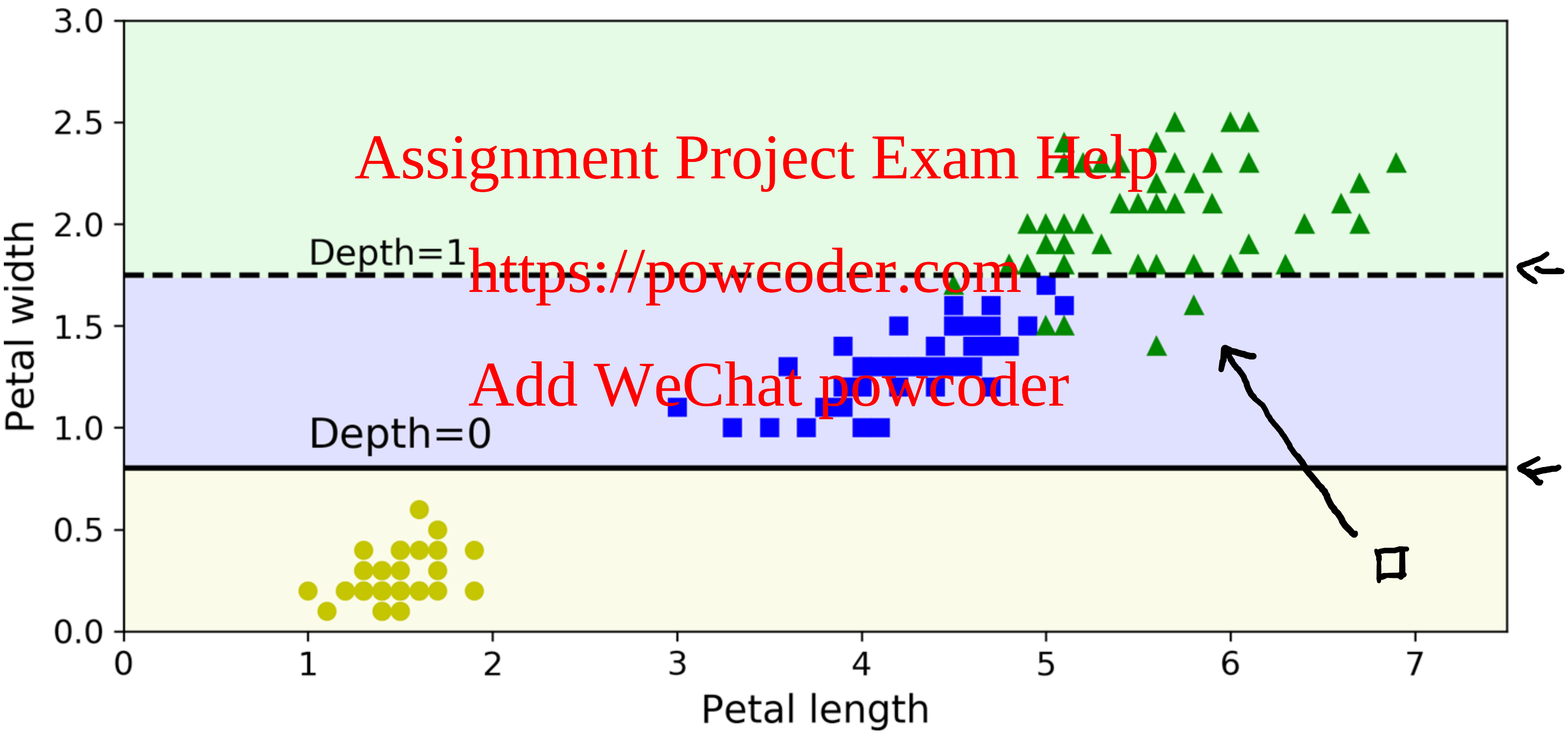


INSTABILITY

UCL SCHOOL OF MANAGEMENT



SENSITIVITY TO TRAINING SET



R2 D3

A visual introduction to machine learning

English

In machine learning, computers apply **statistical learning** techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions.

Keep scrolling. Using a data set about homes, we will create a machine learning model to distinguish homes in New York from homes in San Francisco.

Source: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

QUESTIONS

UCL
SCHOOL OF
MANAGEMENT

- If a Decision Tree is overfitting the training set, is it a good idea to try decreasing **max_depth**?

- If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?

LECTURE 3 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

A P MOORE