WEEK 2 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

LECTURE 2

A P MOORE

Review

MACHINE LEARNING JARGON

UCL SCHOOL OF MANAGEMENT

- Model
- Interpolating / Extrapolating
- Data Bias
- Noise / Outliers
- Learning algorithm
- Inference algorithm
- **Supervised learning**
- Unsupervised learning
- **Classification**
- **Regression**
- Clustering
- Decomposition

- Parameters
- Optimisation
- Training data
- Testing data
- Error metric
- Linear model
- Parametric model
- Model variance
- Model bias
- Model generalization
- Overfitting
- Goodness-of-fit
- Hyper-parameters

- Failure modes
- Confusion matrix
- True Positive
- False Negative
- Data density
- Partition
- Hidden parameter
- High dimensional space
- Low dimensional space
- Separable data
- Manifold / Decision surface
- Hyper cube / volume / plane

机器学习行话

UCL SCHOOL OF MANAGEMENT

- 模型
- 内插 / 外推
- 数据偏差
- 噪声 / 离群值
- 学习算法
- 推断算法
- 监督学习
- 无监督学习
- 分类
- 回归
- 聚类
- 分解

- 参数
- 优化
- 训练数据
- 测试数据
- 误差指标
- 线性模型
- 参数模型
- 模型方差
- 模型偏差
- 模型泛化
- 过拟合
- 拟合优度
- 超参数

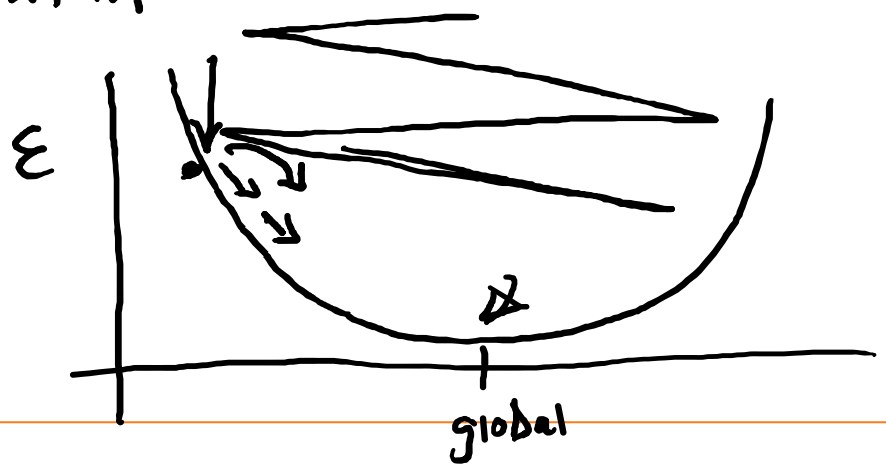
- 失败模式
- 混淆矩阵
- 真正例
- 假反例
- 数据密度
- 划分
- 隐藏参数
- 高维空间
- 低维空间
- 可分数据
- 流形/ 决策面
- 超立方体 / 超体积 / 超平面

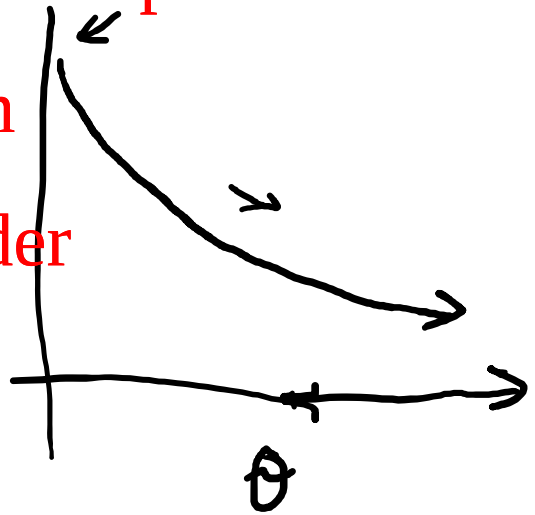
OPTIMISATION / 优化

YŌUHUÀ

UCL SCHOOL OF MANAGEMENT

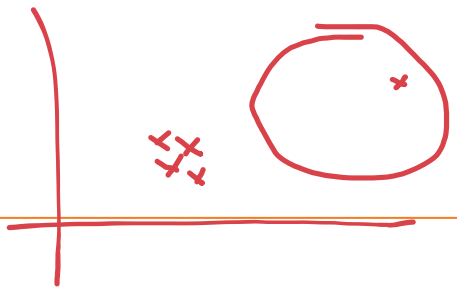
- gradient decent
- initial

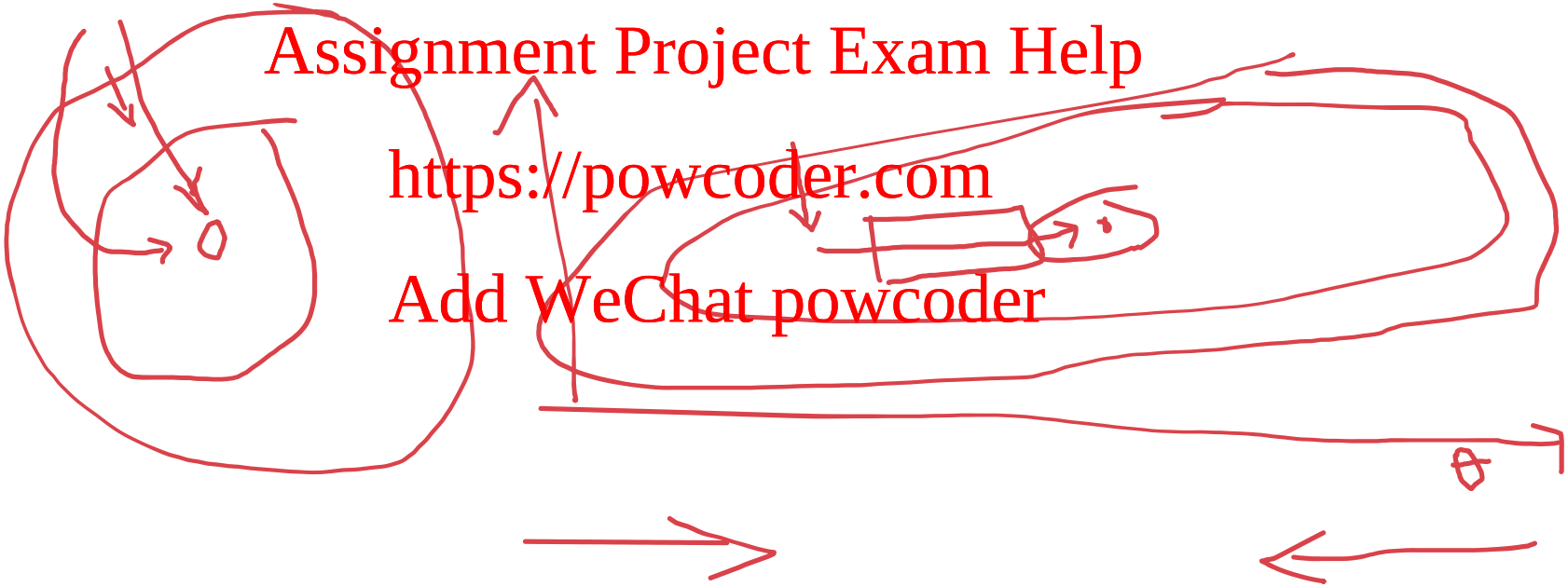




SCALING

UCL
SCHOOL OF
MANAGEMENT

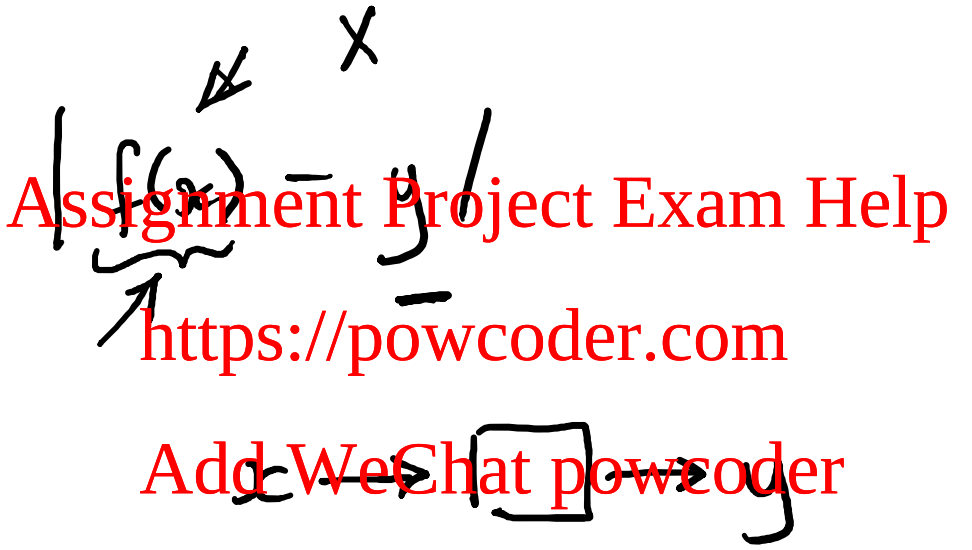




ERROR METRIC / 误差指标

WÙCHĀ ZHǏBIĀO

UCL SCHOOL OF MANAGEMENT



HYPER-PARAMETERS / 超参数

CHĀO CĀNSHÙ

UCL SCHOOL OF MANAGEMENT

learning rate

batch size

PARAMETRIC MODEL / 参数模型

CĀNSHÙ MÓXÍNG

UCL SCHOOL OF MANAGEMENT

MODEL BIAS / 模型偏差

MÓXÍNG PIĀNCHĀ

UCL SCHOOL OF MANAGEMENT

assumptions

MODEL VARIANCE / 模型方差

MÓXÍNG FĀNGCHĀ

UCL SCHOOL OF MANAGEMENT

MODEL GENERALIZATION / 模型泛化

MÓXÍNG FÀN HUÀ

UCL SCHOOL OF MANAGEMENT

OVERFITTING / 过拟合

GUÒ NǏ HÉ

UCL
SCHOOL OF
MANAGEMENT

FAILURE MODES / 失败模式

SHĪBÀI MÓSHÌ

UCL SCHOOL OF MANAGEMENT

COMMON CLASSIFICATION METRICS

UCL SCHOOL OF MANAGEMENT

- Accuracy
- Precision (P)
- Recall (R)
- F1 score (F1)
- Area under the ROC (Receiver Operating Characteristic) curve or simply AUC (AUC) - Log loss
- Precision at k (P@k)
- Average precision at k (AP@k)
- Mean average precision at k (MAP@k)

COMMON REGRESSION METRICS

UCL SCHOOL OF MANAGEMENT

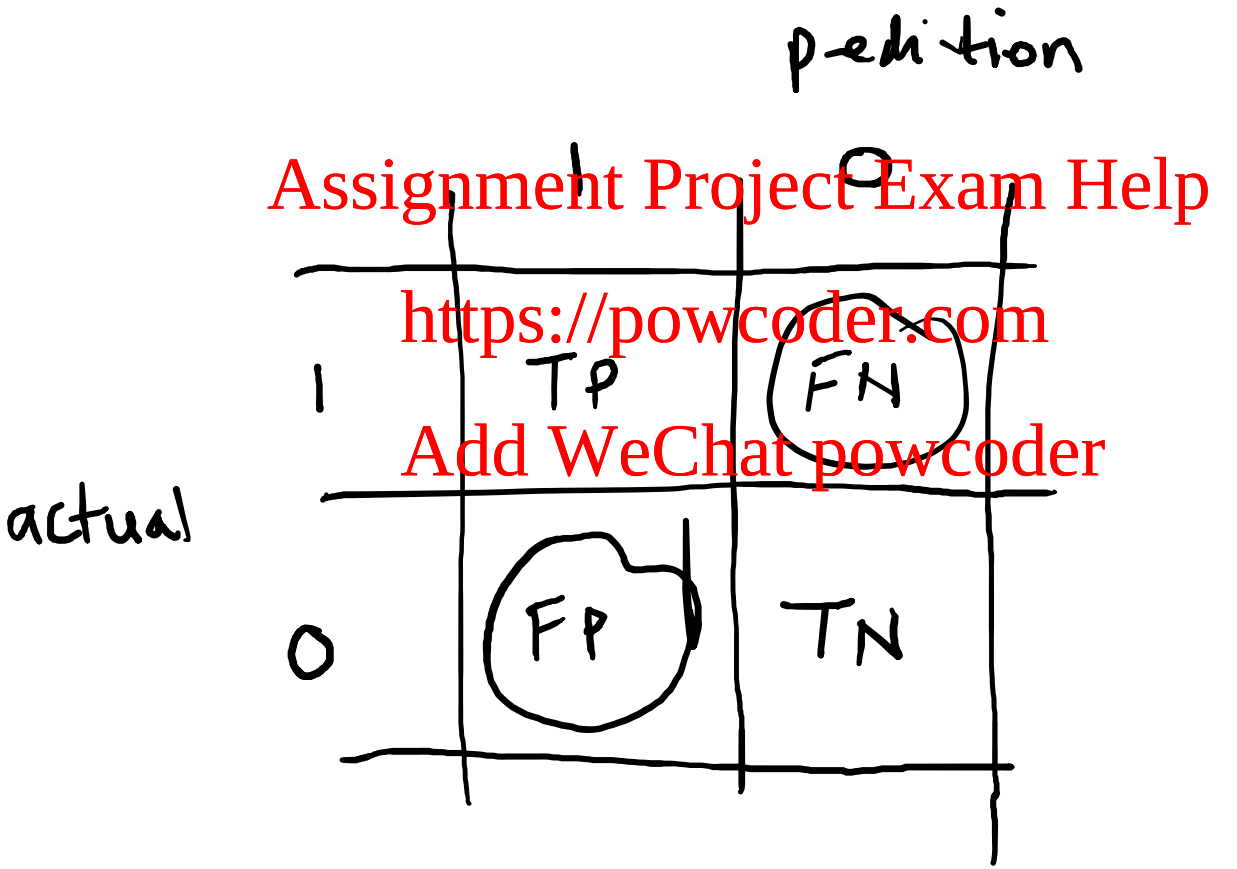
- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean squared error (RMSE)
- Root mean squared logarithmic error (RMSLE)
- Mean percentage error (MPE)
- Mean absolute percentage error (MAPE)
- R2

Measuring performance

CONFUSION MATRIX / 混淆矩阵

HÙNXIÁO JǓZHÈN

UCL SCHOOL OF MANAGEMENT

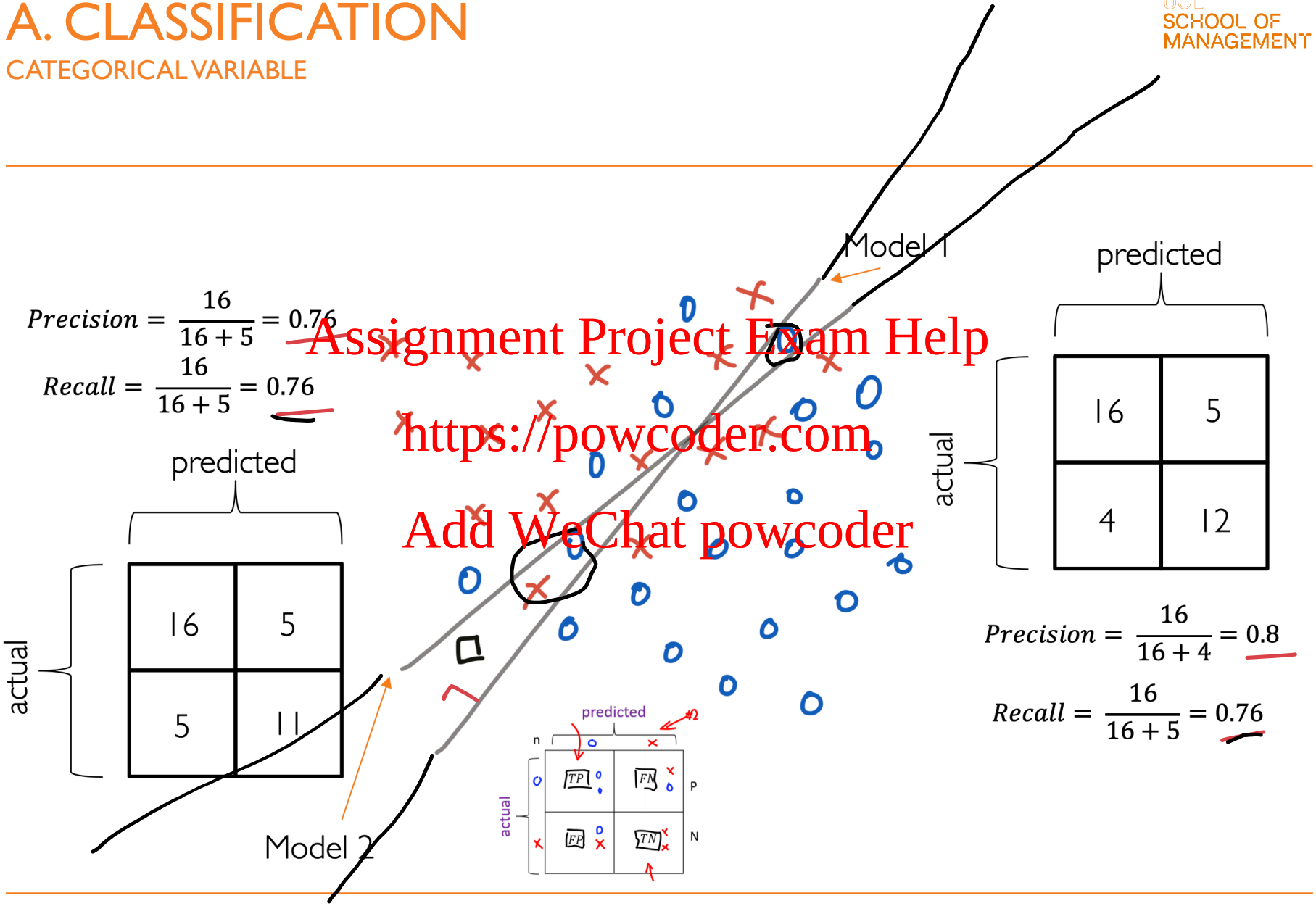



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

A. CLASSIFICATION

CATEGORICAL VARIABLE

UCL SCHOOL OF MANAGEMENT



$$Precision = \frac{16}{16+5} = 0.76$$

$$Recall = \frac{16}{16+5} = 0.76$$

$$Precision = \frac{16}{16+4} = 0.8$$

$$Recall = \frac{16}{16+5} = 0.76$$

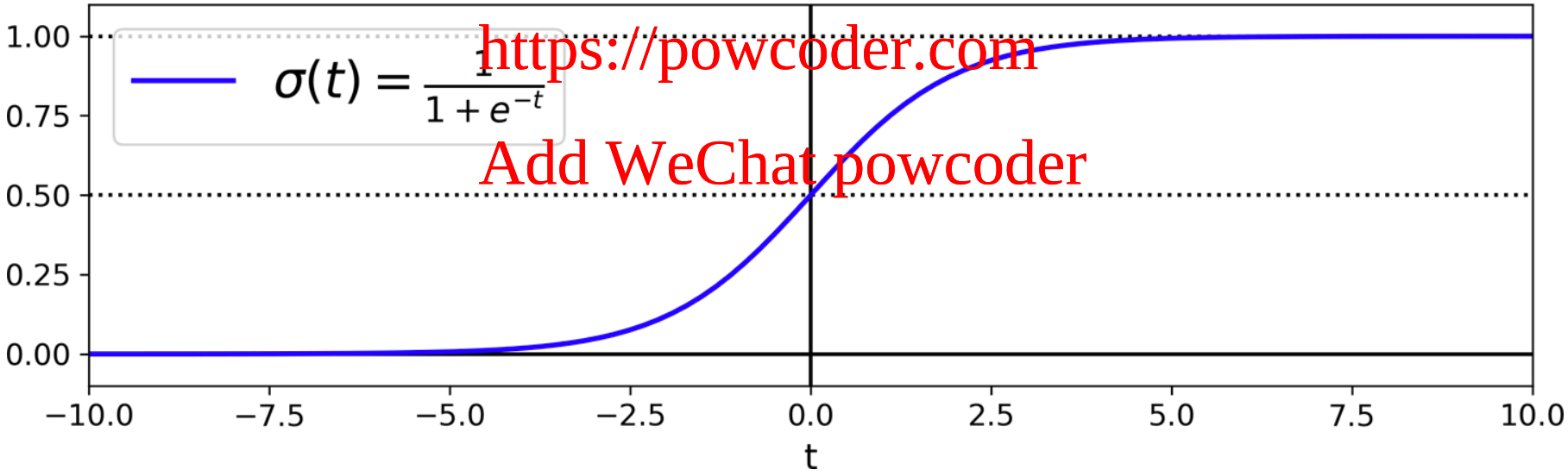
Logistic Regression

LOGISTIC REGRESSION (CLASSIFICATION !!!!!)

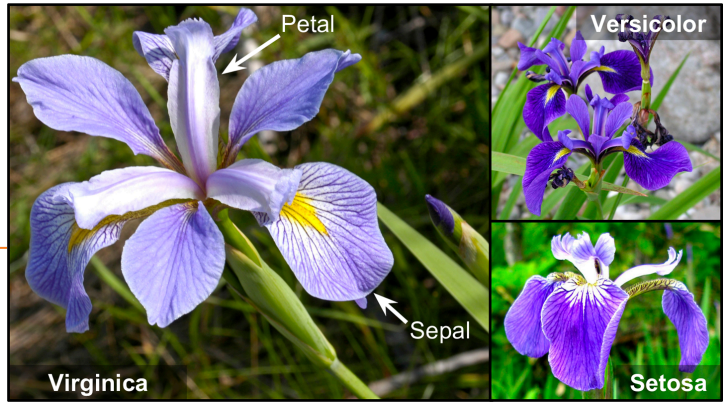
UCL SCHOOL OF MANAGEMENT

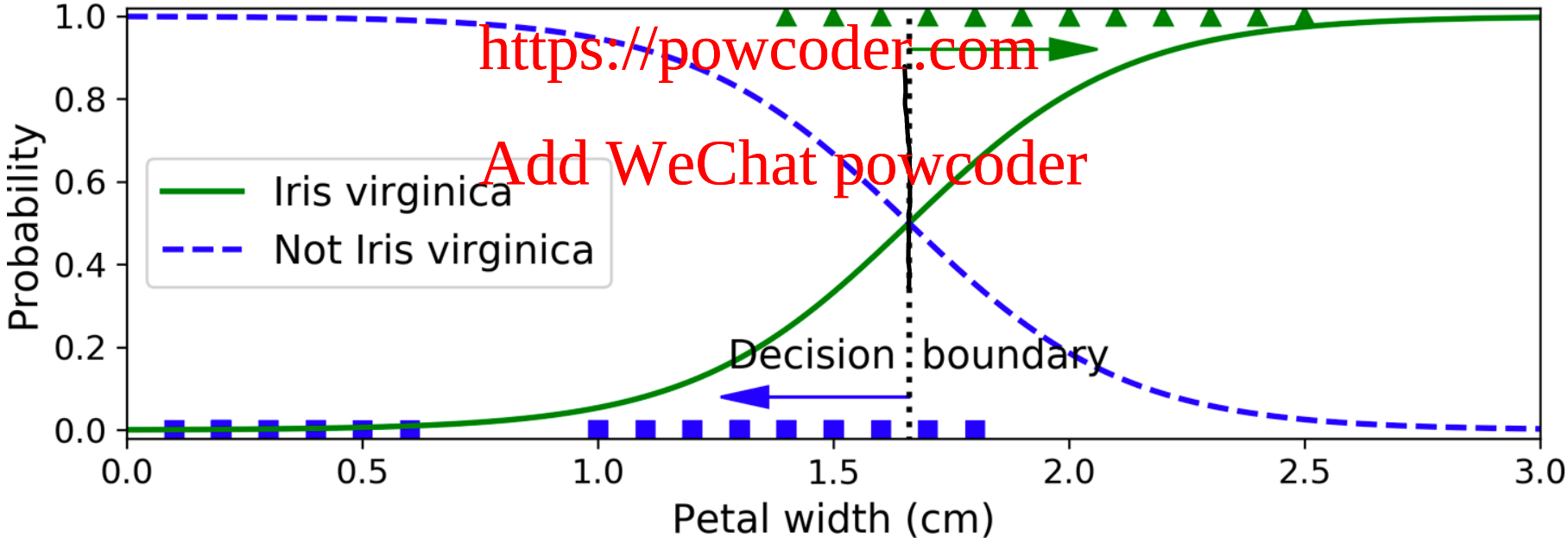
$$\hat{p} = h_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma\left(\mathbf{x}^{\intercal}\boldsymbol{\theta}\right)$$

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$



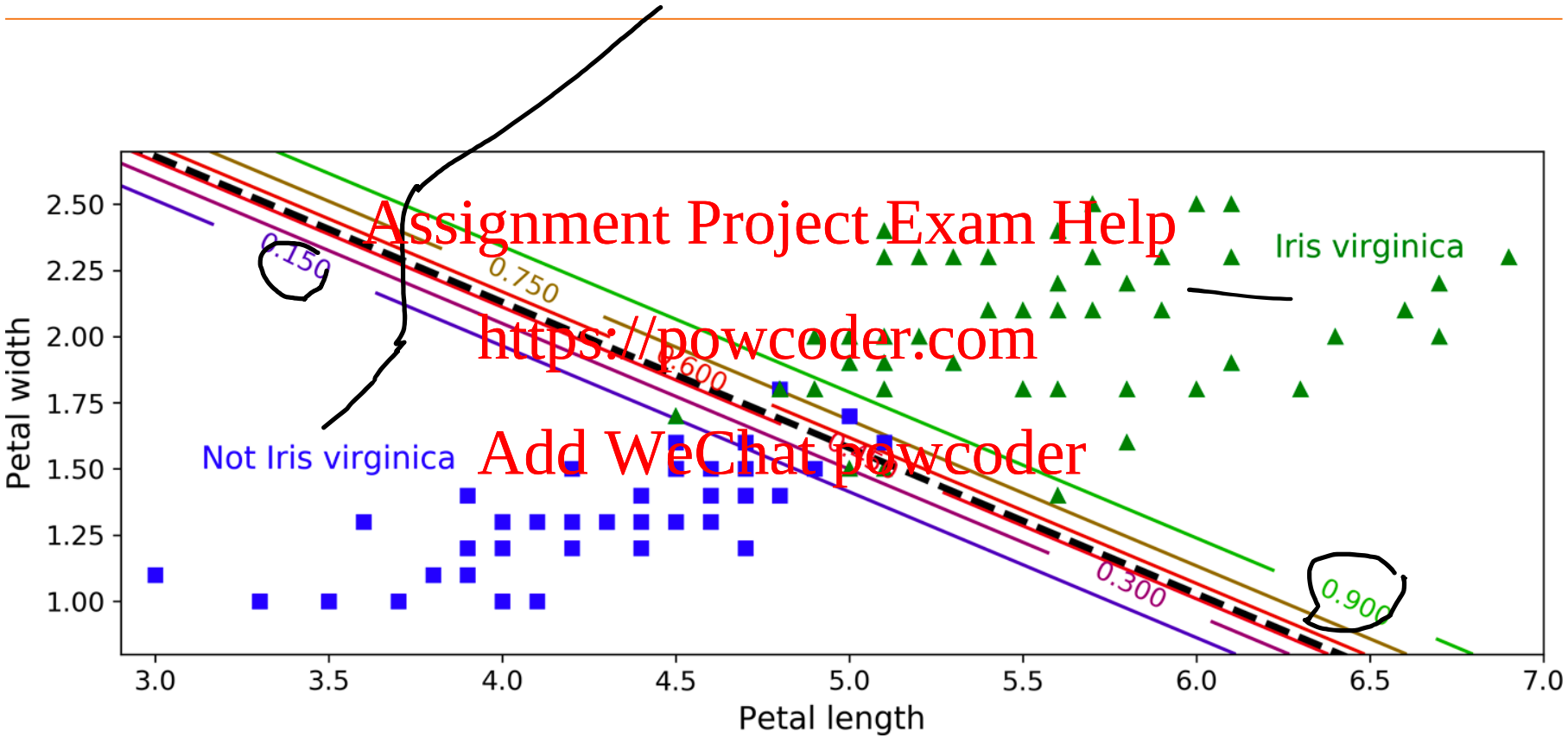
DECISION BOUNDARIES





DECISION BOUNDARIES

UCL SCHOOL OF MANAGEMENT



Problem 1

~15 mins group work

~15 mins discussion

REVIEW

UCL SCHOOL OF MANAGEMENT

- Select good metrics for classification tasks
- How to pick the appropriate precision/recall trade-off
- How to compare classifiers
- Different classification systems for a variety of tasks

- **What business problems can you think of that are classification tasks?**
- **Can you think of some business problems that are multilabel and multioutput?**

The Machine Learning Canvas (v0.4)

Designed for: | Designed by: | Date: | Iteration:

Decisions

How are predictions used to make decisions that provide the proposed value to the end-user?

ML task

Input, output to predict, type of problem.

Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

Data Sources

Which raw data sources can we use (internal and external)?

Collecting Data

How do we get new data to learn from (inputs and outputs)?

Making Predictions

When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?

Offline Evaluation

Methods and metrics to evaluate the system before deployment.

Features

Input representations extracted from raw data sources.

Building Models

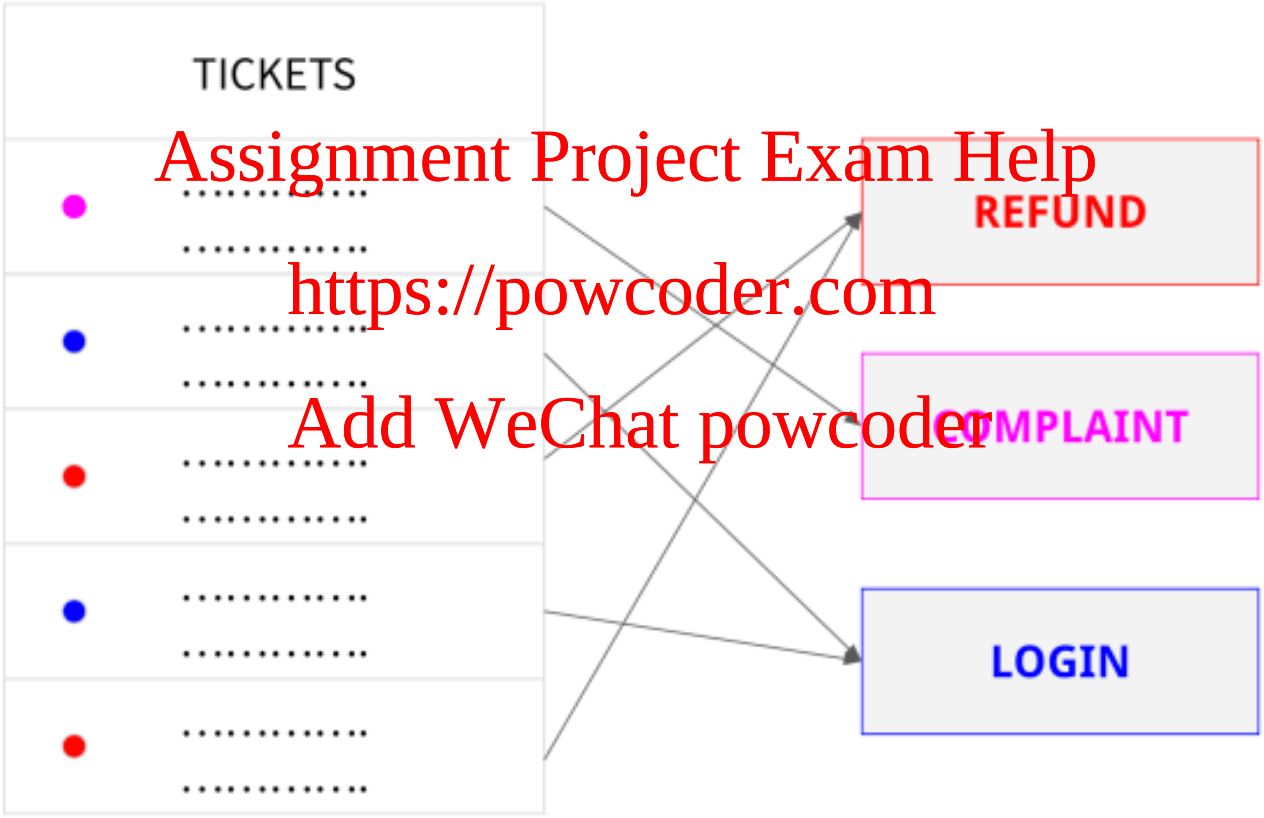
When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?

Live Evaluation and Monitoring

Methods and metrics to evaluate the system after deployment, and to quantify value creation.

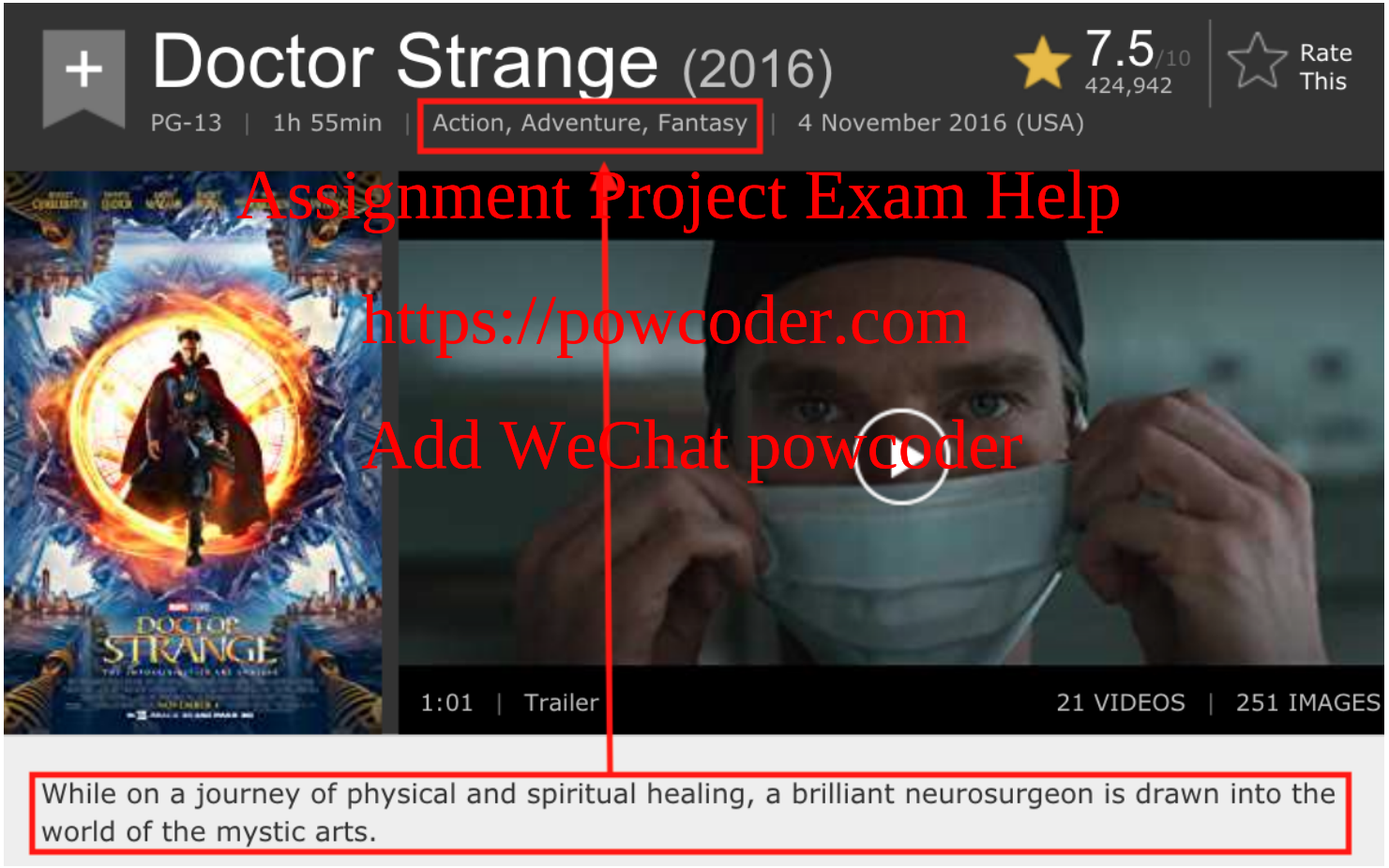
TEXT CATEGORIZATION

UCL
SCHOOL OF
MANAGEMENT



FILM GENRE CLASSIFICATION

UCL SCHOOL OF MANAGEMENT



https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff

MULTI-OUTPUT LEARNING

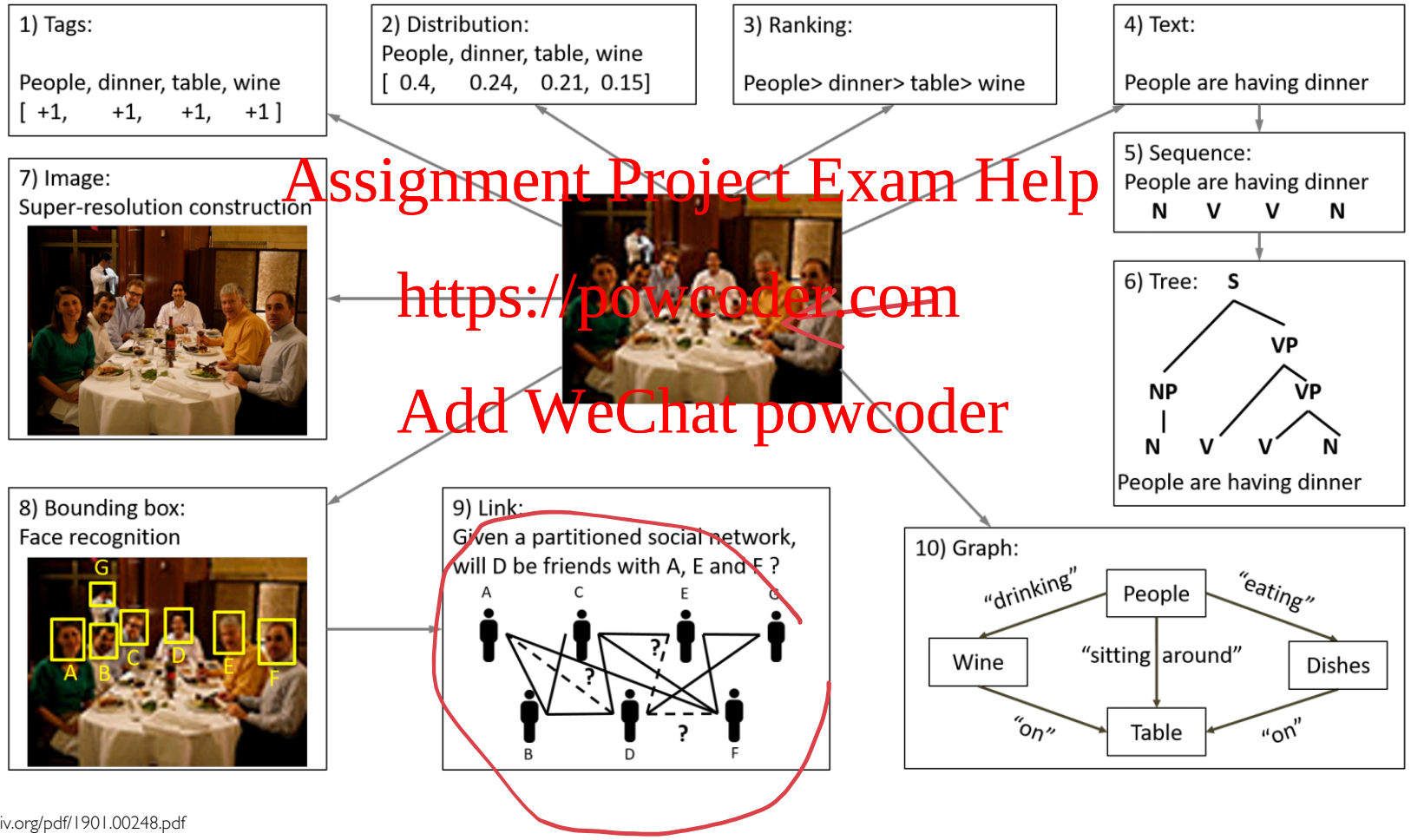
UCL SCHOOL OF MANAGEMENT

Subfield	Output Structure	Application	Discipline
Multi-label Learning	Independent Binary Vector	Document Categorization [19]	Natural Language Processing
		Semantic Scene Classification [20]	Computer Vision
		Automatic Video Annotation [21]	Computer Vision
Multi-target Regression	Independent Real-valued Vector	River Quality Prediction [22]	Ecology
		Natural Gas Demand Forecasting [23]	Energy Meteorology
		Drug Efficacy Prediction [24]	Medicine
Label Distribution Learning	Distribution	Head Pose Estimation [25]	Computer Vision
		Facial Age Estimation [26]	Computer Vision
		Text Mining [27]	Data Mining
Label Ranking	Ranking	Web Categorization Ranking [28]	Information Retrieval
		Question Answering [29]	Information Retrieval
		Visual Object Recognition [30]	Computer Vision
Sequence Alignment Learning	Sequence	Protein Function Prediction [31]	Bioinformatics
		Language Translation [32]	Natural Language Processing
		Named Entity Recognition [33]	Natural Language Processing
Network Analysis	Graph	Scene Graph [34]	Computer Vision
	Tree	Natural Language Parsing [35]	Natural Language Processing
	Link	Link Prediction [36]	Data Mining
Data Generation	Image	Super-resolution Image Reconstruction [37]	Computer Vision
	Text	Language Generation	Natural Language Processing
	Audio	Music Generation [38]	Signal Processing
Semantic Retrieval	Independent Real-valued Vector	Content-based Image Retrieval [39]	Computer Vision
		Microblog Retrieval [40]	Data Mining
		News Retrieval [41]	Data Mining
Time-series Prediction	Time Series	DNA Microarray Data Analysis [42]	Bioinformatics
		Energy Consumption Forecasting [43]	Energy Meteorology
		Video Surveillance [44]	Computer Vision

https://arxiv.org/pdf/1901.00248.pdf

OUTPUT STRUCTURES

UCL SCHOOL OF MANAGEMENT

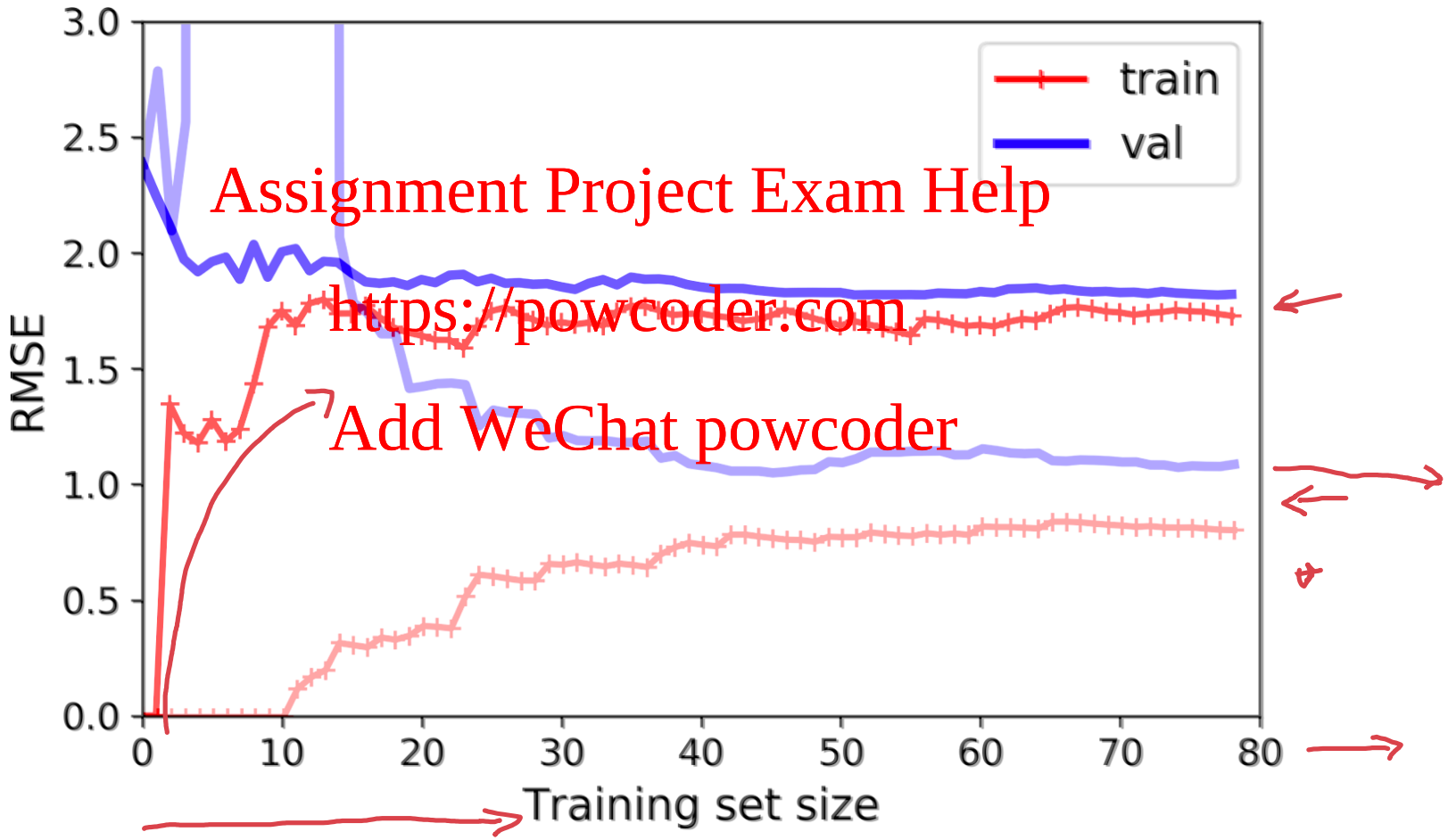


https://arxiv.org/pdf/1901.00248.pdf

Decision boundaries

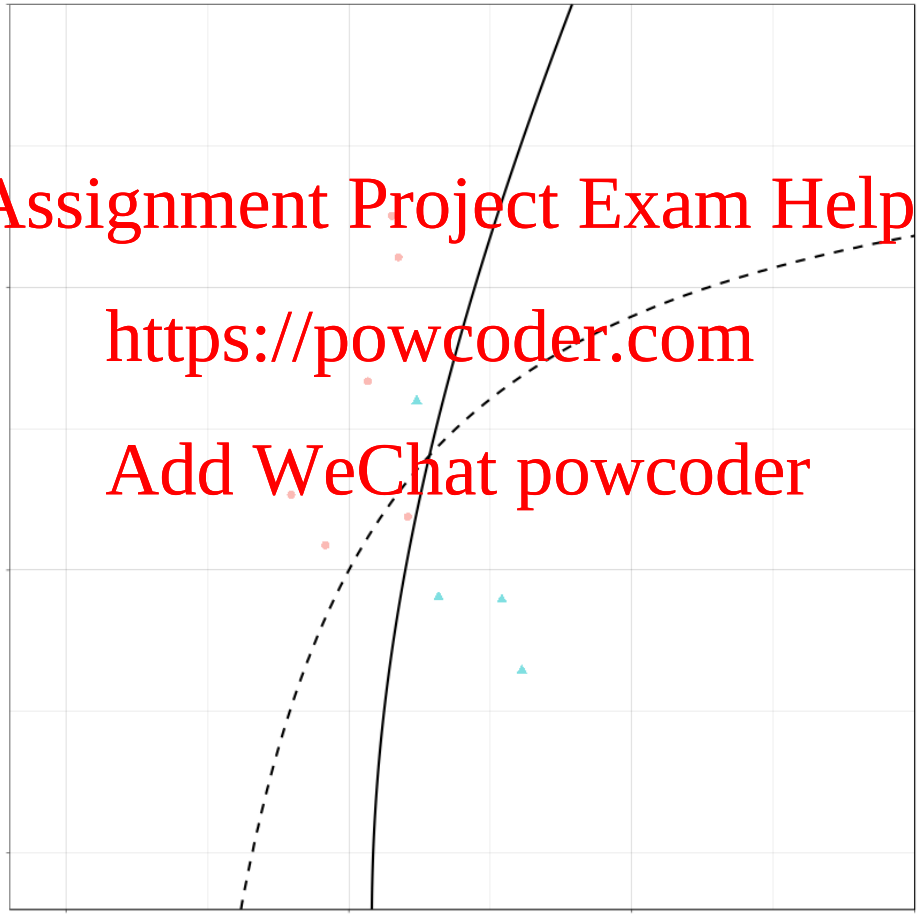
LEARNING CURVES

UCL SCHOOL OF MANAGEMENT



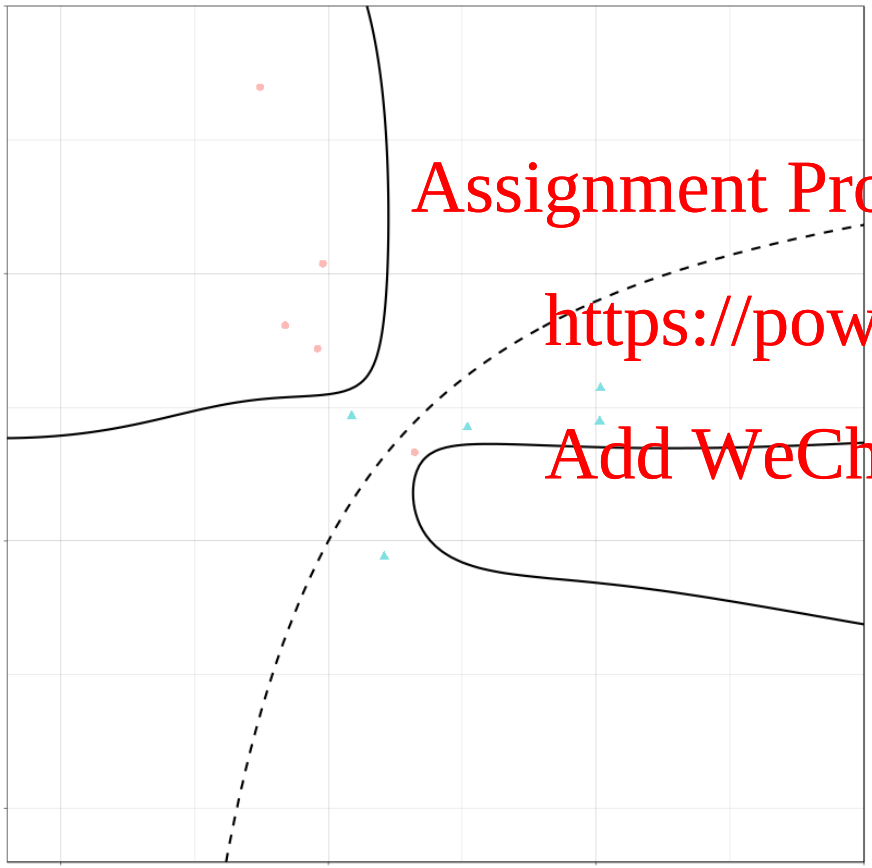
DECISION BOUNDARIES

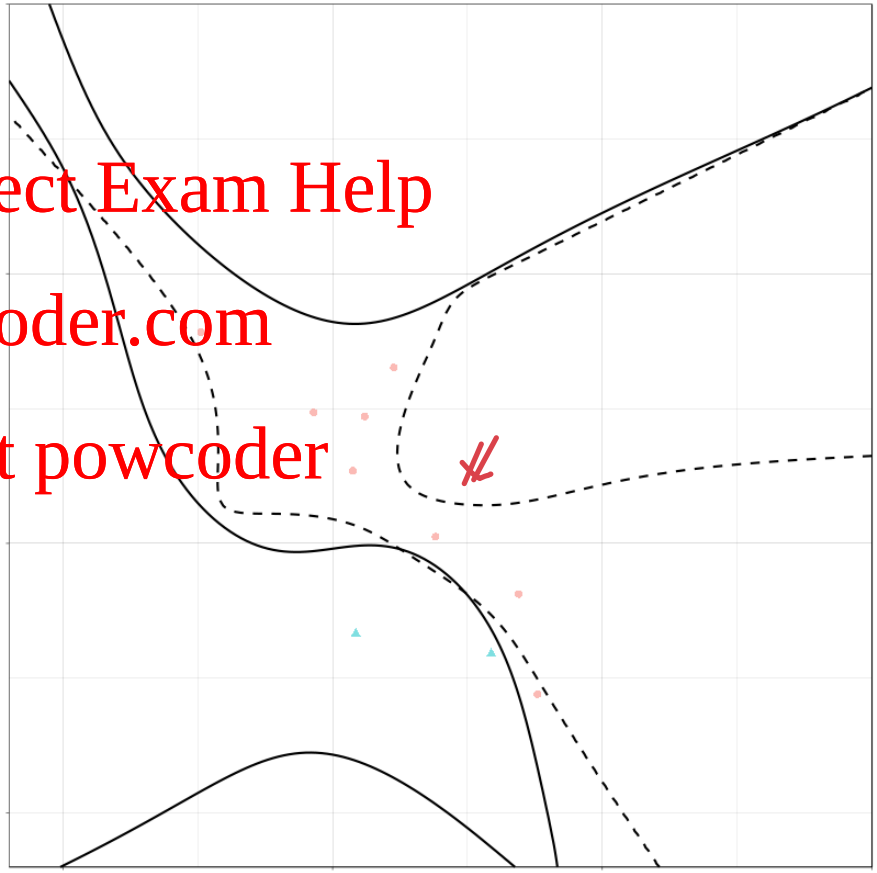
UCL
SCHOOL OF
MANAGEMENT



DECISION BOUNDARIES

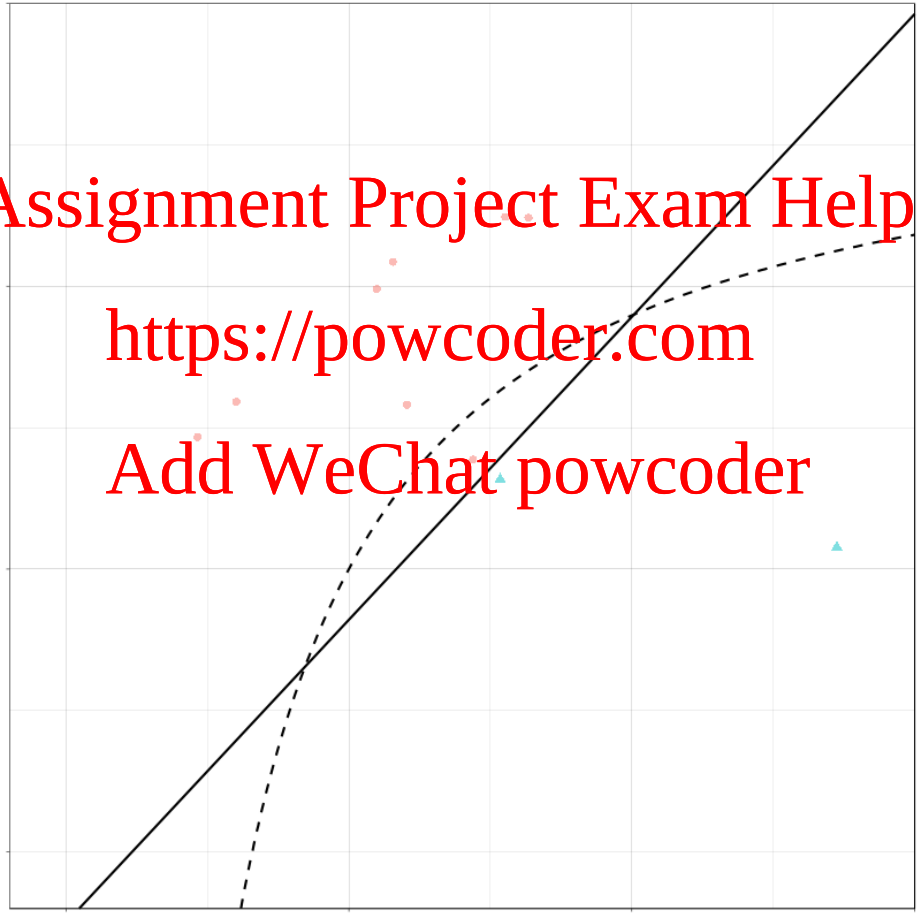
UCL
SCHOOL OF
MANAGEMENT





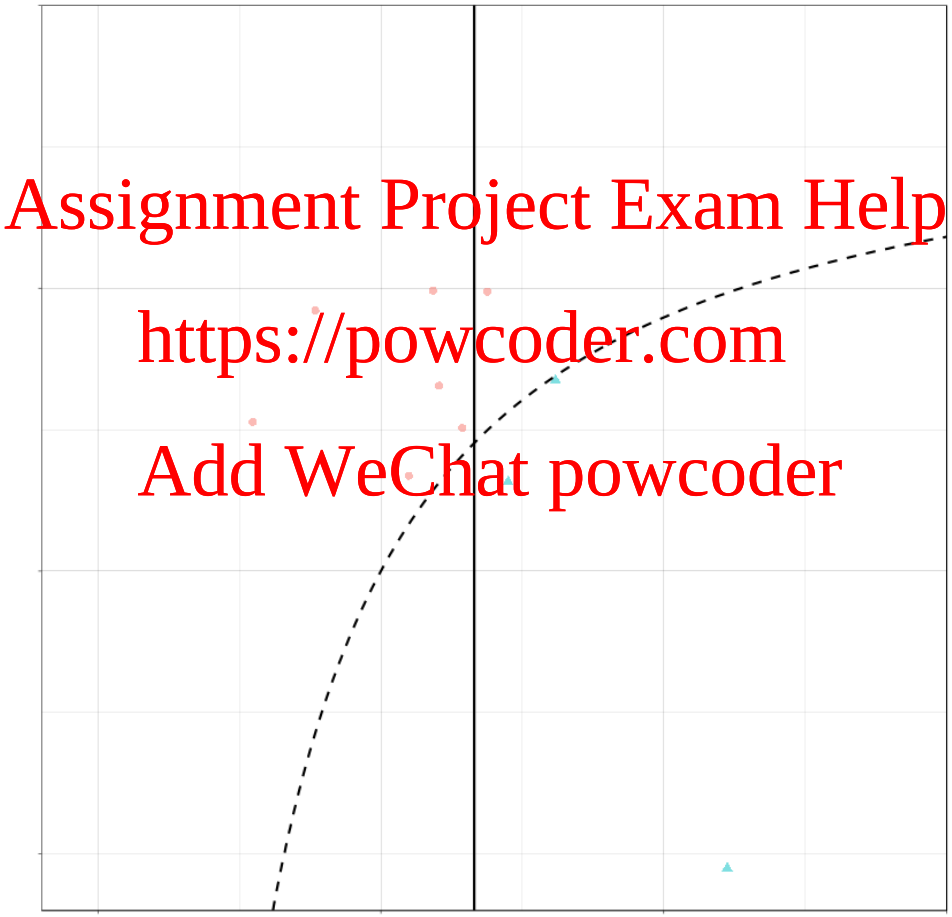
DECISION BOUNDARIES

UCL
SCHOOL OF
MANAGEMENT



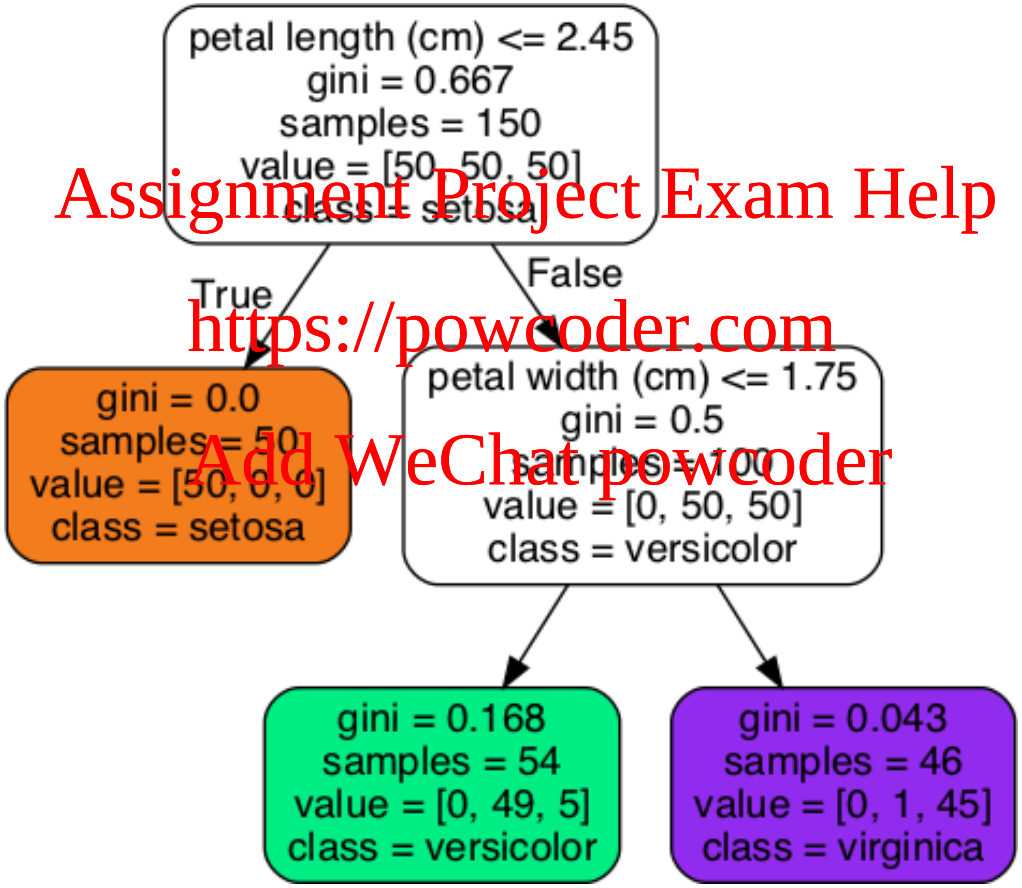
DECISION BOUNDARIES

UCL
SCHOOL OF
MANAGEMENT



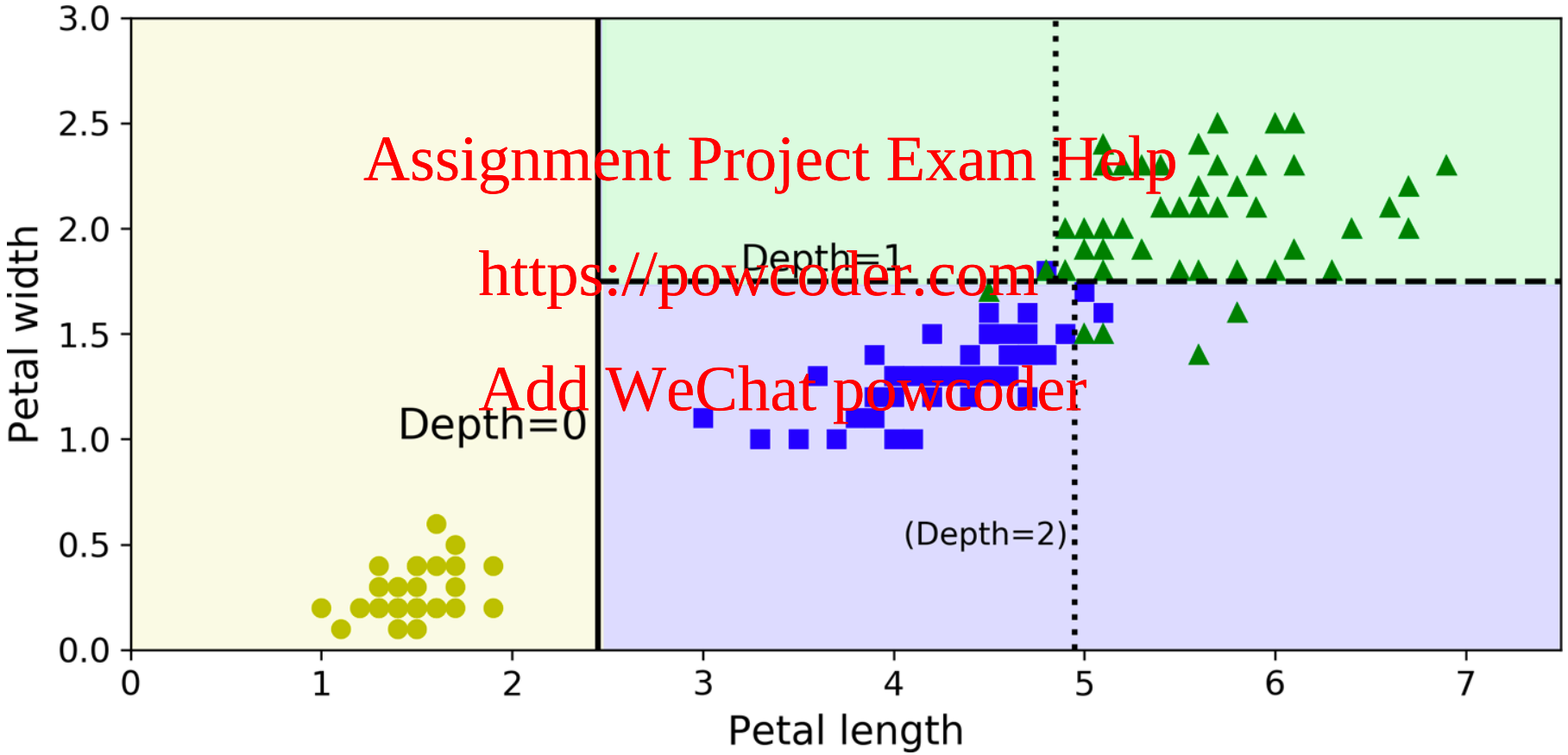
IRIS DECISION TREE

UCL
SCHOOL OF
MANAGEMENT



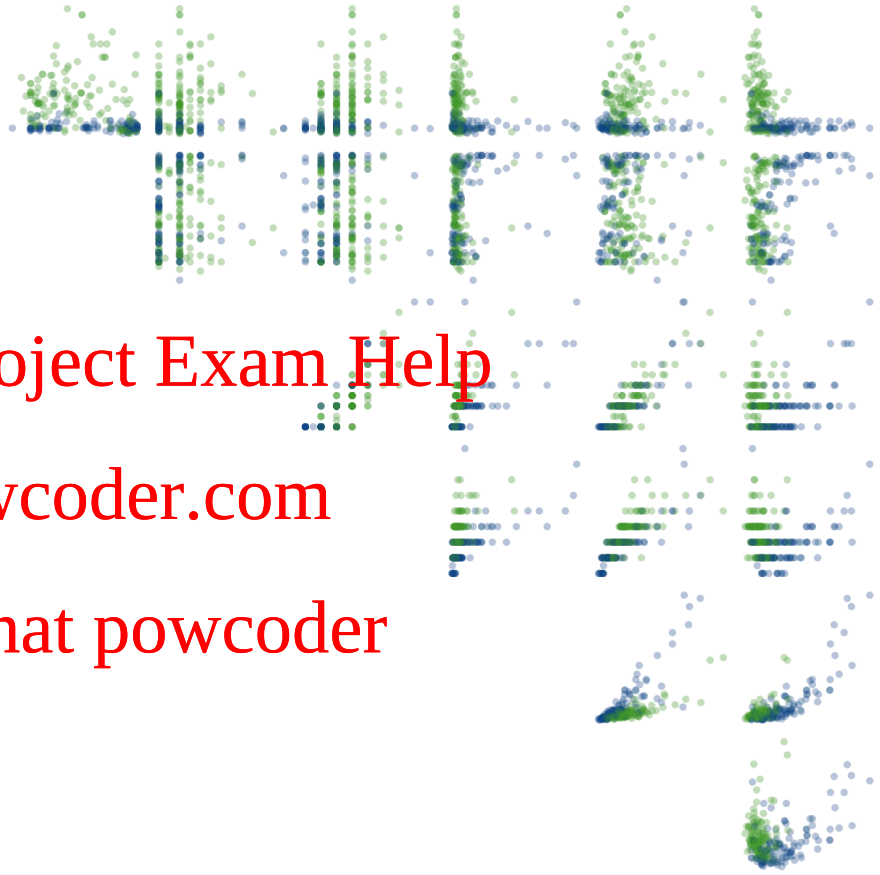
DECISION TREES

UCL SCHOOL OF MANAGEMENT



UCL
SCHOOL OF
MANAGEMENT

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/

Individual Assessment

LECTURE 1 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

A P MOORE