Assignment Project Exam Help

Introduction

https://powcoder.com

MAST90083 Computational Statistics and Data Mining

Dr Karim Seghouane

School of Mathematics & Statistics

The University of Melbourne

Add WeChat powcoder

Outline

Assignment Project Exam Help

§i. Admin

https://powcoder.com

§ii. Introduction & overview

Add WeChat powcoder

§iii. Basic concepts

Admin

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

- Lectures - Dr Karim Seghouane
  - Mon. 14:15-16:15
  - Baldwin Spencer Theatre or via zoom
- Practical lab - Jiadong Mao
  - 8 Groups
  - f2f: G2/Tues. 9:00-10:00, G3/Tues. 14:15-15:15 in PAR-Peter Hall-G70 (Wilson Laboratory) G4/Fri. 10:00-11:00 in PAR-Peter Hall-G69 (Thompson Lab)
  - Online: G1/Wedn. 15:15-16:15, G5/Tues. 16:15-17:15, G6/Thur. 16:15-17:15, G7/Wedn. 16:15-17:15 and G8/Thur. 12:00-13:00
- Consultation time
  - Wedn. 08:00-09:00
  - Frid. 08:00-09:00

Admin - Assessment

Assignment Project Exam Help

- Problem solving assignments. 45%
  - Three assignments due early, mid and late semester
  - Each written assignment is worth 15%
- Exam. 55%

https://powcoder.com

Add WeChat powcoder

Admin - LMS

Assignment Project Exam Help

- All relevant teaching material will be posted on LMS (including the supplementary and the additional material)
- Due to the time limits during the lectures, you will need to go over some mathematical details & deepen your knowledge outside of the lectures time.
- Your assignment must be submitted via LMS.
- Discussion board?

https://powcoder.com

Add WeChat powcoder

Admin - References

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

- **Elements of Statistical Learning** by Hastie Trevor, Tibshirani Robert & Friedman Jerome (2009).
- **An Introduction to Statistical Learning** by James Gareth Witten Daniela, Hastie Trevor & Tibshirani Robert (2013).
- **Introducing Monte Carlo Methods** with R by R.P. Christian & G. Casella (2010).
- **Computational Statistics** by G.H. Givens & J.A. Hoeting (2005).
- Academic **articles**, links to **blogs** & **videos** on LMS.

Admin - Communication

Assignment Project Exam Help

- Note that emails regarding the course material, laboratories and assignments will be addressed during office hours and if time permits.
- It is expected that questions regarding these matters will be asked during consultation hours or during the laboratories.

https://powcoder.com

- There will be no consultation hours during non-teaching periods.

Add WeChat powcoder

- I will be out of office during SWOT Vac week and the first two weeks of the examination period.
- I will provide an extra consultation time on week 12.
- Students should plan ahead with any queries regarding assignments or course material.

Lecture schedule (provisional)

Assignment Project Exam Help

- **Data mining** (7 weeks):
  - linear model selection and regularisation; kernel and local regression; basis expansion and spline regression; general additive models (GAM); classification and regression trees; bagging, random forests and boosting; support vector machines (SVM); component analysis and deep learning.

https://powcoder.com

- **Computational statistics** (5 weeks):
  - EM algorithm; Bayes computing; Monte Carlo methods; and bootstrap methods.

Add WeChat powcoder

Question to the class...

Assignment Project Exam Help

How would you explain in a few sentences to a general audience

https://powcoder.com

- What a statistician / data scientist does?
- What is the main purpose of that?

Add WeChat powcoder

Some answers

Assignment Project Exam Help

- What a statistician / data scientist does?
- **Creat a model (box) to understand the releathinseep between several variable**

https://powcoder.com

- **make sense of data, extract important patterns and trends, we can call this learning from data**
- What is the main purpose of that?

Add WeChat powcoder

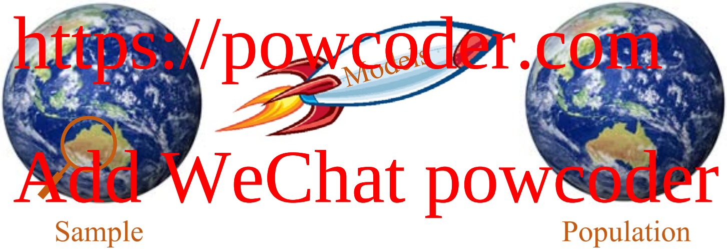
- **Predict/decide or describe/understand**

The power of (statistical) model

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder



Once upon a time ...

Assignment Project Exam Help

Having the data → holding the power !!!

https://powcoder.com

Add WeChat powcoder

Image source: http://freshlearners.blogspot.com.au/2015/07/most-recent-communications-technology.html

Today

The real power is in KNOWING WHAT TO DO with the data

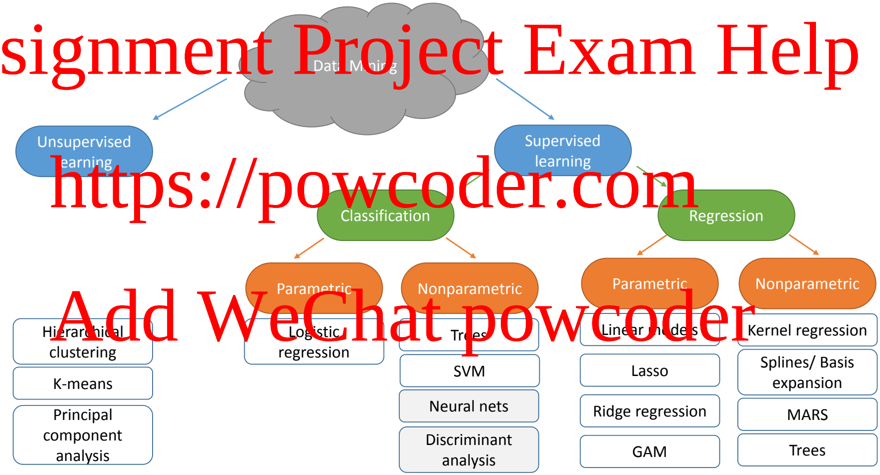
Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Image source: http://psmit.com/about.html

Taxonomy of covered methods



Data Mining

Unsupervised learning

Supervised learning

Classification

Regression

Parametric

Nonparametric

Parametric

Nonparametric

Hierarchical clustering

K-means

Principal component analysis

Logistic regression

Trees

SVM

Neural nets

Discriminant analysis

Linear models

Lasso

Ridge regression

GAM

Kernel regression

Splines/ Basis expansion

MARS

Trees

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

Regression vs. Clasification

Assignment Project Exam Help

https://powcoder.com

- **Regression** - The aim is to predict the numerical outcome of a subject based on its features.
- **Clasification** - The aim is to predict the class belonging of a subject based on its features.

Add WeChat powcoder

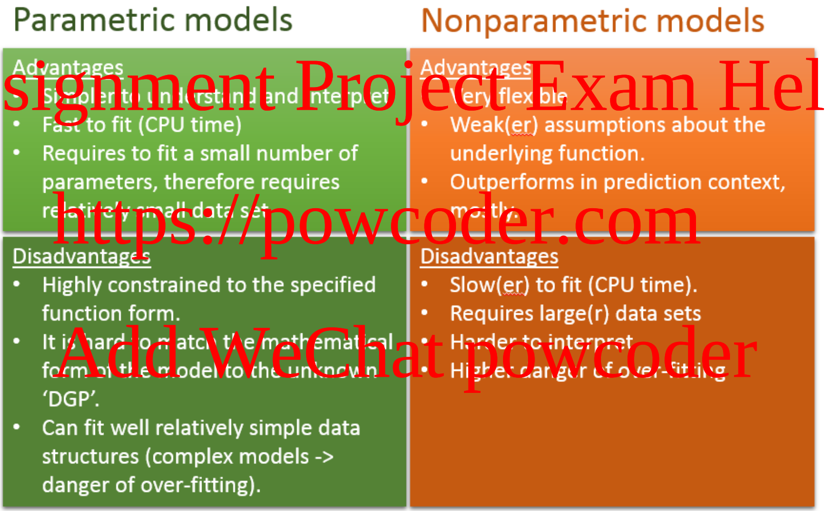
Parametric vs. Nonparametric

Assignment Project Exam Help

- **Parametric** approach - Makes an explicit assumption about the functional form. (Restriction on the shape)
- **Nonparametric** approach - does not assume any functional form for the underlying model structure.

https://powcoder.com

Add WeChat powcoder



Parametric models	Nonparametric models
Advantages • Simpler to understand and interpret • Fast to fit (CPU time) • Requires to fit a small number of parameters, therefore requires relatively small data set	Advantages • Very flexible • Weak(er) assumptions about the underlying function. • Outperforms in prediction context, mostly.
Disadvantages • Highly constrained to the specified function form. • It is hard to match the mathematical form of the model to the unknown 'DGP'. • Can fit well relatively simple data structures (complex models -> danger of over-fitting).	Disadvantages • Slow(er) to fit (CPU time). • Requires large(r) data sets • Harder to interpret • Higher danger of over-fitting

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

Supervised vs. Unsupervised

Assignment Project Exam Help

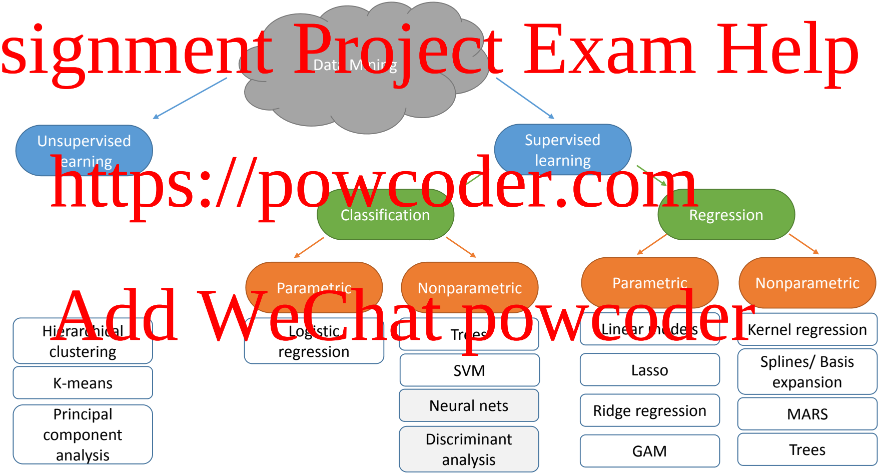
- **Supervised** learning - In case our data set contains the response (outcome) measurements, the fitted model relates the different features to the response.

https://powcoder.com

- **Unsupervised** learning - In case our data set contains only information about the different features of the subjects, the fitted model aims to segment the subjects into groups or just learn about the relationship of the features.

Add WeChat powcoder

Taxonomy of covered methods



- Data Mining
  - Unsupervised learning
    - Hierarchical clustering
    - K-means
    - Principal component analysis
  - Supervised learning
    - Classification
      - Parametric
        - Logistic regression
      - Nonparametric
        - Trees
        - SVM
        - Neural nets
        - Discriminant analysis
    - Regression
      - Parametric
        - Linear models
        - Lasso
        - Ridge regression
        - GAM
      - Nonparametric
        - Kernel regression
        - Splines/ Basis expansion
        - MARS
        - Trees

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

How models are fitted

Assignment Project Exam Help

- By optimization

https://powcoder.com

- We minimize some criterion (e.g. squared error with some additional penalty)

Add WeChat powcoder

Trade-offs

Assignment Project Exam Help

https://powcoder.com

- Flexibility (predictability) - Interpretability trade-off.
- Bias - Variance trade-off.
  - variance - how much $\hat{f}$ changes with different data set.
  - bias - the error from the approximation.

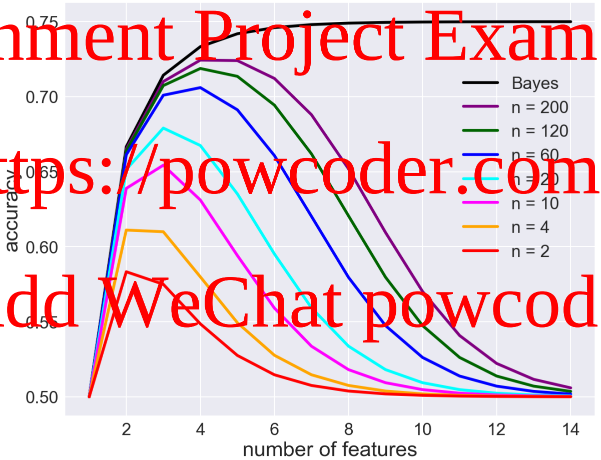
Add WeChat powcoder

Complexity

- Complexity trade-offs involving sample size, dimensionality and empirical performance. It is a characteristic feature of supervised learning methods.
- Notion of *curse of dimensionality*: for a fixed sample size, the expected classification error will improve by increasing the number of features, but eventually will decrease. This is a consequence of the large size of high-dimensional spaces, which require correspondingly large training sample sizes.
- *Scissors effect*: the expected error typically decreases as sample size increases, and more complex classification rules achieve smaller error for large sample sizes; however, simpler classification rules can perform better under small sample sizes, by virtue of needing less data.

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

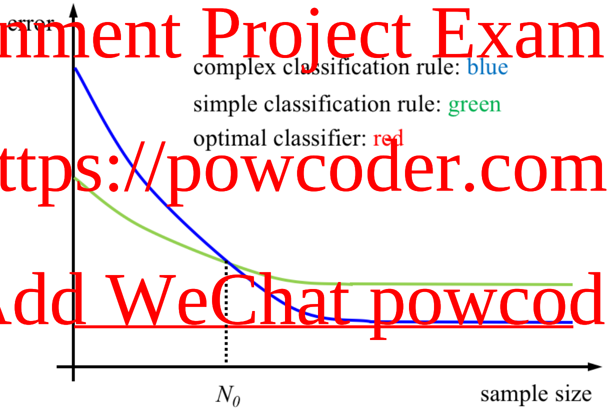
Complexity



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

Expected accuracy in a discrete classification problem for various training sample sizes as a function of the number of predictors.

Complexity



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

Expected error as a function of sample size for two classification rules. There is a problem-dependent critical sample size N0, under which one should use the simpler classification rule.