**Course Syllabus: MA234 Intro to Theoretical and Practical Data Science**

Semester 2024 Spring

**Instructor:** ZHANG, Zhen （张振）

Office: 理学院5014

Phone: 88018753

Email: zhangz@sustech.edu.cn

**Office hours:** Wednesday afternoon, 15:00-17:00; or send email to make an appointment for other time.

**Course webpage:** <http://cookdata.cn>(数据酷客)

**Lecture:** 3 credits, 3 hours per week,

location: 三教204 (双周周三3-4节)，三教204(每周五5-6节)，三教508机房（实验课，每周五9-10节）

**Prerequisite:** Calculus I&II, MA101b&MA102b, (or Mathematical Analysis I&II, MA101a&MA102a); Linear Algebra I, MA103b; Probability Theory, MA215 (or Probability Theory and Mathematical Statistics).

**ITA:** Yang Li (李杨) will be in charge of experimental lectures

**TA:** Lingxiao Che (车凌霄) will be in charge of grading homework and quizzes.

**Requirements:** Approximately 6 homework assignments (including online programming assignments and written problems) and quizzes. In the case of online live lectures, please hand in your completed homework online in the BB system. There is one programming project and one closed-book final exam.

**Your grade will be based on assignments (30%), quizzes (15%), programming project (20%), and the final exam (35%).**

**Main Contents:** This course is intended for undergraduate students who are interested in pursuing industrial work and research in big data science. It provides a concise and self-contained introduction to mathematical aspect of big data science, including theoretical analysis, algorithms and programming with python. Major topics include *introduction to python programming,* *data preprocessing, classification, regression, clustering, model selection, dimensionality reduction, and hot topics such as neural network and deep learning, reinforcement learning and recommender systems if time permits.*

**References:**

1. (textbook) 数据科学导引，欧高炎等著，高等教育出版社，2017.
2. (major reference) 机器学习，周志华 著，清华大学出版社，2016.
3. (Both theory and programming, introductory level) An Introduction to Statistical Learning with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2013.
4. (Theory, middle level) Pattern Recognition and Machine Learning, by Christopher M. Bishop, Springer, 2006.
5. (Theory, middle level) The Elements of Statistical Machine Learning: Data mining, Inference and Prediction, 2nd Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Springer, 2009.
6. (Theory, advanced level) Understanding Machine Learning, by Shai Shalev-Shwartz and Shai Ben-David, Cambridge University Press, 2018.
7. Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, MIT Press, 2016.
8. Foundations of Machine Learning, by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, The MIT Press, 2018.

**Course Schedule**

1. Introduction
2. Introductory example: diagonosis
3. Concepts in data science and machine learning
4. Mathematical representation
5. Key problems
6. Mathematical preliminary
7. Linear algebra: norms, eigen-decomposition, power method for extremal eigenvalue
8. Numerical methods: root finding, interpolation, optimization (gradient descent and subgradient)
9. Probability: discrete/continuous random variables, probability mass/density function, conditional probability, expectation, variance and covariance, Bernoulli and Gaussian distribution, Dirac and empirical distribution
10. Information theory: entropy, cross-entropy, and Kullback-Leibler (KL) divergence
11. Statistics: point estimation, unbiased estimator, maximum likelihood estimate (MLE), Beyesian rule, Maximum A Posterior (MAP)
12. Data preprocessing and python programming
13. Data types and statistics
14. Metrics and distances
15. Data scaling and discretization
16. Missing value treatment
17. Outlier detection
18. Python syntax
19. Programming examples in python
20. Classification
21. K-Nearest Neighbors (kNN)
22. Decision Tree
23. Naïve Bayes
24. Logistic regression
25. Support vector machine
26. Linear discriminant analysis
27. Model assessment
28. Regression
29. Linear regression by least square
30. Linear regression by MLE
31. Geometric interpretation and matrix algebra
32. Algorithm
33. Bias-variance decomposition
34. Overfitting and regularization: Ridge and Lasso
35. Model assessment
36. Ensemble Learning
37. Bagging and random forest
38. Boosting and AdaBoost
39. Gradient Boosting Decision Tree and XGBoost
40. Clustering
41. K-Means
42. Hierarchical clustering
43. DBSCAN
44. Expectation-Maximization (EM)
45. Spectral clustering
46. Model assessment
47. Dimensionality Reduction
48. Principal component analysis
49. Linear discriminant analysis
50. Nonlinear methods and manifold learning (Optional)
51. Autoencoder (Optional)
52. Learning Theory (Optional)
53. Feature and Model Selection
54. PAC learning
55. VC-dimension
56. Neural Network and Deep Learning (Optional)
57. Feedforward neural network and functional composition
58. Backpropagation and gradient descent
59. Convolutional neural network (CNN)
60. Recurrent (recursive) neural network (RNN)
61. Residual neural network (ResNet)
62. Generative model
63. Reinforcement Learning (Optional)
64. Multi-arm bandits
65. Markov decision process
66. Dynamical programming
67. Monte Carlo and Temporal-Difference Learning
68. Recommendation System (Optional)
69. Content-based recommendation
70. Collaborative Filtering
71. Latent Factor Model

**Experiment Schedule**

1. Python introduction
2. Install demo: jupyter notebook, python
3. Python essential exercise
4. Python Object-Oriented Programming exercise
5. Class exercises
6. Data representation and visualization
7. Web spider: Basic concepts for crawlers and common Python modules, including Urllib, Requests modules.
8. Data parsing：Regular expressions, Xpath, Css and BeautifulSoup parsing, Html parsing and more.
9. Data representation and visualization: Pandas, numpy, Matplotlib, Seaborn, Wordcloud and more
10. Analysis case: (1)[Pandas practical cases](http://cookdata.cn/note/view_static_note/c3ab079639eba7c664a2d5cf9e8183fc/), (2)[COVID-19Epidemic data visualization](http://cookdata.cn/note/view_static_note/84230091c4104946ab3d22dd675df58d/)
11. Data preprocessing
12. Data types, Data conversion
13. Data discretization and standardization
14. Missing value treatment
15. Outlier detection
16. Analysis case: (1)Passengers on Titanic, (2)Hypertension data analysis
17. Classification
18. K-Nearest Neighbors (kNN)
19. Decision Tree
20. Naïve Bayes
21. Logistic regression
22. Support vector machine
23. Linear discriminant analysis
24. Common Python modules: sklearn
25. Analysis case: (1)optical character recognition using SVM, (2)eighbor neighbor algorithm to build an automatic breast cancer diagnosis model, (3)use the decision tree to establish a personal credit risk assessment model. (4) mobile phone spam message filtering and gender prediction of Chinese names based on Naive Bayes algorithm.
26. Regression
27. Linear regression by least square, MLE
28. Overfitting and regularization: Ridge and Lasso
29. Analysis case: (1)the model for predicting medical expenses;(2) use logistic regression to classify iris varieties
30. Ensemble Learning
31. Bagging and random forest
32. Boosting and AdaBoost
33. Gradient Boosting Decision Tree and XGBoost
34. Analysis case: red wine quality classification
35. Clustering
36. K-Means
37. Hierarchical clustering
38. DBSCAN
39. Expectation-Maximization (EM)
40. Spectral clustering
41. Analysis case: classifies teenagers' information and interests
42. Dimensionality Reduction
43. Principal component analysis
44. Linear discriminant analysis
45. Nonlinear methods and manifold learning (Optional)
46. Autoencoder (Optional)
47. Analysis case: feature analysis of automobile on 11 indexes
48. Learning Theory (Optional)
49. Feature and Model Selection
50. PAC learning
51. VC-dimension
52. Analysis case: Application of PCA in face recognition task
53. Course project analysis and discussion.
54. Neural Network and Deep Learning (Optional)
55. Feedforward neural network and functional composition
56. Backpropagation and gradient descent
57. Convolutional neural network (CNN)
58. Recurrent (recursive) neural network (RNN)
59. Residual neural network (ResNet)
60. Generative model