STAT4001 Data Mining and Statistical Learning
Essay
Due: 5pm, December 16

Guidelines

- You need to summarize chapters 6, 7 and 9 in the essay. The essay must be a single pdf file. You may include formulas whenever you feel necessary.
- The essay must be submitted through VeriGuide. Students should pay attention to the academic honesty and plagiarism policy of the University. Marks will be deducted for any hints of plagiarism reported by VeriGuide. Cases with sufficient evidence of plagiarism will be forwarded to the Disciplinary Committee of the Science Faculty.
- We will not answer the questions below.

Chapter 6: Nonlinear methods (2 pages, Times New Roman, 12-point font size)
Here are some suggestions on what to write. You do not need to answer all these questions. You may write other contents that you feel important and worth mentioning.

- Summarize polynomial regression, regression with step functions, regression splines, local regression, generalized additive model
- What are the drawbacks of polynomial regression?
- What are the drawbacks of regression with step functions?
- What are the differences between cubic spline and natural cubic spline? Compared with cubic spline, what is the benefit of natural cubic spline?
- What methods do we use to find the regression coefficients in polynomial regression, regression with step function, linear spline, and cubic spline?
- Are there any tuning parameters in the methods? How to tune these parameters?

Chapter 7: Tree-based methods (2.5 pages, Times New Roman, 12-point font size)
Here are some suggestions on what to write. You do not need to answer all these questions. You may write other contents that you feel important and worth mentioning.

- Summarize regression tree, classification tree, bagging, random forest, boosting for regression tree, AdaBoost
- What is the key idea behind bagging?
- Why does random forest significantly improves over bagging?
- What are the differences between random forest and boosting?
- What are the key components in boosting for regression tree and AdaBoost?
- Are there any tuning parameters in the methods that have been introduced? How to tune these parameters?

Chapter 9: Unsupervised learning (2 pages, Times New Roman, 12-point font size)
Here are some suggestions on what to write. You do not need to answer all these questions. You may write other contents that you feel important and worth mentioning.

- Summarize K-means clustering, K-medoids clustering, and hierarchical clustering
- How to solve the objective function in K-means clustering?
- What are the differences between K-means clustering and K-medoids clustering?
- What are the pros and cons of K-means clustering and K-medoids clustering?
- How to choose the number of clusters in K-means clustering?