MSA220/MVE440 - Final Spring 2018

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

Due June 8

Question 1 (50p)

You have conducted 5 Minis. You now have a chance to revisit the Minis, perhaps update and improve on your analysis/investigation since you have acquired more knowledge about the topics as the semester progressed.
If you are unhappy with one of the Mini topics you can switch topics at this point.
This is an individual assessment. You may have worked as a team for the Mini presentation. For this examination you have to work independently but can of course base your write-up on work conducted as a team.

Write a 1-2 page (excluding figures) report for each of the Minis. Make sure to clearly define your approach (e.g. simulation setup, data set used, methods and models). Focus on model/method critique, insight into performance (comparing methods, for different data, for different simulation settings). What's the take-home message for each of the Minis?

Question 2 (25p)

Consider the TCGA gene expression data we have looked at in class (data set posted on the class homepage).
a) Investigate the impact of mislabeled observations on classification. You can construct a data set with mislabeled data by randomly selecting observations and allocating them to different classes (either completely randomly or with structure where class A is always or mostly mislabeled as B). For a minimum of 3 classification methods, investigate the classification performance as a function of proportion of mislabeled observations for the two cases of mislabeling (random or structured).
b) For a moderate degree of mislabeling, suggest and implement a method for detecting which observations are mislabeled.

Question 3 (25p)

Conduct simulations studies to investigate questions a and b. Requirements for the simulation in the itemized list below.

a) Consider the case where data comprises K clusters and L classes, $L > K$. Investigate different methods for clustering and classification in this setting and discuss your results and findings.
b) Consider the case where data comprises K clusters and L classes, $L < K$. Investigate different methods for clustering and classification in this setting and discuss your results and findings.

- you can start from scratch or use real data as a basis for the simulation
- at least 20 features. you can choose to have some of them be unrelated to clusters/classes
- you can opt to have clusters and classes be related or not

Since you can choose your simulation settings, clearly state how you made your choices and why.