Math336 Project Description

Florian Pein

22/23

In this module we mostly deal with supervised learning problems. A typical example of an unsupervised learning problem is clustering, which is one of the most widely used techniques across all disciplines from social sciences to computational biology and finance only to name a few. In this project you will explore clustering as outlined below.

k-means Clustering

Familiarise yourself with clustering, and how it is different from classification. There are many algorithms for clustering. Your task in this project is to explore the particular algorithm called k-means clustering. Describe the model, give a pseudo-code for the procedure, explaining how it works.

Clustering as an Image Compression Tool

One application of clustering is image compression. In a 24-bit colour image each pixel is represented as three 8-bit integers between 0 to 255 indicating its RGB intensity values. Therefore, an image contains thousands of colours. One way to reduce the size of an image is to reduce its colour count. Propose an approach based on k-means clustering to reduce the colour count of an image to $k = 16$. Implement an R script that takes as input an image, and reduces its colour count to 16 colours using k-means clustering. You may start with the provided R script. You are invited to write the clustering algorithm yourself, but it is not required. Run your script with the provided image. You may want to use your own images as well. Comment on the effect of the choice of k on the compression quality, i.e. how the compression quality changes as k is increased or decreased.

Submission

Each project should consist of a written report compiled collaboratively with your fellow group members. Each group will make only one submission by the specified deadline; this will be done by one of the group members via Moodle. The submission should contain,

1. a contribution sheet (ideally as an appendix of the report): This should list all group members and their contribution. At the bare minimum it should include who has written and who has proofread which parts of the code and of the report. Each of you should be contributing to the writing of the report. Each of you should proofread the work of at least one other person. You are also responsible for the quality of the text you have proofread. The only exception is plagiarism. Only the person who has written the part of the code / report will be responsible for potential plagiarism (see below).
2. the main report as a pdf file. Your goal as a group is to carry out an investigation of the k-means clustering problem. You should start by providing some background information and a brief discussion of benefits and shortcomings of k-means clustering. Remember to reference this appropriately. This also includes all images you have used. The expectation is to have an introduction, sections on the clustering problem with focus on the k-means clustering algorithm, and on the application of k-means to image compression, and a brief conclusion. As a rough guide, aim for about half a page on the introduction and conclusion, and one page on the other three sections. The introduction would give a

brief background to clustering as an unsupervised learning problem, and outline the contents of the report, the conclusion would discuss limitations of the k-means clustering.

3. a zip that should contain any R code used. The R code must produce all the results mentioned in the pdf. If you have time consuming steps, please provide intermediate results as well, e.g. simulation results saved as a csv or RDS file. If you use any images that are not easily available online, then include them as well.

Reports which contain computer code in the main body will be penalised. The main body of the report without references, i.e. the report excluding the references and the contribution sheet, should not be more than 5 pages, written in size 12 font with sensible margins and spacing. This includes all graphs and tables but excludes the appendix.

Reports with a main body over 5 pages will be penalised!

A shorter report, with material effectively communicated, is acceptable. Writing concisely is an important skill. So it is part of the project and its assessment.

Assessement

Marks will be assigned for the presentation, organisation, and content of your report. Furthermore, your R code will be assessed. The code has to be correct, produce the results given in the report and follow minimal coding standards such as basic level of annotation, identification of lines, etc. You will be given one mark as a group, but I will make individual adjustments if the contribution of members differ significantly. Make sure that your contributions are listed in the contribution sheet. Please note that this is an openended project, creativity and thinking outside the box will be highly valued! Marks above 80% will require some extra work in addition to the listed tasks.

Timeline

1. Week 6: Assignment of the groups
2. Week 7: First meeting: provisional breakdown of tasks; who is responsible for which tasks. Write it down in the contribution sheet.
3. Week 8-10: Working on the project and writing. I recommend that you start writing mid week 9 at the latest. Ensure that there is enough time to proofread each others work.

Submission deadline is Friday 16 December at 23:59pm.

Working as a group

You have to work together as a group. This includes helping each other. It is also required that you check each others work (proofreading). Effective group work will be rewarded.

If major problems occur such as non contributing group members, missing of meetings etc., let me know **as early as possible**. Any issues, e.g. someone exceeds his writing time and others cannot proofread, have to be communicated to me **before** submission deadline or listed on the contribution sheet if the group agrees on it. If needed, a viva may be held in Lent term to sort out individual contributions.

Plagiarism

Plagiarism is presenting someone else's work or ideas as your own, with or without their consent, by incorporating it into your work without full acknowledgement. All published and unpublished material, whether in manuscript, printed or electronic form, is covered under this definition. Plagiarism may be intentional or reckless, or unintentional. As usually, it is strictly forbidden and will be penalised. University procedures will be followed. Note that the university runs automatic tools to help us detecting plagiarism. So write the report yourself and cite all references that you have used. This includes course material, but not the provided R code for the project. Simply copying resources with giving references is not plagiarism, but of

course not enough for the project. You must do “significant work” beyond the referenced resources. If you have any questions, let me know.