Machine Learning 2020-2021

Home Assignment 6

Yevgeny Seldin **Christian Igel**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **12 January 2021, 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your source code in the PDF file.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.
- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in speed grader. Zipped pdf submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted, please use the provided latex template to write your report.

Unsupervised Learning – Clustering

The following task is partially taken from a previous exam. The data are based on a research project funded by Miljøstyrelsen involving researchers from DIKU.

Selected results from the project are described by Rasmussen et al. (2016) and Olsen et al. (2017). While the problem setting is inspired by Olsen et al. (2017), the data were generated differently, in particular without preprocessing to compensate for color-balancing and a number of illumination effects.

Introduction to the problem. We start by giving some background information explaining the underlying real-world problem. However, understanding the data generating process is not necessary for solving the assignment.

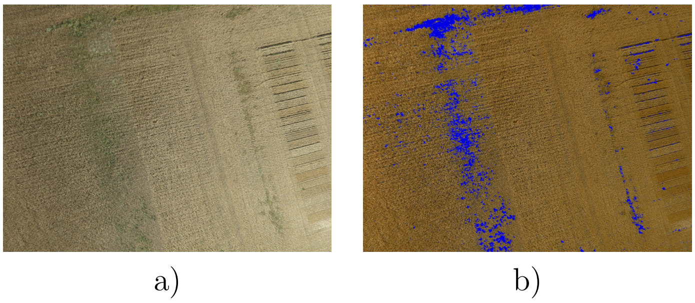
Pesticide regulations and a relatively new EU directive on integrated pest management create strong incentives to limit herbicide applications. In Denmark, several pesticide action plans have been launched since the late 1980s with the aim to reduce herbicide use. One way to reduce the herbicide use is to apply site-specific weed management, which is an option when weeds are located in patches, rather than spread uniformly over the field. Site-specific weed management can effectively reduce herbicide use, since herbicides are only applied to parts of the field. This requires reliable remote sensing and sprayers with individually controllable boom sections or a series of controllable nozzles that enable spatially variable applications of herbicides. Preliminary analysis (Rasmussen et al., 2016) indicates that the amount of herbicide use for pre-harvest thistle (Cirsium arvense) control with glyphosate can be reduced by at least 60 % and that a reduction of 80 % is within reach. See Figure 1 for an example classification. The problem is to generate reliable and cost-effective maps of the weed patches. One approach is to use user-friendly drones equipped with RGB cameras as the basis for image analysis and mapping.

The use of drones as acquisition platform has the advantage of being cheap, hence allowing the farmers to invest in the technology. Also, images of sufficiently high resolution may be obtained from an altitude allowing a complete coverage of a normal sized Danish field in one flight.

Data. Your data is taken from a number of images of wheat fields taken by a drone carrying a 3K by 4K Canon Powershot camera. The flying height was 30 meters. A number of image patches, all showing a field area of 3×3 meters were extracted. Approximately half of the patches showed crop, the remaining thistles. For each patch only the central 1×1 meter sub-patch is used for performance measurement. The full patch was presented to an expert from agriculture and classified as showing either weed (class 0) or only crop (class 1).

In Figure 2 two patches classified as crop and two patches classified as weed are shown. Two of the patches are easy to classify (expert or not), while the remaining two less clearly belong to either of the classes.

For each of the cental sub-patches (here of size 100×100 pixels), 13 rotation and



a) b)

Figure 1: Example from another approach. a) An original image. b) Initial pixel based detection.

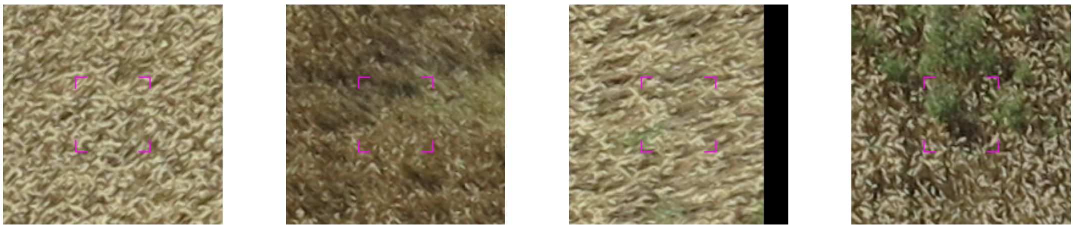


Figure 2: The two images on the left are classified as crop. The two images on the right are classified as weed. The classification of the middle two patches is debatable. The central area used for performance evaluation is indicated by the small magenta markers.

translation invariant features were extracted. In more detail, the RGB-values were transformed to HSV (hue, saturation, value; a color representation model better reflecting human perception than standard RGB) and the hue values were extracted. The 13 features were obtained from a 13-bin histogram of the relevant color interval.

The data you need for this assignment are stored in `MLWeedCropTrain.csv`. The last column corresponds to the class label. Remember to remove the class label from the data when computing the k-means prototypes. **Do not use normalization as preprocessing in this assignment.**

Task 1: Principal component analysis and visualisation. Perform a principal component (PCA) analysis of the data in `MLWeedCropTrain.csv`. Plot the eigenspectrum (eigenvalue vs. rank of corresponding eigenvector after sorting by eigenvalue; be careful not to mix up singular values and eigenvalues). You find an example of a plotted eigenspectrum on the slide “Explained variance” of the lecture on PCA. How many components are necessary to “explain 90 %

of the variance"? Visualize the data by a scatter plot of the first two principal components. Use different colors or symbols for weed and crop. Write down the equation mapping the 13-dimensional points to two dimensions with a very concise explanation.

Deliverables: Description of software used; plot of the eigenspectrum; number of components necessary to explain 90 % of variance; scatter plot for first two principal components with different colors indicating the different classes; equation mapping the 13-dimensional points to two dimensions with a very concise explanation

Task 2: Clustering. Perform 2-means clustering of `MLWeedCropTrain.csv`. For the submission, initialize the cluster centers with the first two data points in `MLWeedCropTrain.csv` (that is not a recommended intialization technique, but makes it easier to corrrect the assignment). Plot the cluster centers using the transformation described above. That is, add the cluster centers to the plot from the previous question. Briefly discuss the results: Did you get meaningful clusters?

Deliverables: Description of software used; one plot with cluster centers; short discussion of results

References

S. Olsen, J. Nielsen, and J. Rasmussen. Thistle detection. In P. Sharma and F. Bianchi, editors, *Scandinavian Conference on Image Analysis (SCIA 2017)*, volume II, pages 413–425. Springer, 2017.

J. Rasmussen, J. Nielsen, S. I. Olsen, K. Steenstrup Petersen, J. E. Jensen, and J. Streibig. Droner til monitorering af flerårigt ukrudt i korn. Bekæmpelsesmiddelforskning 162, Miljøstyrelsen, Miljø- og Fødevareministeriet, 2016.