Project 3, Exercises

Veronika Cheplygina

IT University of Copenhagen

This is all the exercises for mini-project 3. Doing these exercises will help you understand the material, and give you ideas for your project, but you do not have to do every single exercise to finish the project. Use the exercise time to start (with help from TAs) and become familiar with the data and methods.

For several questions, there will be no single correct answer, as your interpretation will depend on the images/functions/etc you used. Therefore, it is important you motivate your choices in your project.

1 System setup

Set up your Github repository "fyp2021p03gXX" where XX stands for your group number, similar to the other repositories. You can use the project repository (`https://github.com/vcheplygina/fyp2021p3`) as a template to get started.

1.1 Explore an image

You can use a notebook or a Python script for this. Try the following:

- Use matplotlib.pyplot package to load an image and the corresponding segmentation. Display the images side by side (as subplots)
- Inspect the dimensions and values inside the images, what do you notice?
- How can you use the mask to measure the width or height of the skin lesion at different points in the image? Can you now find the maximum width and height? What about finding the maximum possible diameter of the lesion?
- Use the segmentation to mask out / erase the outside of the skin lesion. Can you do this without a for loop?
- Try to extract the color of a single pixel inside the lesion, and store this color. Can you now fill the entire lesion, just with this color?

2 Features

2.1 Area and perimeter

Review the provided functions for measuring the area and perimeter, then try the following:

- What do you expect from these values for a perfect circle?

- Verify your hypothesis by drawing a circle (you can do this in Python or just draw it yourself), and measuring the features on it. Do your measurements confirm your hypothesis? Why/why not? Hint: try to think what happens for a tiny circle, with a diameter of only a few pixels.
- Create a scatter plot for the area and perimeter of the images you are provided with. Do you see some patterns already? Are there some outliers? If yes, can you explain why some images have such values?
- Try to modify an image from your dataset in some way, like adding noise to the image or to the mask. How does this affect the features?
- Try to modify the code of measuring the area and perimeter (for example, morphology.disk with parameter 1, vary the parameter or choose a different shape, or look at morphology functions which is not erosion). Investigate how this affects the masks (you might want to visualize intermediate steps here), and how this then affects the area/perimeter measurements.

2.2 Other features

Choose a property you want to measure (asymmetry, border, colour) and search the literature to find out how it can be measured, and/or look at an existing implementation online.

- Are there different (conceptual) ways to measure the same feature? Are there different implementations (for instance, the steps are similar, but different functions or parameters are used) of the same concept?
- Adapt an implementation to work with the skin lesion images, and examine the feature values this outputs for your data. Do you see differences for healthy and cancer lesions? What if you vary the parameters of the method?
- Create some artificial images (like a perfect circle, an image of a single colour, etc) and measure the feature for those. Do you get values that you would expect?
- Create scatter plots for pairs of features you have so far. You might want to normalize the features first. Can you see (some) separation between the healthy and cancer lesions?

3 Classifiers

Once you have a selection of features (either measuring different things, or measuring the same thing in different ways), you can experiment with classifiers.

3.1 Data size

- Try splitting your dataset into two parts, and look at the distributions of the features (with histograms, scatter plots, etc) of each part. What do you think of these differences?

- Split your data, and fit several classifiers on the training set. You can either change the features used, the classifier, and/or the parameters of the classifier. Now evaluate these classifiers on the training set - which combination of features/classifier/parameters is the best?
- Evaluate the same (trained) classifiers on the test set. Is your answer still the same?
- Do your answers change, if you had less data to start with (before splitting the data)?

3.2 Overfitting

There are some parallels between p-value hacking and overfitting. This is a thought exercise to understand why.

- Go to `http://tylervigen.com/spurious-correlations` and explore some correlations. Find some strange/funny ones that you don't think can be related.
- If you think of the entire set of data on the website, what do the sample size and dimensionality correspond to?
- Create a dataset with random numbers (there is a `random` package), such that the dataset has low sample size, but high dimensionality. Choose on the features as the label you want to predict. Your "classifier" is simply a loop through the features, that checks the p-value of each feature, and uses that feature to predict the label.
- Experiment with different sample size/dimensionality, are you able to get good predictions? Did you expect this? What would happen if your classifier was more flexible (for example, you can use multiple features, transform them, etc)?

References