Assignment 1

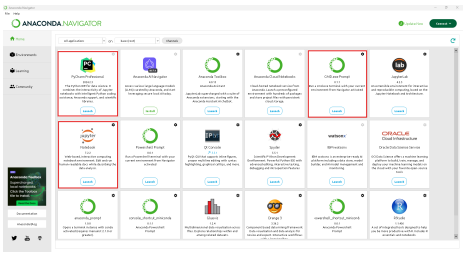
Task 1: Summarize Chapters 1, 2 and 3

Chapter 1 is named "The Science of Information. This chapter talks mainly about data science and highlights 2 different paradigms. One is the big data paradigm which looks used for problems with data volume, velocity and variety. The second paradigm that is looked at thoroughly is the machine learning paradigm which is used for understanding, modeling and making predictions from data. Chapter 1.1 specifically starts by defining data science as a "as the management and analysis of data sets, the extraction of useful information, and the understanding of the systems that produce the data". I think this part was to start the visualization about how data science is applied in real world scenarios. Chapter 1.2 talks about data characteristics. There were 3 main words that I picked up on that I mentioned earlier in this section. These 3 words are Volume, Variety and Velocity. Volume refers to the amount of data that is generated and collected. Variety refers to the diversity of formats and also the diversity of the data itself. Finally, velocity is basically the speed at which the data is either generated or processed. Chapter 1.3 Talks about the 2 main paradigms that I mentioned earlier. The Big Data Paradigm and the Machine Learning Paradigm. The Big Data Paradigm is said to deal with large and complex datasets. This chapter talks about the tech that is used to manage typical big data and it specifically talks about distributed file systems that are used to store and analyze data efficiently. The machine learning paradigm on the other hand gives the algorithms and models that are necessary for predictive analytics which enables data-driven decision-making without explicit programming. Overall this chapter talks about the importance of collaboration and firmly implies that professionals in mathematics, computer science, stats and other related fields are needed for effective data science. Some other topics the chapter also mentions various practical applications of data science which include financial sectors, academic institutions, in-formation technology divisions, health care companies, and government organizations

Chapter 2 is named Big Data Essentials. It goes into detail about the essentials of big data analytics, and it establishes a solid foundation for understanding the problems and techniques that are associated with handling large datasets. This chapter also mentions three familiar words from the previous chapter. Volume, velocity and variety. The chapter overall talks a lot about the importance of data classification and how important it is for scalable systems to be able to process big data efficiently. It brings up a couple technologies such as Apache, Hadoop and Spark which are all big data applications that are commonly used in the world. The chapter also talks about differentiating between descriptive, predictive and prescriptive analytics. Descriptive analytics mainly focuses on analyzing past data to understand TRENDS. Predictive analytics mainly focuses on historical data to PREDICT future outcomes and finally perscriptive analytics is used to RECOMMEND certain actions based on the data that was analyzed. The chapter later talks about how machine learning is extremely important to big data analytics by highlighting the importance of the role of using algorithms to identify patterns and make a prediction. It talks about many different common machine learning techniques which include supervised learning and unsupervised learning. For a quick explaination, supervised learning uses labeled training data and unsupervised does not. So this basically means that when a model is training unsupervised, this means that it does not understand the meaning of the data it is looking at. Supervised is when you basically tell the model, "you are looking at different types of hardwood". The purpose of chapter 2 was basically to prepare us for deeper discussions about practical applications in future chapters.

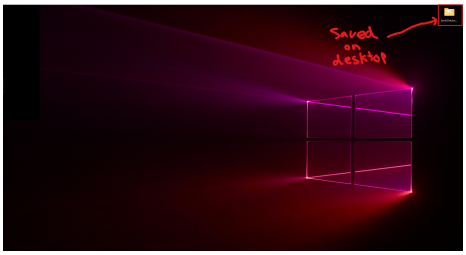
Chapter 3 talks about the basics of Machine learning which is detremental to datascience. This is mentioned several times in the book so far because it basically makes computers learn and can be mostly automated. The chapter starts off my talking about what machine learning is and how it is different from traditional programming. Basically instead of having to write specific instructions and hard code for stuff, machine learning algorithms will essentially learn by themselves with minimal supervision. The chapter mentions 3 main types of machine learning. The first is unsupervised. Supervised learning is when you train a model using labeled data—basically, you give it a set of input and the expected output, so it learns to make predictions on new, unseen data. On the flip side, unsupervised learning deals with unlabeled data and is all about finding patterns on its own. This type of learning is often used for tasks like clustering similar items together. Lastly reinforcment learning is when an agent learns by interacting with its environment. The model will either be rewarded or penalized based on its actions, kind of like training a dog or a baby. Other topics mentioned in this chapter also include, feature extraction, selection and also other key topics like cross-validation which is used to increase the variety. To conclude, the purpose of the chapter was to introduce the basics of machine learning for future chapters.

Task 2: Install and Setup Environment



Installed pycharm and some other stuff shown in the screenshot. All of this was installed through AnacondaNavigator which was downloaded through the link in the assignment pdf. I didn't end up using spyder even though I installed it because it was running terribly on my computer. Pycharm ended up working a lot better.

Task 3: Download and Store Data Set



Simply downloaded and extracted the zip file from the github link on the assignment pdf.

Task 4: Convert images to RGB

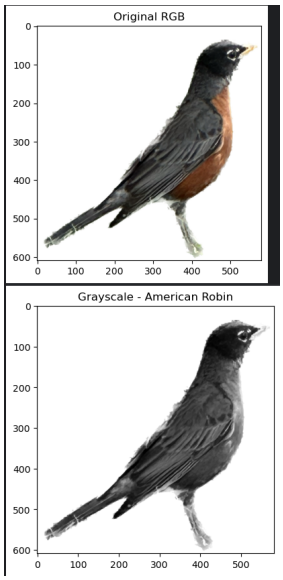


Figure 1: Robin

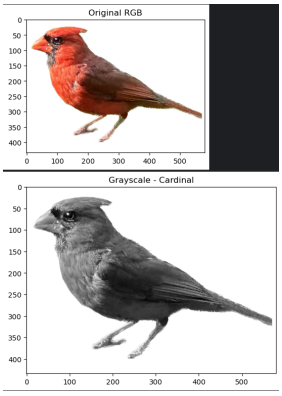


Figure 2: Cardinal

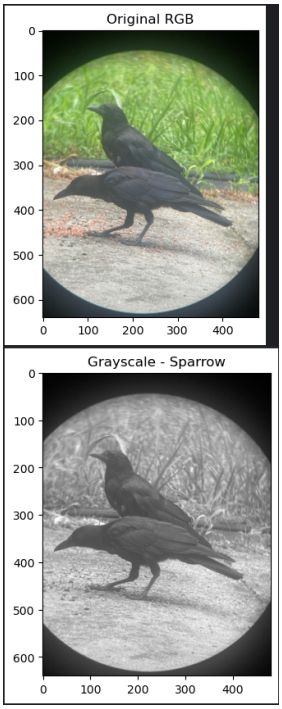


Figure 3: Sparrow

Wrote some code to convert the image to RGB because it was in BGR format originally for some reason. Displayed the RGB version and then the grayscale image for each bird. Used methods directly from the files section on canvas.

Task 5: Resize and display resized images

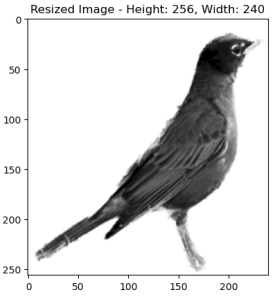


Figure 4: Robin

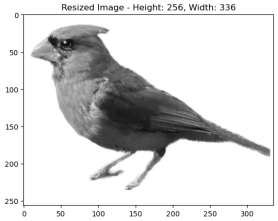


Figure 5: Cardinal

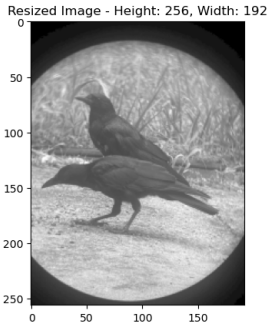


Figure 6: Sparrow

Wrote some code to resize the image in accomodation to the assignment. The instructions were to make the height 256 and then make the width divisible by 16 while maintaining aspect ratio. For this one, the assignment said to specifically make it a function so I wasn't sure if this was a mistype or not but I was still basically able use the code from the files section on canvas again.

Task 6: Block Feature Vectors

Wrote some code to generate block feature vectors in a specified location on my computer. The vectors were stored in CSV format so that I would be able to analyze it in a spreadsheet application of choice. I used google sheets. I noticed that a lot of the features were showing as 255 and I looked up why this was and I also experimented with my own pictures at some point. It seems that the 255 value might be the white space in the background of the bird image. It makes sense why the image of the sparrow had a lot less of these values. I put a png of a cartoon character to test and I concluded that this was the reason. CSV file that was generated is in my zip file labeled accordingly.

Task 7: Sliding Block Feature Vectors

Wrote some code to generate sliding block feature vectors in a specified location on my computer. The vectors were stored in CSV format so that I would be able to analyze it in a spreadsheet application of choice. I used google sheets. This part was very similar in code to the last part and I noticed that sliding block method created way more features. This is likely due to the fact that this method does not care about about overalapping with pixels that it has already scanned. CSV file that was generated is in my zip file labeled accordingly.

Task 8: Statistical Analysis

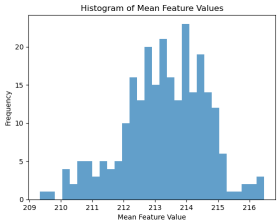


Figure 7: Robin

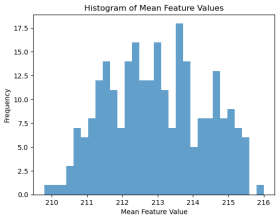


Figure 8: Cardinal

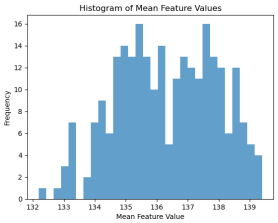


Figure 9: Sparrow

Robin Features (Image 1)

- **Distribution:** The distribution is slightly right-skewed, indicating a longer tail on the right side.
- **Comparison:** Compared to the other two images, the Robin features have a slightly higher overall mean value.

Cardinal Features (Image 2)

- **Distribution:** The distribution is relatively symmetric, with a slight left-skewness.
- **Comparison:** The Cardinal features have a mean value that is slightly lower than the Robin features but higher than the Sparrow features.

Sparrow Features (Image 3)

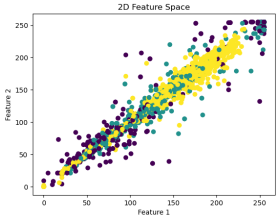
- **Distribution:** The distribution is slightly left-skewed, with a longer tail on the left side.
- **Comparison:** The Sparrow features have the lowest overall mean value among the three images.

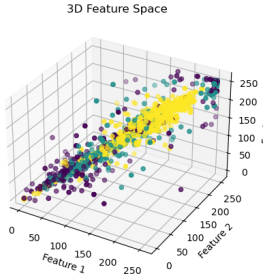
Task 9

Wrote some code to merge the block vectors with the sliding block vectors. Merged results were also randomly shuffled and then saved into a CSV file on my computer. CSV file that was generated is in my zip file labeled accordingly. Used this stack overflow discussion to figure out the randomization part in a much simpler way: https://stackoverflow.com/questions/71758460/effect-of-pandas-dataframe-sample-with-frac-set-to-1

Task 10

Wrote some code to generate a 2d and 3d plot. Couldn't find a method for the 3D plot anywhere in the files and I was struggling for a while till I found this: https://stackoverflow.com/questions/8722735/i-want-to-use-matplotlib-to-make-a-3d-plot-given-a-z-function. This stack overflow discussion shows a good method to create a 3D plot for the feature space.





Based on the 2D and 3D feature space plots, the data shows a strong positive correlation between the features in both plots. While it looks like there is a lot of overlap in the 2D plot, the 3D plot shows us that the are not actually overlapping as much. This shows us how important the 3rd feature is in this context because we would never know that some of the values were in fact distinct.

Task 11

Did not do this part.

Task 12

Changing the block size affects how we describe images. Bigger blocks mean we get more detailed descriptions, like looking at a picture closer to see more details. This means we have more features, which can make it harder for a computer to understand the picture. However, it can also help the computer learn better because it has more information. It's like trying to describe a person based on just their face versus their whole body. More details can make it easier to tell them apart, but it can also make it harder to decide who they are. I think this is what is mentioned a couple times in the book. Too much noise is generated when there are too many features and it will negatively affect the model's performance.

Task 13

did this