**Module 6 – Machine Learning**

**PreLab 6.2: Introduction to Machine Learning**

Now that you have some experience with Python, we’re going to introduce you to some basic machine learning techniques. This lab will have you explore three fundamental machine learning algorithms by applying them to real-world datasets.

The following pre-lab exercises will provide context for the algorithms we’ll be working with. Their aim is to build a little bit of intuition before coming to lab. Each part will have you watch a video or two, then answer a question to verify your understanding.

**Part I: Introduction to Linear Regression**

Watch Cassie Kozyrkov’s lecture on [Simple Linear Regression](https://www.youtube.com/watch?v=j8VjRnaHRBM&list=PLRKtJ4IpxJpDxl0NTvNYQWKCYzHNuy2xG&index=11) (≈ 9 minutes long).

Next, watch her lecture on [Multiple Linear Regression](https://www.youtube.com/watch?v=VqhafjGTDI8&list=PLRKtJ4IpxJpDxl0NTvNYQWKCYzHNuy2xG&index=12) (≈ 5 minutes long).

Suppose that you are a meteorologist tasked with predicting the total rainfall in College Park on any given day. You have access to a (fictional) dataset of historical daily rainfall amounts and other weather-related features for the past five years; the dataset looks something like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Average Cloud Cover (%)** | **Average Temperature (oF)** | **Average Wind Speed (mph)** | **Average Humidity (%)** | **Average Barometric Pressure (inHg)** | **Total Rainfall (inches)** |
| January 1st, 2017 | 5% | 34 oF | 3 mph | 38% | 30.20 inHg | 0.00 inches |
| January 2nd, 2017 | 90% | 39 oF | 10 mph | 93% | 29.80 inHg | 0.39 inches |
| … | … | … | … | … | … | … |
| March 27th, 2022 | 40% | 58 oF | 2 mph | 91% | 30.01 inHg | 0.01 inches |
| March 28th, 2022 | 95% | 63 oF | 26 mph | 99% | 28.79 inHg | 1.23 inches |

If you were to construct a linear regression model to predict total rainfall given the features above, which feature(s) would you use? Why? Based on the feature(s) you chose, would your model be considered *simple* or *multiple* linear regression?

I would use could cover, average humidity, barometric pressure, and total rainfall. This, this would be considered a multiple linear regression.

**Part II: Introduction to Classification**

Watch Cassie Kozyrkov’s lecture on [*k*-Nearest Neighbors Classification](https://www.youtube.com/watch?v=HnCHdeyJNOM) (≈ 4 minutes long).

Suppose that we have a dataset consisting of fifteen 2D points separated into five unique classes, displayed on the plot below:

Chart, scatter chart

Description automatically generated

Suppose that a new datapoint arrives with Feature1 = 2 and Feature2 = 3 (displayed as the black circle at coordinates (2,3) in the figure above). For each of the following values of *k*, determine which class this new datapoint will be classified as using *k*-nearest neighbors (either by inspection, a.k.a. “eyeballing it”, or by using the [2D distance formula](https://www.calculatorsoup.com/calculators/geometry-plane/distance-two-points.php)):

|  |  |
| --- | --- |
| ***k*** | **Class** |
| 1 | 2 or 3 |
| 2 | 2 or 3 |
| 3 | 3 |
| 4 | 2 or 3 |

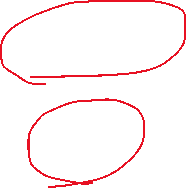
**Part III: Introduction to Clustering**

Watch Cassie Kozyrkov’s lecture on [*k*-Means Clustering](https://www.youtube.com/watch?v=lQ39ZRFfYbI) (≈ 8 minutes long).

Consider the seven datasets below consisting of 1500 datapoints each. Suppose that we wish to separate them into three distinct clusters using the *k*-means algorithm (i.e., *k* = 3). Hypothesize which points will belong to which of the three resulting clusters. You can do this by circling the points on each plot corresponding to each of your three hypothesized clusters. (We will compare your guesses to the actual clusters during the lab).

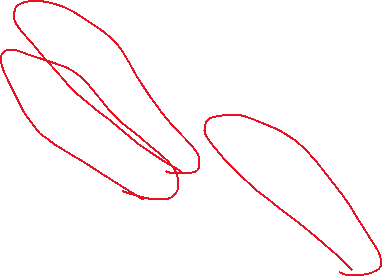
Chart, scatter chart

Description automatically generated



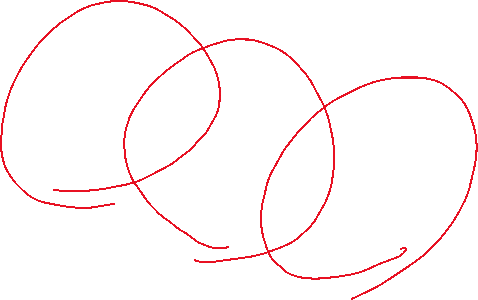
Chart, scatter chart

Description automatically generated



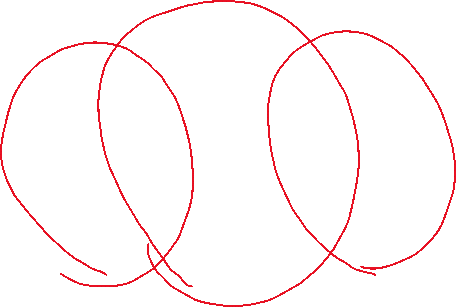
Icon

Description automatically generated



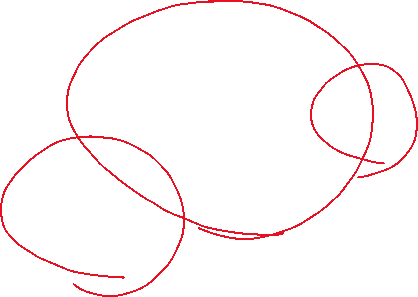
Icon

Description automatically generated



Chart, scatter chart

Description automatically generated



A picture containing text, fabric

Description automatically generated

