Section 1

**Getting Started with Machine Learning (From the outline)**

***Section Description (from the outline):*** This volume starts explaining how to set up our environment to build a machine learning application by leveraging a toolkit of open source libraries. First, we’ll discuss proper setup our Anaconda environment and from there, we will dive deeper into exploring a simple dataset (preferably IRIS), some important ML concepts and familiarizing ourselves with our tools.

Video 1.1

Demo machine learning product (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| Demonstrate pre-fit machine learning model; this is the format our final product will resemble | POST “test” data to the REST API | Retrieve predictions from the server | Examine predictions from ML model |

Video 1.2

Setting up our Anaconda Environment (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| Install Anaconda and use it to create an environment to manage our dependencies | Download Anaconda for Python 3.6 | Create the environment using the environment.yml file | Activate the environment |

Video 1.3

Introduction to the iPython notebook (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| We will be using Jupyter notebooks to interactively explore our datasets and generate visualizations. | Ensure Anaconda environment is active | Launch Jupyter from the command line | Get familiarized with the interface, notebooks, and creating new kernels |

Video 1.4

Loading and manipulating data with Pandas (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| Pandas allows us to query data in a SQL-esque fashion. We will use it to load data from disk and into memory, and then we will get familiar with various transforming functions | Load our two datasets from differing formats | Explore various slicing/dicing operations | Explore grouping functions and visualizations |

Video 1.5

ML Objective + Data splitting & common pitfalls (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| We need to define the broader objective of machine learning, and specifically the difference between supervised and unsupervised learning. This will help us understand why data splitting is absolutely critical in machine learning | Define supervised vs. unsupervised machine learning | Introduce data splitting concept | Use scikit-learn to create a valid train/test split |

Video 1.6

Descriptive statistics using Pandas (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| There are things we will not know about our dataset until we’ve begun to explore it both visually and statistically | Summarize the dataset in terms of several informative metrics | Create various visualizations to aid analysis | Identify relationships between existing variables |

Video 1.7

Planning our preprocessing stages (From the outline)

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| Now that we’ve summarized relationships in our data both visually and statistically, it’s time to plan how to clean up the dataset before we can model | Identify strategies for imputing missing data | Consider options for feature transformations | Discuss potentially valuable new features to engineer |