Wine Quality Prediction

Links

* [Github](https://github.com/manankumbhani/CS643PA2)

* [Docker](https://hub.docker.com/repository/docker/manankumbhani/pa2)

1. Parallel training implementation
* Create a cluster

We create a cluster of **5 nodes (1 master 4 slaves)** to train a ML model to predict the quality of the wine.
We use **AWS EMR** cluster management tool to achieve this.

* Upload files to the s3 bucket

Creating a Cluster

Step 1: click `EMR` in the AWS dashboard under `analytics` section
Step 2: Click `Create Cluster`
Step 3: Type desired cluster name in the `General Configuration` for `Cluster Name` .
***Step 3.1**: Under ``Software configuration` in the application column click the button which shows `Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.8.2``.
***Step 3.2:** Under `Hardware Configuration` click `m4.large` rather than the default `m5.xlarge` as the default m5.xlarge incurs a cost of $0.043/hr in contrast to the $0.03 for m4.large. Keep in mind that EMR incurs an additional 25% cost post first usage.
***Step 3.4:** Select `4` instances under the column `Number of instances`
Step 3.5: Under `Security and access` click the EC2 key pair if already created else create a new one
Step 4: Click Create Cluster button. Wait for around 15 minutes for the cluster to start functioning.

Run this command to parallelly train the data.
spark-submit --master spark://ip-172-31-27-247.ec2.internal:7077 train.py hdfs:///data/TrainingDataset.csv hdfs:///model

Then recompile your project

Docker Setup
1. Install latest docker.
2. Start docker.
3. Adding ec2 instance to docker.
4. Push image to Docker repository (docker push mak86/wine_quality:tagname)

5. Pull image from the Docker repository (docker pull mak86/wine_quality:tagname)

Then go to your ec2 and use following command to run the docker
```docker run --mount type=bind,source=/home/ec2-user/TestingDataset.csv,target=/TestingDataset.csv manankumbhani/pa2```
```