Machine Learning in the Cloud

TJ CCC

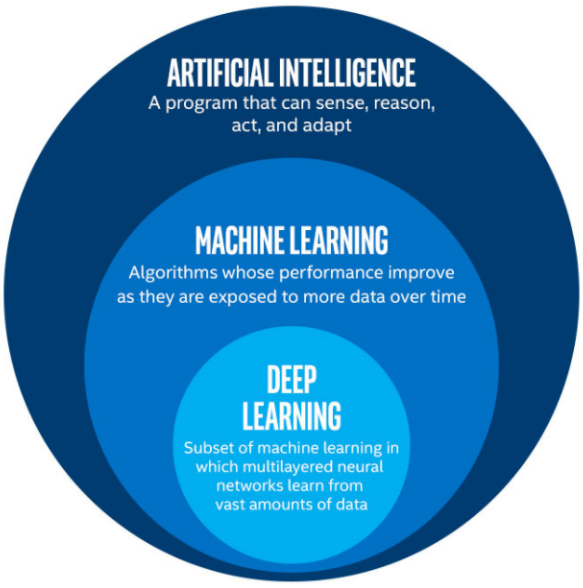
January 2020

1 Introduction

Cloud computing and machine learning are two very powerful, growing technologies that become even more powerful as complements. Machine Learning's few drawbacks are ameliorated with the benefits of cloud computing. We will use Azure to demo how powerful ML in the cloud actually is.

2 Machine Learning

Machine Learning is is a rapidly growing field that has become prevalent in all fields with applications everywhere. This powerful tool of making computers do tasks previously thought only humans could do has been able to accomplish extraordinary feats such as detecting cancer in early stages, driver cars, detect fake news, and beat world-class masters in chess.



2.1 Basics

1. First of all, you need data to train your model on. This is generally labeled data (supervised) such as an image with a label for what the image is of, or it can be non-labelled data (unsupervised) such as many images that need to be categorized by the objects in them.

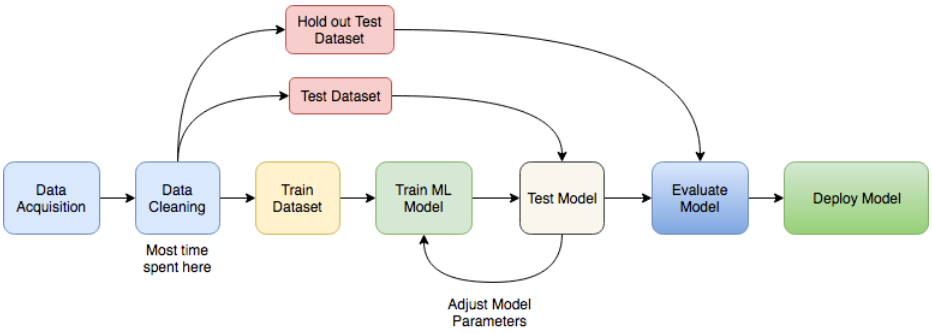
2. Clean, purify, and format the data

3. Split your data into 80% training data, 20% testing data. Your model will use the training data to "learn" the patterns of the data and then will be tested against the testing data to measure how good of a model it is.

4. Choose a model (e.g. neural network or one of its many variations, SVMs, Decision Trees etc.) and choose optimal hyperparameters. Hyperparameters are variables that the programmer needs to manually define that dictates how the model will train.

5. Train your model on the training data and evaluate the model using the testing data.

6. Deploy your model so it can be used in real-time!



2.2 Drawbacks

- Lots and lots of data – hard to store and access
- Extremely computationally expensive to train models
- Extensive knowledge is required to train models by coding them straight-up and still requires lots of time to achieve a useful model accuracy

hint, these problems can be solved with cloud computing :)

3 Benefits of ML in Cloud

1. **Compute Power** - As mentioned earlier, Machine learning is a very computationally expensive task. The resources needed to train models quickly and effectively are only readily accessible over the cloud. These resources are also easily scalable in the cloud making your machine learning tasks as efficient and affordable as possible.

2. **Storage Resources** - Machine learning also requires immense amounts of data (the more the data the better!) which is more easily stored and accessed in the cloud than on local machines.

3. **AutoML** - With cloud computing, anyone can train and test ML models with ease even with minimal AI or ML knowledge. Experts in machine learning are constantly working on making ML more easily accessible for everyone, creating AutoML services and deploying them in the cloud for us to use.

4 Public Cloud Provider Comparison

AWS, GCP, and Azure all provide services for speech, vision, natural language processing, video intelligence, and AutoML. AWS and GCP additionally offer services for translation and conversational interface (e.g. alexa).

4.1 AutoML

- AWS Sagemaker Studio
- Azure Machine Learning Studio
- GCP AutoML

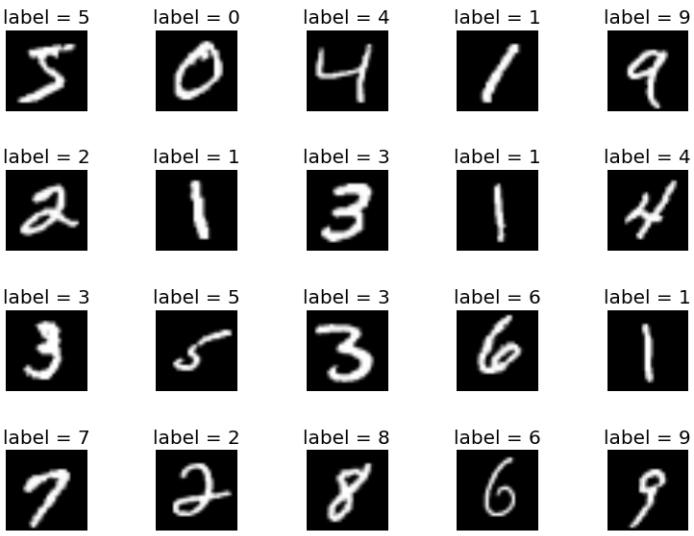
5 Demo - AutoML

5.1 General AutoML Procedure

1. Gather, purify, and format data according to service specifications
2. Pass data to AutoML service and set parameterse
3. Evaluate your model and deploy!

5.2 Demo in Azure

1. Sign into Azure and go to all services → Machine Learning
2. Create a new Azure Machine Learning Workspace (might take a few minutes)
   (a) Name: whatever you want
   (b) resource Group: Create one if you don't already have on
   (c) Location: East US (doesn't actually matter though)
   (d) Edition: Enterprise (note if you don't use Enterprise you won't have access to Azure ML Studio)
3. Go to Experiments and press *Launch the Studio*
4. Press *Automated ML* in the left pane then *New Automated ML Run* to train a new model
5. For your dataset, we will be using MNIST, a widely used beginner ML dataset. Press *Create Dataset* and choose *From Open Datasets*, then search for MNIST.
   (a) Subset: All
   (b) Register as: Tabular



6. Next, configure your run (training job)
   (a) Name: whatever you want
   (b) Target Column: Column785 (labels are in the last column, 785, for this specific dataset)
   (c) Compute Target: Make a new compute if you don't already have one – if you go to *Advanced settings* you can see references to VPCs and Subnets which we covered last week!
7. Choose Classification since we are classifying numbers
8. Press *View additional configuration settings* and under *Exit Criterion* set the metric score threshold to 0.90. This will make the run finish faster, but if you're ready to wait a few hours for an optimal model then leave this blank for now.
9. Press *Finish* and let your model train! Come back in a few minutes to see how it did