1. Favorite language:

My favorite languages are Python and Scala. Python is relative easy to learn but extremely useful. there are so many machine learning libraries playing an important role supporting the development of whole data science industry. Learning Scala is a tough experience, but once getting to it, I found it’s function programming attribute is very powerful. I also learn R, it is almost the best among statistic software.

2. Version of spark have you worked with? Latest version compared to spark 1.0

I always try to use the latest version of spark once it comes out. To be honese I’m not very familiar with old spark, but I believe, compared to older version, spark has fewer bugs now and has more library to use.

3. Favorite plotting library in Scala?

My favorite plotting library in Scala is Plotty. Actually not only in Scala, in python my favorite plotting library is also Plotty because its graph is very smooth and the color looks very comfortable. Especially it is very easy to build 3D graph by using Plotty.

4. Solve a binary classification problem?

I would like to use logistic regression to solve classification problem, since binary classification problem is easy compared to multi-class classification problems, then I will focus on the interpretation of model. Logistic regression’s advantage is its speed which benefits s lot in online deployment.

5. Deep learning methods have created/utilized?

I used CNN to do images labeling. Deep learning is kind of “black box” usually to deal with features that human can’t really directly understand such as pixels, sounds.

6. RNN? LSTM? CNN? FNN? Use cases?

RNN stands for recurrent neural net, it is usually used in time-sequenced dataset since it predicts time t based on result of last period (t-1) like predicting what will people say based on his/her last sentence.

LTSM is long-short term memory. It is an advanced version of RNN since we can control what information we will put into the model or not. It can be used to predict change stock price which is supposed to be affect a lot by investor’s sentiments. There are a lot Investors usually make decision only based on stock’s change one day before.

CNN is usually used to do image labeling.

FNN

7. Difference between LSTM/GRU, is it a vanishing gradient? why does that matter on our prior topic?

8. Difference between bidirectional LSTM and LSTM?

9. Difference between PCA and autoencoder? They are same or not? Would one linear project ever be preferred over the other?

Both of them can be used to do dimension reduction and data compression. The difference is PCA is usually not used to do feature selection but autoencoder do. PCA is often used to avoid collinearity. In linear project, they are basically exchangeable

10. How to prevent overfitting with neural networks? Dimension reduction? When would a non-linear form of that method be required? When to utilize it? How do we know optimal dropout rate? V.S L1, L2 regression? Lost functions of those two forms of regression?

There are many common ways to prevent overfitting in deep learning: dropput, adding data, cross validation and regularization. But we usually don’t use dimension reduction methods like PCA to solve overfitting, because we don’t want to lose many useful features in the beginning. Non-linear? We usually set dropout rate as 0.5. Dropout is better than L1, L2

11. Last two projects? Machine learning pipelines, algorithms used? Why should I care about your prior work?

You invited me to this interview probably because you find something interesting on my profile, you are interested in my experience and believe that I could be fitting to this position but was still have some doubt. My job is to use all I’ve experienced, learned in my prior work to release your doubt to show you why I fit this position.

12. Server-less architectures? How does that compare to traditional systems you have worked with in the past? How could that architectures effect how we deploy machine learning algorithms today?

Yes, I have used cloud platform such as google cloud platform, AWS a lot before. I think the best thing of these platforms is we can collaborate with team members much easier.