# Class Notes - SSW 2016 Data Science

Class Feedback Form: <http://tinyurl.com/FeedbackTD1>

This document can be found at c

Python + Scikit-Learn RULZZ

Table of contents

[Class Notes - SSW 2016 Data Science](#h.zc8yektdl270)

[Class Schedule](#h.e0jsmzgjjya5)

[Instructors & TAs](#h.ic0ko6odkux)

[File Sharing](#h.87rzswstlb96)

[Getting Help](#h.sk0iruvh98wy)

[Class Notes](#h.dbc10va2blmm)

[Q&A Parking Lot](#h.vpzddob8w5as)

## Class Schedule

**Location:** Imperial B, Floor 2, Hilton

**Training: 9:00am - 5:00pm**

* Coffee Break: 10:45am - 11am
* Lunch (provided): 12:00pm - 1:00pm
* Soda Break: 2:15pm - 2:30pm
* Coffee Break: 4pm - 4:15pm

## Instructors & TAs

*Instructors*

* Doug Bateman [doug.bateman@databricks.com](mailto:doug.bateman@databricks.com)
* Waleed Zaghloul

*TAs*

* Richard Garris [richard@databricks.com](mailto:richard@databricks.com)
* Peyman Mohajerian
* Silvio Fiorito

## File Sharing

* This document can be found at <http://tinyurl.com/Spark-DataScience-TD1>
* Your thumb drives have a copy of the labs & slides. They are for you to take with you.
* Any updates we made after class will be uploaded to the [class file share](https://drive.google.com/folderview?id=0B09cDg18tuRhTXlmNkptc1ZJMjA&usp=sharing).Surveys
* We value your feedback. Please do not leave without submitting the post-class feedback & evaluation. The conferences likes us to get a 95% feedback rate.
* If you haven’t already, it’s not too late to complete the pre-class skills assessment: <http://tinyurl.com/06-06-16TD1>

## Getting Help

* If you have trouble logging in or if the lab environment is broken, email [training-support@databricks.com](mailto:training-support@databricks.com)
* If you need help completing a lab, ask any TA.
* If you need help with the building facilities or have an emergency, contact the event coordinator at the door.
* If you have a question you’d like answered asynchronously, please post the question at the bottom of this document.
* Alert a TA immediately if
  + You cannot get on the WiFi at all, or you are on but cannot load websites
  + The presentation font size is too small and you can’t see the screen
  + The presenter is speaking too too fast and is hard to follow (sometimes we get excited)
  + If the main idea of the concepts just covered are unclear. (More advanced follow-up questions should be put in the Q&A section at the bottom of this document.)

## Class Notes

Notes the instructor creates during class will go here.

## Q&A Parking Lot

**Question from Doug Bateman**: What is this question parking lot?

**Likes this question:** Jacob Parr, Aaron Maxwell

**Answer from Doug Bateman:** With a very large class, we’ll have students post questions here. The TAs will then help answer the questions. And the questions with the most likes will also be answered by the instructor. So put your questions here using the format below.

**Question from Doug Bateman:** Here’s a sample question.

**Likes this question:** Taggart McCurdy, Andy Konwisky, +1 from Anonymous

**Answer from Doug Bateman:** Here’s a sample answer.

Q1: How long are you going to make <https://ssw2016.cloud.databricks.com> alive? When we go back after summit, can we still use access to it?

Answer: It will be up for at least a month. After that, you can upload the labs (available in the filestore) to your own Databricks CE account (Free).

Q2: Is the Spark dataframe identical to the pandas dataframe?

Ans: No :) Although acts similarly from code perspective, but under the hood is a distributed collection

Q3: In the Workspace what version of Jupyter did you setup?

A: This is using a Databricks notebook, not Jupyter

Q: What happens when you do a collect (i.e., action) on a very large dataset...does the data need to reside within the master memory (or driver)?

A: Yes, if you call collect on an RDD or DataFrame, it will bring ALL the data down to the Driver. So you need to use that very carefully or ensure your Driver has been assigned enough memory to keep all that data.

Q. Is there a driving reason to prefer Scala over Python?

A: ML features get written in Scala and then get migrated to Python later. And calling UDFs and map functions in scala is faster than Python. But for the most part, Python works great.

Q: Dataframes seem to be useful for loading/querying data to a form you can manipulate. Can you write ML algorithms using dataframes?

A: If you want to implement custom algorithms not supported out of the box, the best thing to do is use the spark.ml API to create custom transformers and estimators. This will let you define reusable libraries which operate on DataFrames. See more here: <http://spark.apache.org/docs/latest/ml-guide.html> To simply define a model then you would use Pipelines to do feature engineering, training, and get the model to use for predictions.

Q: Where can I find the slides?

Answer from Doug: Slides are now [uploaded here](https://drive.google.com/folderview?id=0B09cDg18tuRhTXlmNkptc1ZJMjA&usp=sharing)… and also on the thumb drive.

Note from Anonymous: Time series weather application @ <https://spark-packages.org/package/killrweather/killrweather>

Q: In DataFrames how do we evaluate models?

A: <https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.ml.evaluation.RegressionEvaluator>

Q: Could you provide a basic code example of how RDDs differ from DataFrame in Python? I’m trying to understand the difference clearly. :)

Answer from Doug: Take a look at page 52 in [this slide deck](https://drive.google.com/a/databricks.com/file/d/0B1BTTPu5-xDOV3dxQmFsUXZQbE0/view?pli=1)…

Using RDDs:

rdd = sc.textFile(...).map(\_.split(" "))

rdd.map { x => (x(0), (x(1).toFloat, 1)) }

.reduceByKey { case ((num1, count1), (num2, count2)) =>(num1 + num2, count1 + count2)}

.map { case (key, (num, count)) => (key, num / count) }.

.collect()

Using DataFrames

df.groupBy("key").agg(avg("value")).collect()

Q: Will the lab notebooks work in Zeppelin?

Answer from Richard at Databricks (@rlgarris) - You can export them as iPython Notebook format. Not sure if Zeppelin supports importing as ipyn format. You can also export as raw python/scala source.

Q: Where is the temp table stored, HDFS?

Answer from Doug: Temp tables are not stored at all. .registerTempTable() simply makes the existing Dataframe have a SQL table name for use in SQL. However “saveTable” does write a table. And it writes it out to whatever path you specify (HDFS, S3, local files, etc)

A: Temp tables are just a way to temporarily persist a name and schema for the table in order to reference it in SQL queries. The data is still the original data loaded via the sqlContext (or if you cache the data, then it’s in memory). So, temp tables were aliases for the data, they are not actually stored

Q: What’s the release date of Spark 2.0?

Answer: TBD, it’s possible projected release dates to be announced tomorrow at the keynote. KeyNotes are going to be AWESOME so you should go.

Q: if I have Avro file with schema ( in json format) embedded, can it be read by dataframe or do i need to define schema in my code for dataframe

A: Avro is supported through a custom data source which you can find here: https://github.com/databricks/spark-avro

Q: Is SVD only runs on driver?

A: No, it’s parallelized.

Q. Why doesn’t this throw an exception?

val sparseVector = Vectors.sparse(**5**, Array(2, **7**), Array(1.0, 5.0)).asInstanceOf[SparseVector]

A: Because sparseVectors actually allow non-uniform dimensions… so you can have a vector of vectors of different lengths. Errors only happen later when you try and do math operations that don’t fit the dimensions.

Q. When creating partitions, is it possible or does Spark provide a way to determine hot spots or skew in my data?

A: You can detect skew while running jobs by looking at the Driver UI and finding longer running tasks. Another good indicator is if you’re using a non-splittable input format (e.g. GZip) then each task will get one whole file. So if you’re pulling in a bunch of files of differing sizes you’ll have skew based on the input file sizes. If you know you may have skew you can repartition the data to better distribute it across your executors after it’s been loaded. This also applies to DataFrames, where you can repartition by one or more columns and specify the number of partitions.

Q. How can you add code completion to the notebook?

A: Code completion will work as long as you have the proper imports and simply using the tab key.

Addendum from Doug: In Python, you need to define the variable, run the cell, and then it will remember what type the variable was. (Python doesn’t have static typing, so running the cell is required in order to know what the variable is.)

Q: What does the parqUUID in the “Saving our DataFrame” section of the 2nd part of the kmeans tutorial do ( df.write.mode('overwrite').parquet('/path/...'.format(parqUUID)) )? +1

A. Generates GUID to save data to (GUID is used so less likely to have a collision where two people use the same filename in the class). This only needed when multiple people share a cluster. (In our class we all had seperate clusters.)

Q: The Databricks notebook tool we are using, can that be downloaded and deployed locally or is it only available in the cloud?

A: This is running on AWS. It cannot be run locally or downloaded as a binary package. If you have questions on running on other cloud providers, please stop by the booth.

Q: Is the Python interface of Spark Python 2.7 or Python 3.x?

2.7

A: Spark supports both 2.7 and 3.x. Databricks let you pick which one you want by configuring a global setting.

Q. Search for nan (Not a Number)

Look up is nan in fn

Q: Iris data set - What do the features in the Iris dataset represent? What does the label represent?

Features

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

see <https://archive.ics.uci.edu/ml/datasets/Iris> for more on the data set

Question: Where can I get the presentation on DataSets

Answer: <https://www.youtube.com/watch?v=pZQsDloGB4w>

Title: “RDDs, DataFrames and Datasets in Apache Spark - NE Scala 2016”

Presenter: “Brian Clapper”

Type: “Video”

Length: “48:04”

URL: “<https://www.youtube.com/watch?v=pZQsDloGB4w>

Question : Arrays vs. Vectors? Is there a source / your views on when to use? Is it more Vectors work best with ML?

A: Many MLlib machine learning algorithms require vectors as their input, but that is changing to DataFrames in ML.

Q. Does help in python have an equivalent in scala?

A: ScalaDoc (online). [So no. But IntelliJ lets you cover and get scaladoc inline.]

Question: Can you use Normalizer to normalize by partition and not the whole dataset?

Answer: Generally you don’t want to be grouping by partition as that requires control at the RDD level and is easier solved by groupBy. In this case you’d filter the RDD for each group, standardize each part, and then union them back together. (Normalizing on the other hand wouldn’t matter, as each row is normalized independently of the other.)

Q: Discretizer - roughly fixed length bucketizer

A: QuantileDiscretizer uses a sample of the data to generate roughly equally-sized buckets. Remember this is only an estimate - only available in Scala. .binmode in R

<https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.ml.feature.QuantileDiscretizer>

Q. If you don’t execute until an action is called, what is the benefit of creating a pipeline vs just describing each transformation?

A: Defining the pipeline serves another purpose as well: automation. The pipeline can be used against the same data and/or other data sets by just executing the pipeline stages.

Q. What was the significance of the date April 20 with the wiki traffic prediction walkthrough you discussed before break? 420?

A: 420.

Q. Is it possible to use a `when()` statement with multiple conditions? For example, if the value in one column is greater than 5 in a certain column, and if a value is less than 2 in another column, then return 1 (`otherwise()` return some other value).

Yes. We support standard SQL Case When Syntax.

Q. Is it possible to use the `agg()` function on grouped data to get two different aggregate types (i.e., a new ‘max’ column and a new ‘min’ column based on the same grouped data).

A: Yes

df.groupBy(“col2”).agg(count(col1), min(col1), max(col1))

Q. I see Java API/samples in spark documentation , how good are those api coverages compared to scala ( both run on jvm though!!)

Answer from Richard at Databricks: We generally maintain Java APIs as a best effort but Scala is the most up to to date.

Q. where can I find a ton of data science jokes and cartoons?

Answer from Anonymous:

<http://discuss.analyticsvidhya.com/t/friday-funalytics-share-your-favourite-data-science-analytics-statistics-jokes-or-cartoons/435/4>

Comment from Doug at Databricks: Thanks!

Q. I tried to use Window functions with data frames in pyspark. It worked fine in the shell, but when I ran the app, Spark throws an error that I should use HiveContext! It seems some features in HiveContext is not in SqlContext! How I can use Window functions in a PySpark app?

Answer from Anonymous: You will need to instantiate a Hive Context and create a HiveContext with a SQLContext backend.

Answer from Doug: When you instantiate the SqlContext, instantiate a HiveContext instead. **val** sqlContext **=** **new** org.apache.spark.sql.hive.**HiveContext**(sc)

Q. For the Wikipedia demo, I don’t think you are plotting the RMSE (in the histogram) as the root-mean-square-error is non-negative [ see <https://en.wikipedia.org/wiki/Root-mean-square_deviation> ]. Mostly likely you are plotting just the residual = observed value - predicted value.

Answer from Doug: You are completely correct. It’s plotting the residual. The RMSE is simply the std-dev of the residuals.

Q. What are the best practices for creating/sharing sparkcontext and/or sqlcontext between driver and executor nodes? (I’ve had spark complain that I’m trying to do transforms when I’m not in executor nodes when doing spark streaming and trying to query using elasticsearch-hadoop)

Answer from Doug: The SparkContext and and SqlContext are only available on the driver. The executors can do compute work, but cannot kick of spark jobs their own. If you share more about what you’re trying to accomplish we might be able to suggest a different approach to accomplish your goals.

Q: Can we get email address to Doug Bateman, in case we have some questions afterwards?

A: It’s at the top of this doc. And here: [doug.bateman@databricks.com](mailto:doug.bateman@databricks.com). I’m notorious for letting his inbox pile up… so feel free to email me multiple times if needed.

Q: Do we have a slack group or something similar?

A: We can use the Databricks forums:

<https://forums.databricks.com>

Q: I am trying to grab my labs and data off the data science training accounts we used. However, I can no longer access using the given email and password. Can someone please help especially all the data sets we used in the labs?