Getting started with SFrame and GraphLab Create

Why SFrame & GraphLab Create

There are many excellent machine learning libraries in Python. One of the most popular one today is scikit-learn. Similarly, there are many tools for data manipulations in Python; a popular example is Pandas. However, most of these tools do not scale to large datasets, including some we will tackle in this Specialization. In addition, in this specialization, we will cover a wide range of ML models, feature engineering transformation, and evaluation metrics. With most existing packages, you will have to install a combination of packages to get the tools that we need to tackle the use cases in this course. This is possible, but requires advanced knowledge of Python, which we feel will slow down most people's learning of the core concepts.

The main goal of this course is to learn core ML concepts, not how to use a specific software package. Thus, in this course, we recommend you use GraphLab Create, a package we have been working on for many years now, and has seen an exciting adoption curve, especially in industry with folks building real applications. GraphLab Create is a highly scalable machine learning library for Python, which also includes the SFrame, a highly-scalable library for data manipulation. A huge advantage of SFrame over Pandas is that with SFrame, you are not limited to datasets that fit in memory, which allows you to deal with large datasets, even on a laptop. (The SFrame API is very similar to Pandas' API. ➔https://turi.com/learn/translator/.)

Licenses for SFrame & GraphLab Create

The SFrame package is available in open-source under a permissive BSD license. So, you will always be able to use SFrames for free.

GraphLab Create is free on a 1-year, renewable license for educational purposes, including Coursera. This software, however, has a paid license for commercial purposes. You can get the GraphLab Create academic license at the following link:

https://dato.com/learn/coursera/

For full disclosure!

GraphLab Create is very actively used in industry by a large number of companies. This package was created by a machine learning company called Dato. This company is spin off from a popular research project called GraphLab, which Carlos Guestrin, one of your two instructors, and his research group started at Carnegie Mellon University. In addition to being a professor at the University of Washington, Carlos is the CEO of Dato.

The reason we suggest you use GraphLab Create is not because Carlos is the CEO of Dato :), but because we very strongly believe using this software will make it much easier for us to follow the "case-study approach" we are taking in this specialization. In particular, it will let you focus

on exploring each case study in this first course, without having to implement your own algorithms from scratch, and benefiting from the performance advantages that GraphLab Create provides. ***In subsequent courses in the specialization, you will be implementing many of these algorithms from scratch, having had the foundation of seeing them perform in practice on real applications.***

We are happy, however, for you to use any tool(s) of your liking, by following the steps below. As you will notice, we are only grading the output of your programs, so the specific software tool is not the focus of the course.

It's important to emphasize that this specialization is **not** about providing training for a specific software package. The goal of the specialization is for your effort to be spent on learning the fundamental concepts and algorithms behind machine learning in a hands-on fashion. These concepts transcend any single package. What you learn here you can use whether you write code from scratch, use any existing ML packages out there, or any that may be developed in the future. We are happy to hear that so many of you are enjoying this approach so far!

Using other ML packages

We strongly encourage you to use SFrame for this course.

You are welcome to use other ML packages, like scikit-learn, instead of GraphLab Create. However, we believe this will significantly slow down the your implementation tasks, especially for this first course.

The first course is focused on exploring the use cases we'll tackle throughout the specialization. A huge goal here is to familiarize ourselves with the core ML concepts that we will use the 5 follow-on courses. In those courses, there will be much more implementation of ML algorithms, so the specific ML package becomes less important. But, in this first course, we want to move quickly through all the use cases, and GraphLab Create will help us do just that.

If you choose to use a different package, we will provide the data sets and the assignment questions will not depend specifically on GraphLab Create.
