Using MLlib

One of the reasons we use spark is for easy access to powerful data analysis tools. The MLlib library gives us a machine learning library that is easy to use and utilizes the scalability of the Spark system.

It has supported APIs for Python (with NumPy), R, Java and Scala.

We will use the Python version in a generic manner that looks very similar to any of the above implementations.

There are good example documents for the clustering routine we are using, as well as alternative clustering algorithms, here:

http://spark.apache.org/docs/latest/mllib-clustering.html

And an excellent API reference document here:

http://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.clustering.KMeans

I suggest you use these pages for all your Spark work.
Clustering is a very common operation for finding grouping in data and has countless applications. This is a very simple example, but you will find yourself reaching for a clustering algorithm frequently in pursuing many diverse machine learning objectives, sometimes as one part of a pipeline.

Coin Sorting

Size

Weight
As intuitive as clustering is, it presents challenges to implement in an efficient and robust manner.

You might think this is trivial to implement in lower dimensional spaces.

But it can get tricky even there.

Sometimes you know how many clusters you have to start with. Often you don’t. How hard can it be to count clusters? How many are here?

We will start with 5000 2D points. We want to figure out how many clusters there are, and their centers. Let’s fire up pyspark and get to it...
Finding Clusters

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

```python
>>> rdd1 = sc.textFile("5000_points.txt")
```

```python
>>> rdd2 = rdd1.map(lambda x: x.split())
```

```python
>>> rdd3 = rdd2.map(lambda x: [int(x[0]), int(x[1])])
```

```
Finding Clusters
Read into RDD
Transform to words and integers
```
```python
>>> rdd1 = sc.textFile("5000_points.txt")
>>> rdd1.count()
5000
>>> rdd1.take(4)
['  664159  550946', '  665845  557965', '  597173  575538', '  618600  551446']
>>> rdd2 = rdd1.map(lambda x:x.split())
>>> rdd2.take(4)
[['664159', '550946'], ['665845', '557965'], ['597173', '575538'], ['618600', '551446']]
>>> rdd3 = rdd2.map(lambda x: [int(x[0]), int(x[1])])
>>> rdd3.take(4)
[[664159, 550946], [665845, 557965], [597173, 575538], [618600, 551446]]
```
Finding Clusters

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

```python
rdd1 = sc.textFile("5000_points.txt")

rdd2 = rdd1.map(lambda x: x.split())

rdd3 = rdd2.map(lambda x: [int(x[0]), int(x[1])])

from pyspark.mllib.clustering import KMeans
```

Using KMeans for clustering:

```
class KMeans

New in version 0.9.0.

Methods:

- train(rdd, k, maxIterations=100, runs=1, initializationMode=\"k-means\"
  seed=None, initializationSteps=5, epsilon=0.0001, initModel=None)

Train a k-means clustering model.
```

Parameters:
- `rdd` — Training points as an RDD of Vector or convertible sequence types.
- `k` — Number of clusters to create.
- `maxIterations` — Maximum number of iterations allowed. (default: 100)
- `runs` — This param has no effect since Spark 2.0.0.
- `initializationMode` — The initialization algorithm. This can be either \"random\" or \"k-means\". (default: \"k-means\")
- `seed` — Random seed value for `k-means` initialization. Set as None to generate seed based on system time. (default: None)
- `initializationSteps` — Number of steps for the `k-means` initialization mode. This is an advanced setting — the default of 5 is almost always enough. (default: 5)
- `epsilon` — Distance threshold within which a center will be considered to have converged. If all centers move less than this Euclidian distance, iterations are stopped. (default: 1e-4)
- `initModel` — Initial cluster centers can be provided as a KMeansModel object rather than using the random or \"k-means\" initializationModel. (default: None)
Finding Clusters

What is the exact answer?
Finding Clusters

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

```python
>>> rdd1 = sc.textFile("5000_points.txt")
>>> rdd2 = rdd1.map(lambda x:x.split())
>>> rdd3 = rdd2.map(lambda x: [int(x[0]),int(x[1])])

>>> from pyspark.mllib.clustering import KMeans

>>> for clusters in range(1,30):
...    model = KMeans.train(rdd3, clusters)
...    print (clusters, model.computeCost(rdd3))
```

Let’s see results for 1-30 cluster tries

```
1 5.76807041184e+14
2 3.4183673951e+14
3 2.23097486536e+14
4 1.64792608443e+14
5 1.19410028576e+14
6 7.97690150116e+13
7 7.1645159444e+13
8 4.81469246295e+13
9 4.23762700793e+13
10 3.65230706654e+13
11 3.1699187996e+13
12 2.94369408304e+13
13 2.04031903147e+13
14 1.37018893034e+13
15 8.91761561687e+12
16 3.1833652006e+13
17 1.39010717893e+13
18 8.22806178508e+12
19 8.22513516563e+12
20 7.935929283e+12
21 7.79615059172e+12
22 7.70001662709e+12
23 7.24231610447e+12
24 7.21990743993e+12
25 7.09395133944e+12
26 6.92577789424e+12
27 6.53939015776e+12
28 6.57782690833e+12
29 6.37192522244e+12
```
```python
for trials in range(10):
    ...
    print
    ...
    for clusters in range(12,18):
        ...
        model = KMeans.train(rdd3,clusters)
        ...
        print (clusters, model.computeCost(rdd3))
```

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Compute Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2.45472346524e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.00175423869e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.90313863726e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.52746006962e+13</td>
</tr>
<tr>
<td>16</td>
<td>8.67526114029e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.49571894386e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.62619056924e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.9031673822e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.52308079405e+13</td>
</tr>
<tr>
<td>15</td>
<td>8.91765957989e+12</td>
</tr>
<tr>
<td>16</td>
<td>8.70736515113e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.49616440477e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.5524719797e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.1432949698e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.11070395905e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.47792736325e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.8573695725e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.42795740134e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.31466520037e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.91856542103e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.49332023312e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.3506302755e+13</td>
</tr>
<tr>
<td>16</td>
<td>8.7757678836e+12</td>
</tr>
<tr>
<td>17</td>
<td>1.60075548613e+13</td>
</tr>
<tr>
<td>12</td>
<td>2.5187054064e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.83498739266e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.96076943156e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.4172566214e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.41986217172e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.46755159547e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.38234539188e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.85101922046e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.91732620477e+13</td>
</tr>
<tr>
<td>15</td>
<td>8.91769396968e+12</td>
</tr>
<tr>
<td>16</td>
<td>8.64876051004e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.54677681587e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.5187054064e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.0431903147e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.95213876047e+13</td>
</tr>
<tr>
<td>15</td>
<td>9.3000628589e+13</td>
</tr>
<tr>
<td>16</td>
<td>2.07670831868e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.47797102908e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.39830397362e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.00248378195e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.34867337672e+13</td>
</tr>
<tr>
<td>15</td>
<td>2.0929321238e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.32266735736e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.50857884943e+12</td>
</tr>
</tbody>
</table>
for trials in range(10):                          #Try ten times to find best result
    for clusters in range(12, 16):                 #Only look in interesting range
        model = KMeans.train(rdd3, clusters)   
        cost = model.computeCost(rdd3)         
        centers = model.clusterCenters         
        if cost<1e+13:                             #If result is good, print it out
            print (clusters, cost)            
            for coords in centers:         
                print (int(coords[0]), int(coords[1]))
            break
...
16 Clusters
We are going to find a recurring theme throughout machine learning:

• Our data naturally resides in higher dimensions

• Reducing the dimensionality makes the problem more tractable

• And simultaneously provides us with insight

This last two bullets highlight the principle that "learning" is often finding an effective compressed representation.

As we return to this theme, we will highlight these slides with our Dimensionality Reduction badge so that you can follow this thread and appreciate how fundamental it is.
The problems we are going to address, as well as the ones you are likely to encounter, are naturally highly dimensional. If you are new to this concept, let's look at an intuitive example to make it less abstract.

<table>
<thead>
<tr>
<th>Category</th>
<th>Purchase Total ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children's Clothing</td>
<td>$800</td>
</tr>
<tr>
<td>Pet Supplies</td>
<td>$0</td>
</tr>
<tr>
<td>Cameras (Dash, Security, Baby)</td>
<td>$450</td>
</tr>
<tr>
<td>Containers (Storage)</td>
<td>$350</td>
</tr>
<tr>
<td>Romance Book</td>
<td>$0</td>
</tr>
<tr>
<td>Remodeling Books</td>
<td>$80</td>
</tr>
<tr>
<td>Sporting Goods</td>
<td>$25</td>
</tr>
<tr>
<td>Children's Toys</td>
<td>$378</td>
</tr>
<tr>
<td>Power Tools</td>
<td>$0</td>
</tr>
<tr>
<td>Computers</td>
<td>$0</td>
</tr>
<tr>
<td>Garden</td>
<td>$0</td>
</tr>
<tr>
<td>Children's Books</td>
<td>$180</td>
</tr>
</tbody>
</table>

This is a 2900 dimensional vector.
Why all these dimensions?

If we apply our newfound clustering expertise, we might find we have 80 clusters (with an acceptable error).

People spending on “child’s toys” and “children’s clothing” might cluster with “child’s books” and, less obvious, "cameras (Dashcams, baby monitors and security cams)", because they buy new cars and are safety conscious. We might label this cluster "Young Parents". We also might not feel obligated to label the clusters at all. We can now represent any customer by their distance from these 80 clusters.

<table>
<thead>
<tr>
<th>Customer Representation</th>
<th>80 dimensional vector.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>College Athlete</td>
</tr>
<tr>
<td>Young Parents</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We have now accomplished two things:

- we have compressed our data
- learned something about our customers (who to send a dashcam promo to).
This is a good time to point out how our intuition can lead us astray as we increase the dimensionality of our problems - which we will certainly be doing - and to a great degree. There are several related aspects to this phenomenon, often referred to as the *Curse of Dimensionality*. One root cause of confusion is that our notion of Euclidean distance starts to fail in higher dimensions.

These plots show the distributions of pairwise distances between randomly distributed points within differently dimensioned unit hypercubes. Notice how all the points start to be about the same distance apart.

Once can imagine this makes life harder on a clustering algorithm!

There are other surprising effects: random vectors are almost all orthogonal; the unit sphere takes almost no volume in the unit square. These cause all kinds of problems when generalizing algorithms from our lowly 3D world.
Even the definition of distance (the *metric*) can vary based upon application. If you are solving chess problems, you might find the Manhattan distance (or taxicab metric) to be most useful.

For comparing text strings, we might choose one of dozens of different metrics. For spell checking you might want one that is good for phonetic distance, or maybe edit distance. For natural language processing (NLP), you probably care more about tokens.

For genomics, you might care more about string sequences.

Some useful measures don’t even qualify as metrics (usually because they fail the triangle inequality: $a + b \geq c$).
Alternative DR: Principal Component Analysis

3D Data Set

Maybe mostly 1D!
Alternative DR: Principal Component Analysis

Flatter 2D-ish Data Set

View down the 1st Princ. Comp.
Let's look at one more example today. Suppose we are trying to do a Zillow type of analysis and predict home values based upon available factors. We may have an entry (vector) for each home that captures this kind of data:

<table>
<thead>
<tr>
<th>Home Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
</tr>
<tr>
<td>Longitude</td>
</tr>
<tr>
<td>Last Sale Price</td>
</tr>
<tr>
<td>Last Sale Year</td>
</tr>
<tr>
<td>Width</td>
</tr>
<tr>
<td>Depth</td>
</tr>
<tr>
<td>Floors</td>
</tr>
<tr>
<td>Bedrooms</td>
</tr>
<tr>
<td>Bathrooms</td>
</tr>
<tr>
<td>Garage</td>
</tr>
<tr>
<td>Yard Width</td>
</tr>
<tr>
<td>Yard Depth</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

There may be some opportunities to reduce the dimension of the vector here. Perhaps clustering on the geographical coordinates...
Principal Component Analysis Fail

1st Component Off
Data Not Very Linear

D x W Is Not Linear
But (DxW) Fits Well

Non-Linear PCA?
A Better Approach Tomorrow!
Why Would An Image Have 784 Dimensions?

MNIST 28x28 greyscale images
Central Hypothesis of Modern DL

Data Lives On
A Lower Dimensional Manifold

Maybe Very Contiguous

Images from Wikipedia
import numpy as np
import matplotlib.pyplot as plt
from sklearn import (datasets, decomposition, manifold, random_projection)

def draw(X, title):
    plt.figure()
    plt.xlim(X.min(0)[0], X.max(0)[0]); plt.ylim(X.min(0)[1], X.max(0)[1])
    plt.xticks([]); plt.yticks([])
    plt.title(title)
    for i in range(X.shape[0]):
        plt.text(X[i, 0], X[i, 1], str(y[i]), color=plt.cm.Set1(y[i] / 10.))

digits = datasets.load_digits(n_class=6)
X = digits.data
y = digits.target
rp = random_projection.SparseRandomProjection(n_components=2, random_state=42)
X_projected = rp.fit_transform(X)
draw(X_projected, "Sparse Random Projection of the digits")

X_pca = decomposition.PCA(n_components=2).fit_transform(X)
draw(X_pca, "PCA (Two Components)")

tsne = manifold.TSNE(n_components=2, init='pca', random_state=0)
X_tsne = tsne.fit_transform(X)
draw(X_tsne, "t-SNE Embedding")

plt.show()
How does all this fit together?
As the Data Scientist wanders across the ill-defined boundary between Data Science and Machine Learning, in search of the fabled land of Artificial Intelligence, they find that the language changes from programming to a creole of linear algebra and probability and statistics.