livelessons

video instruction from technology experts

Lesson 3: Your First Spark Application

3.12 Challenges of k-means: Latent Features, Interpretation, and Validation

livelessons

Latent Features

Centroids of each cluster are representative points

- “Average” user
- Document topic
- Movie genre

```
print "Top Words for each Cluster:\n"

for i , v in df_index.loc[topics[top_topics][:,0]].iterrows():
    print "%d: %s" % (i, ", ".join(top_terms[j][0] for j in v.b.argsort()[::-1][:top_n]))
    print "\n"
```

```
Top Words for each Cluster:

32: computer, children, We, project, science, learning, program, work, would, experience, world, trip, use, technolog
y, 's, year, The, class, classroom, digital


66: reading, books, book, read, rug, children, English, classroom, writing, class, materials, love, year, literature,
language, 's, grade, would, learning, time


80: books, library, reading, read, readers, level, book, classroom, levels, children, love, grade, independent, serie
s, second, nonfiction, sets, first, leveled, My


33: math, calculators, overhead, projector, manipulatives, calculator, concepts, Math, graphing, use, mathematics, pr
oblems, learning, fractions, mathematical, materials, solving, hands-on, children, understanding


43: words, writing, word, letters, letter, sounds, machine, children, write, centers, sight, spelling, center, vocabu
lary, alphabet, phonics, reading, literacy, laminating, English


1: science, Science, hands-on, owl, materials, life, microscopes, experiments, Social, kit, Studies, pellets, study,
scientific, kits, learn, hands, explore, animals, curriculum


7: art, paint, supplies, artists, Art, painting, express, paper, projects, crayons, children, creative, artwork, mate
rials, clay, markers, artist, brushes, drawing, work


87: music, instruments, musical, CD, play, band, Music, recorder, instrument, player, program, songs, sound, rhythm,
CDs, drums, singing, guitar, playing, sing
```

Testing an algorithm

- Use a reference implementation (ex: `scikit-learn`)
- Synthetic input (or predictable input)
- Visualize results and inspect manually
- Test components in isolation
    - ex: don’t test BoW at the same time as k-means

Challenges

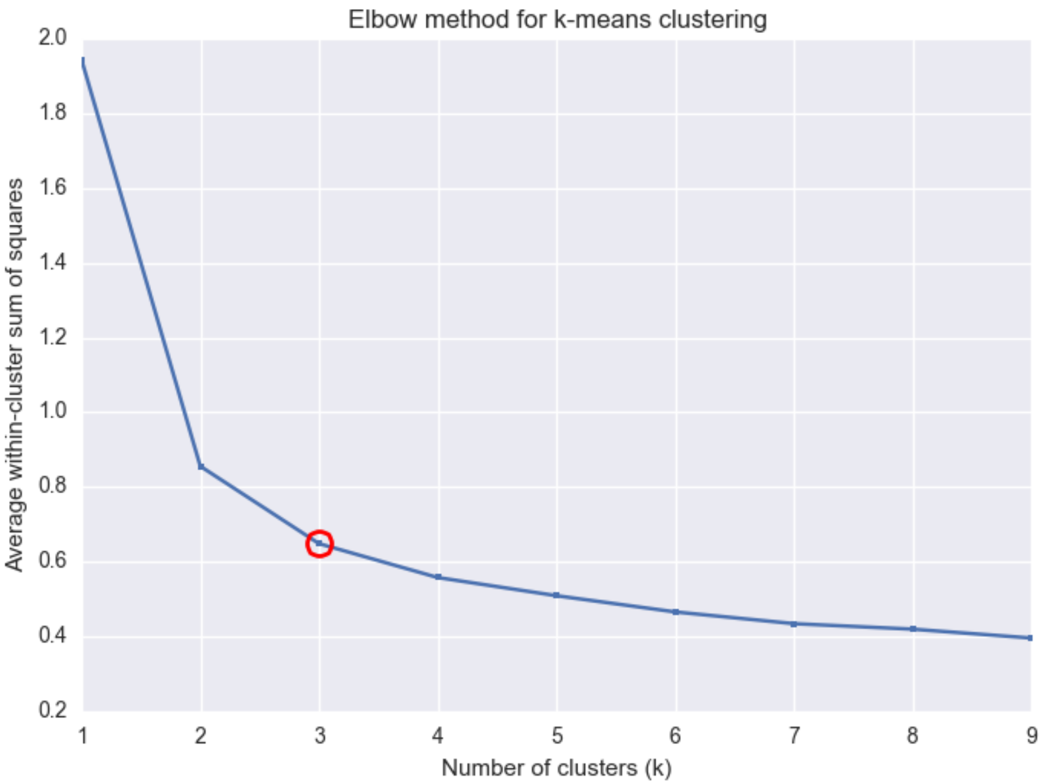
- How many k do we choose?
- Stochastic with local optimum (randomized greedy algorithm)
- Often need to preprocess/scale input features
- Which initial points to choose for centroids?
- Only Euclidean distance (the means of Kmeans)

Choosing K

- Elbow Method
- Silhouette Statistic
- Gap Statistic

Elbow Method

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$



Stochastic Greedy Algorithm

- Multiple runs of algorithm
- Pick assignments with lowest within cluster sum of squares

Initial Centroids

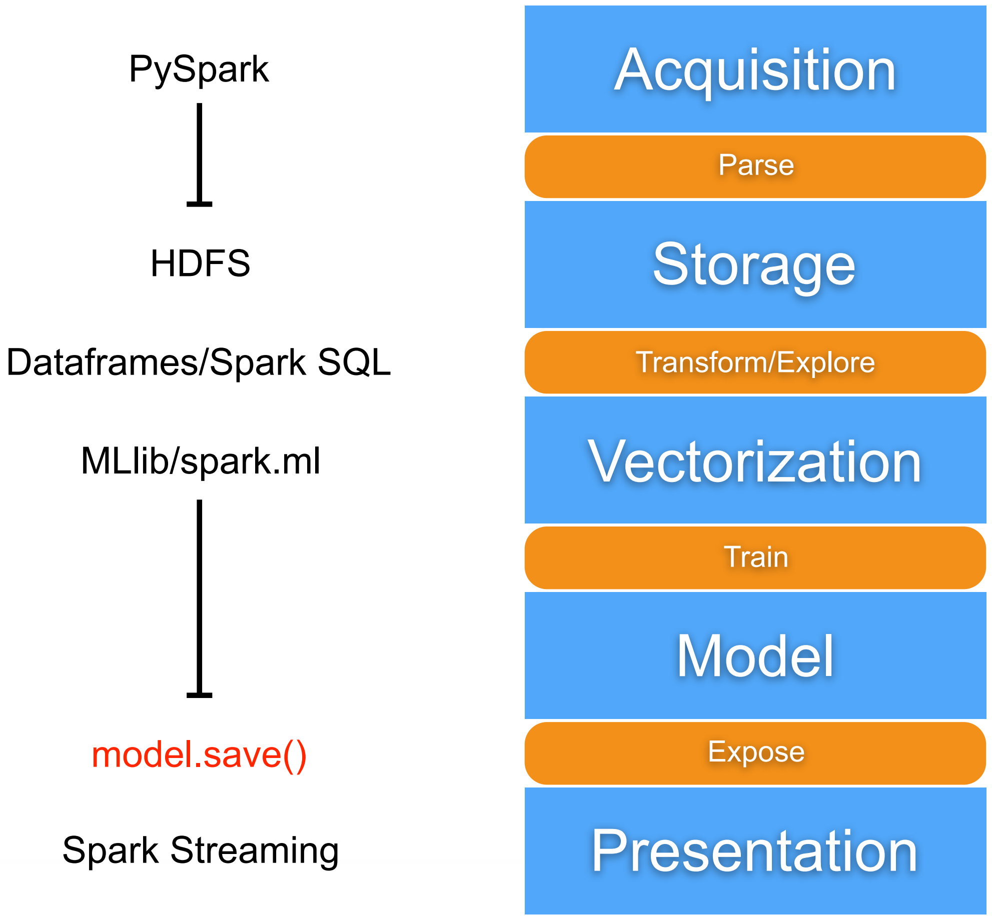
- Multiple runs of algorithm
- `kmeans++` (and `kmeans||`)

k-means Review

- Group similar (in vector space) points
- Discover hidden properties of each cluster
- Iteratively (but greedily) improve clusters
- Difficult to validate without labels (often takes manual inspection)

At Scale

Data Pipeline



PySpark

HDFS

Dataframes/Spark SQL

MLlib/spark.ml

model.save()

Spark Streaming

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

Model

Expose

Presentation

Serialization

A model is just a function….

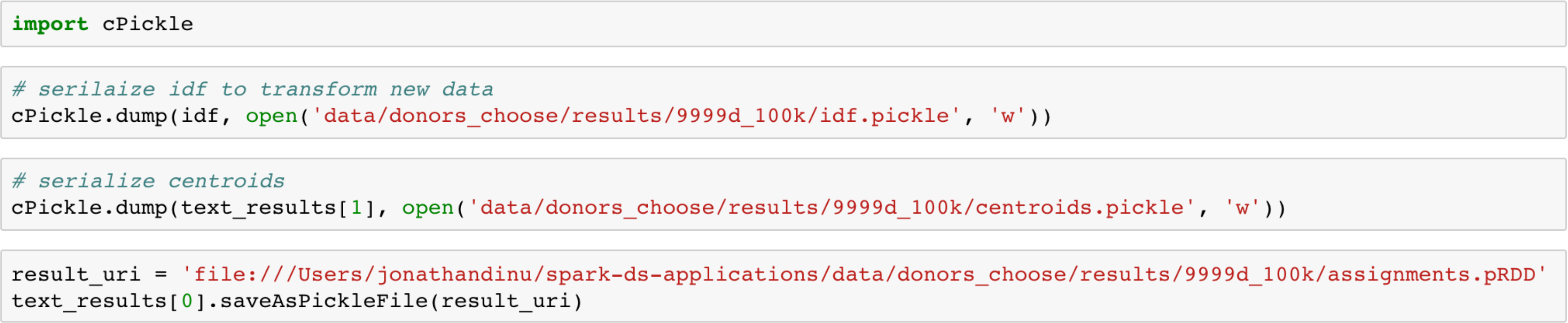
- Inputs
- Outputs

Serialization

You can store its parameters:

- Disk
- Database
- Memory

Serialization



```
import cPickle
```

```
# serilaize idf to transform new data
cPickle.dump(idf, open('data/donors_choose/results/9999d_100k/idf.pickle', 'w'))
```

```
# serialize centroids
cPickle.dump(text_results[1], open('data/donors_choose/results/9999d_100k/centroids.pickle', 'w'))
```

```
result_uri = 'file:///Users/jonathandinu/spark-ds-applications/data/donors_choose/results/9999d_100k/assignments.pRDD'
text_results[0].saveAsPickleFile(result_uri)
```

Note: It is best to serialize to S3 if running your analysis on a cluster

Review

- Often the place you train a model is not where you predict
- A model’s “learning” is capture in its parameters
- We can serialize a trained model to deploy elsewhere
- Once we have the cluster centroids, we can interpret them locally

Next Up: Spark Internals