





Spark Machine Learning

- Spark-ML (new developments)
 - Uses Dataset API
- Spark-MLlib (in maintenance mode)
 - Uses RDD API

Example Problem

Classify protein fold as either all alpha or all beta given the protein sequence.

MHHHHHHSSGRENLYFQGMTVREKTRLEKFRQLLSSQNTDLDELRKCS WPGVPREVRPITWRLLSGYLPANTERRKLTLQRKREEYFGFIEQYYDSR NEEHHQDTYRQIHIDIPRTNPLIPLFQQPLVQEIFERILFIWAIRHPASGYV QGINDLVTPFFVVFLSEYVEEDVENFDVTNLSQDMLRSIEADSFWCMSK LLDGIQDNYTFAQPGIQKKVKALEELVSRIDEQVHNHFRRYEVEYLQFAF RWMNNLLMRELPLRCTIRLWDTYQSEPEGFSHFHLYVCAAFLIKWRKEIL DEEDFQGLLMLLQNLPTIHWGNEEIGLLLAEAYRLKYMFADAPNHYRR





PDB ID: 3DZX - all alpha

GSSGSSGLPQVEAYSPSACSVRGGEELVLTGSNFLPDSKVVFIERGPDG KLQWEEEATVNRLQSNEVTLTLTVPEYSNKRVSRPVQVYFYVSNGRRK RSPTQSFRFLPVICKEE





PDB ID: 2YRP - all beta

Feature Vector Creation

SALHWR... \Rightarrow [S, A, L, H, W, R... \Rightarrow [S A, A L, L H, H... \Rightarrow [-0.867...

Spark-ML Pipeline: RegexTokenizer NGram Word2Vec

++++++			L	L			
0.7194245	alpha	alphal beta	coil foldType	l w	ords I	ngraml	featuresl
0.6904762 0.011904762 0.29761904 alpha [P, T, I, H, D, H [P T, T I, I H, H [-0.786925232194 0.8108108 0.0 0.1891892 alpha [G, A, M, E, P, E [G A, A M, M E, E [-0.835084856674 0.022900764 0.39694658 0.5801527 beta [I, V, N, G, E, E [I V, V N, N G, G [-0.684020347320 0.8021978 0.0 0.1978022 alpha [M, T, P, D, V, L [M T, T P, P D, D [-0.769814390980	0.7194245 0.4090909 0.015151516 0.7916667 0.028846154 0.6904762 0.8108108 0.022900764	94245 0.0 90909 0.0 51516 0.5808081 16667 0.0 46154 0.52884614 04762 0.011904762 08108 0.0 00764 0.39694658	0.28057554 alpha 0.59090906 alpha 0.4040404 beta 0.20833333 alpha 0.44230768 beta 0.29761904 alpha 0.1891892 alpha 0.5801527 beta	I[S, A, L, H, W, I] [S, H, L, K, S, I] [M, A, H, H, H, H, I] [M, G, S, S, H, I] [M, E, K, A, T, I] [P, T, I, H, D, I] [G, A, M, E, P, I] [I, V, N, G, E, I]	R [S A, A L, L K [S H, H L, L H [M A, A H, H H [M G, G S, S K [M E, E K, K H [P T, T I, I E [G A, A M, M E [I V, V N, N	H, H [-0.867888 K, K [-0.217258 H, H [-0.648501 S, S [-0.767287 A, A [-0.730871 H, H [-0.786925 E, E [-0.835084 G, G [-0.684026	36349164 38950821 12823101 74573129 13050851 52321949 18566740

https://github.com/sbl-sdsc/mmtf-spark/blob/master/src/main/java/edu/sdsc/mmtf/spark/ml/SequenceWord2Vector.java



Create Class Labels

	alphal	betal		foldTypel
Ī	0.7194245	0.0	0.28057554	•
	0.40909091	0.01	0.59090906	l alphal
-	0.0151515161	0.5808081	0.4040404	betal
-	0.79166671	0.0	0.20833333	alpha
-	0.0288461541	0.52884614	0.44230768	betal
-	0.69047621	0.011904762	0.29761904	alpha
-	0.8108108	0.01	0.1891892	alpha
-	0.0229007641	0.396946581	0.5801527	betal
-	0.8021978	0.01	0.1978022	alpha

The Dataset for Classification

We save this dataset as .parquet file:

```
data.write().mode("overwrite").format("parquet").save(filename);
```

class label

_	+-													L			
-	alphal	betal	coill	foldTypel				١	wordsl				ngraml	l		feat	uresl
l	0.7194245		0.28057554														
I	0.40909091	0.01	0.590909061	alphal[S,	Η,	L,	Κ,	S,	KI	[S H,	НL,	LK,	KI	[-0.	217258	895082	1
I	0.015151516	0.5808081	0.4040404	betal[M,	Α,	Η,	Η,	Н,	H	[M A,	ΑН,	ΗН,	H	[-0.	648501	282310	1
I	0.79166671	0.01	0.208333331	alpha [M,	G,	S,	S,	Н,	HI	[M G,	GS,	SS,	SI	[-0.	767287	457312	9।
١	0.0288461541	0.528846141	0.442307681	betal[M,	Ε,	Κ,	Α,	Τ,	KI	[M E,	ΕK,	ΚA,	A	[-0.	730871	305085	1
I	0.690476210	0.011904762	0.29761904	alpha [P,	Τ,	I,	Η,	D,	HI	[P T,	ΤI,	ΙH,	H	[-0.	786925	232194	9।
١	0.8108108	0.01	0.1891892														

beta|[I, V, N, G, E, E...|[I V, V N, N G, G...|[-0.6840203473201...|

alpha|[M, T, P, D, V, L...|[M T, T P, P D, D...|[-0.7698143909806...|

0.022900764 | 0.39694658 | 0.5801527 |

0.80219781

0.01 0.19780221

features

Demo 1

Create a dataset

Fit several classification models

Problem 1

- Change the code to a 3-state classification problem:
 - alpha, beta, alpha+beta
- Rerun the classification methods

Resources

- Machine Learning Library (MLlib) Guide
 - https://spark.apache.org/docs/latest/ml-guide.html
- Extracting, transforming and selecting features
 - https://spark.apache.org/docs/latest/ml-features.html
 - N-gram
 - https://spark.apache.org/docs/latest/ml-features.html#n-gram
 - Word2Vec example
 - https://spark.apache.org/docs/latest/ml-features.html#word2vec
 - Word2Vec model
 - https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec
- Classification and regression
 - https://spark.apache.org/docs/latest/ml-classification-regression.html
- Parquet files (columnar format)
 - https://spark.apache.org/docs/latest/sql-programming-guide.html#parquet-files

