

第11课 排序与CTR预估问题

七月在线 寒小阳

2016年11月19日

Outline

- Online advertising and click through rate prediction
- Data set and features
- Spark MLlib and the Pipeline API
- MLlib pipeline for Click Through Rate Prediction
- Random Forest/GBDT/FM/FFM/DNN

Online Advertising

Online advertising

THE WALL STREET JOURNAL

Subscribe Now | Sign In

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Search

What's News

SEC Appeals on the Slow Track

The regulator's use of its own tribunal has coincided with longer delays in the agency's handling of appeals, according to a Wall Street Journal analysis.

SEC to Retrench on SAC's Cohen

The Securities and Exchange Commission said it pared back its case against Steven A. Cohen and disclosed settlement talks with him, a reversal that reflects a major shift in the legal landscape since the government declared victory in its long pursuit of the hedge-fund billionaire's firm.



Valeant Deal With Walgreens Has Unusual Twist

Valeant's distribution deal with Walgreens has an unusual twist that allows the Canadian drugmaker to repurchase its drugs at the pharmacy without physically taking them back.



Bomb Kills Six American Soldiers In Afghanistan



China Unveils Economic Blueprint for 2016

China's leadership has mapped out an economic blueprint for next year that focuses on reducing industrial overcapacity, slashing costs for businesses, cutting unsold property inventory and fending off financial risks.

China's Economic Miracle Hits Tough Times

Front-Runners Trump and Clinton Face Off

Republican Donald Trump demanded an apology from Democrat Hillary Clinton for calling his rhetoric an Islamic State recruitment tool—and her campaign stood its ground.

238



CAPITAL JOURNAL

Regime Change, Good or Bad?



SpaceX Successfully Lands Rocket



Wreath Charity Has Ties to Its Supplier



San Bernardino Shooter's

Markets

U.S.	EUROPE	ASIA	FX	RATES	FUTURES
DJIA	17251.62	0.72%			
S&P 500	2021.15	0.78%			
Nasdaq	4968.92	0.93%			
Russell 2000	1127.74	0.60%			
DJ Total Mkt	20864.64	0.73%			
Global Dow	2301.28	-0.03%			

Dec 21 '15, 4:32 PM EST

Opinion

Another Time, Another Trump

By Dave Shiflett | Commentary

Treasury's Latest Inversion Failure

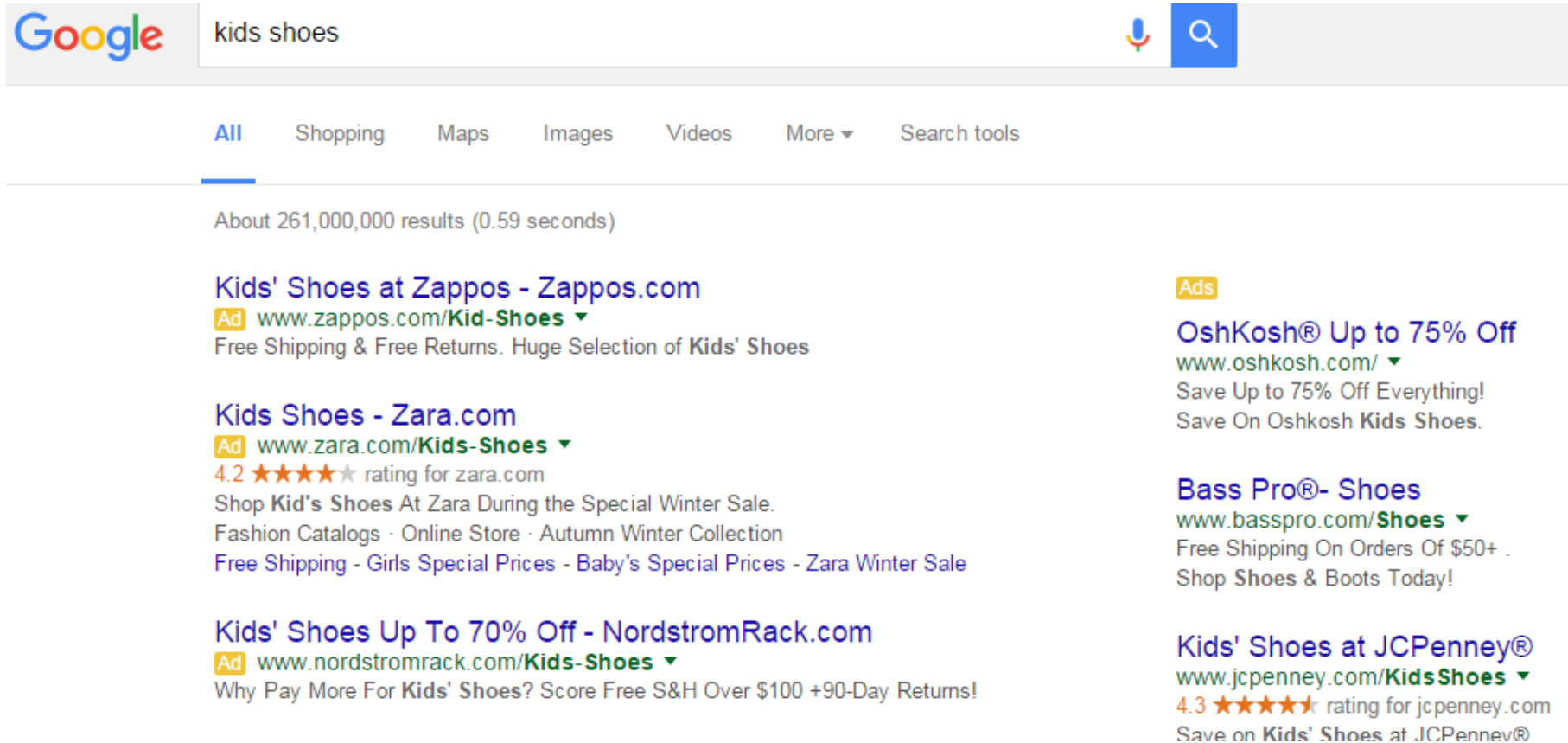
Review & Outlook

Let's Elect Hillary Now

By Bret Stephens | Global View



Online advertising on search engines



The screenshot shows a Google search for "kids shoes". The search bar at the top contains the text "kids shoes" and a blue search button with a magnifying glass icon. Below the search bar, there are tabs for "All", "Shopping", "Maps", "Images", "Videos", "More", and "Search tools". The "All" tab is selected. Below the tabs, it says "About 261,000,000 results (0.59 seconds)".

The search results are organized into two columns. The left column contains three search results, each with a title, a link, and a description. The right column contains three advertisements, each with a title, a link, and a description.

Search Results (Left Column):

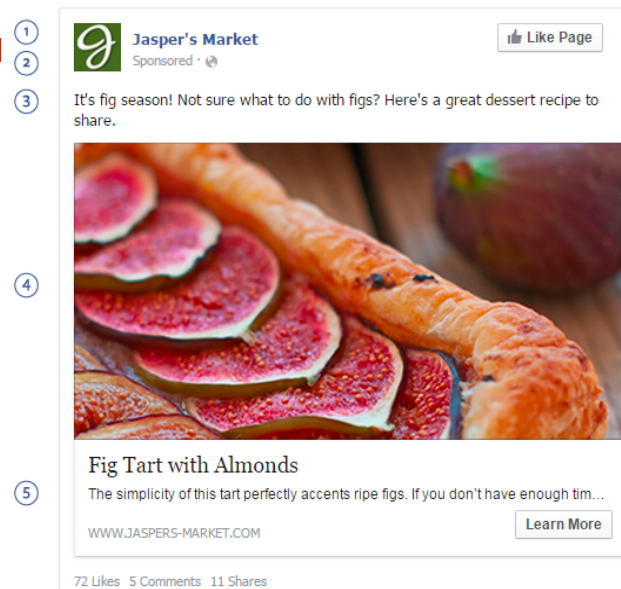
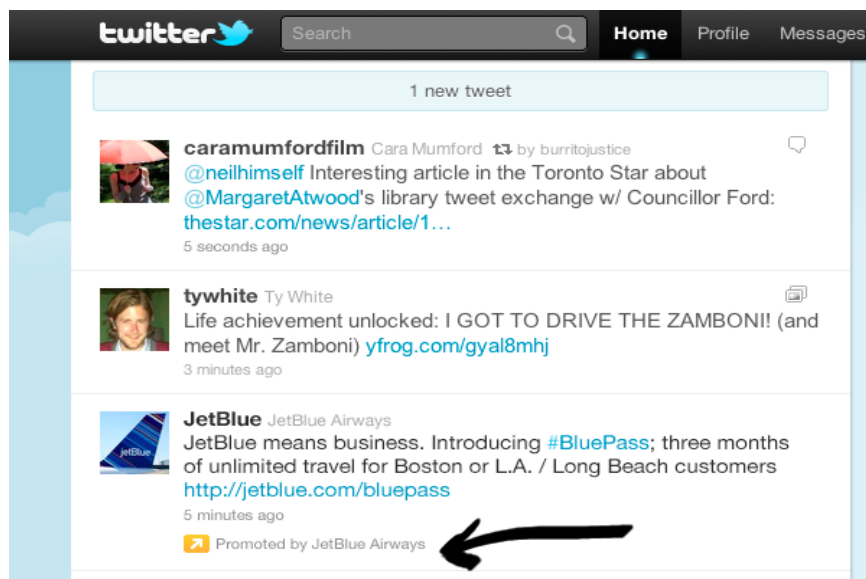
- Kids' Shoes at Zappos - Zappos.com**
Ad www.zappos.com/Kid-Shoes ▼
Free Shipping & Free Returns. Huge Selection of Kids' Shoes
- Kids Shoes - Zara.com**
Ad www.zara.com/Kids-Shoes ▼
4.2 ★★★★★ rating for zara.com
Shop Kid's Shoes At Zara During the Special Winter Sale.
Fashion Catalogs · Online Store · Autumn Winter Collection
Free Shipping - Girls Special Prices - Baby's Special Prices - Zara Winter Sale
- Kids' Shoes Up To 70% Off - NordstromRack.com**
Ad www.nordstromrack.com/Kids-Shoes ▼
Why Pay More For Kids' Shoes? Score Free S&H Over \$100 +90-Day Returns!

Advertisements (Right Column):

- OshKosh® Up to 75% Off**
www.oshkosh.com/ ▼
Save Up to 75% Off Everything!
Save On Oshkosh Kids Shoes.
- Bass Pro®- Shoes**
www.basspro.com/Shoes ▼
Free Shipping On Orders Of \$50+ .
Shop Shoes & Boots Today!
- Kids' Shoes at JCPenney®**
www.jcpenney.com/KidsShoes ▼
4.3 ★★★★★ rating for jcpenney.com
Save on Kids' Shoes at JCPenney®

Social Media advertising

- ❑ Facebook's "Sponsored Stories"
- ❑ LinkedIn's "Sponsored Updates"
- ❑ Twitter's "Promoted Tweets"



1. Social Information

When available, people will see if their friends have engaged with your business.

2. Business Name

The name of your business always shows prominently.

3. Text

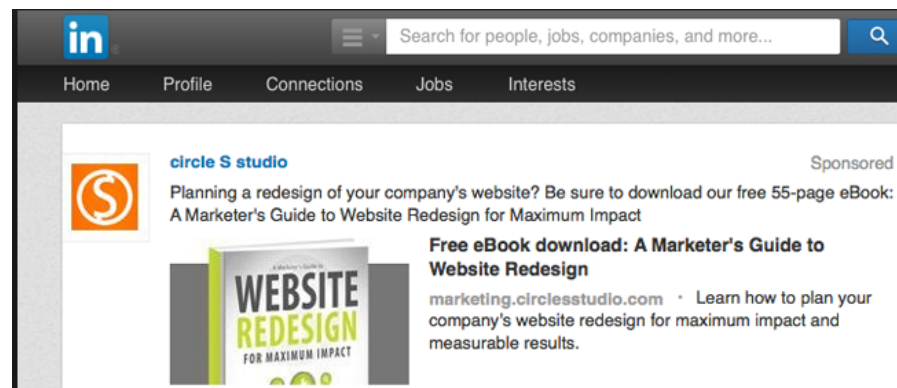
Grab interest with more info about what you're advertising.

4. Images and Videos

Compelling images and videos encourage your target audiences to engage.

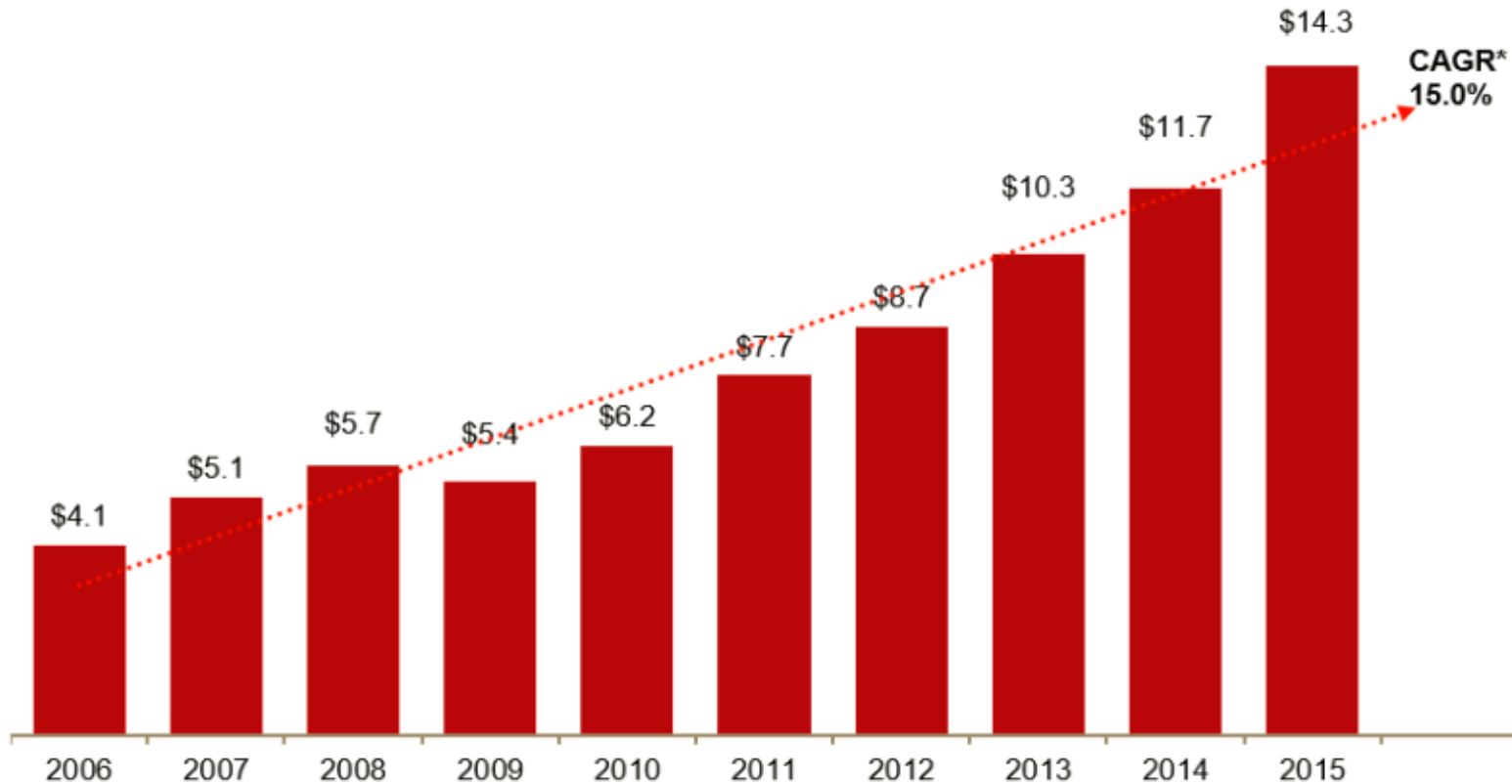
5. Call to Action (optional)

A customizable button encourages people to click.



Online advertising is big business

Second-quarter revenue 2006-2015 (\$ billions)



Source: IAB/PwC Internet Ad Revenue Report, HY 2015

* CAGR: Compound Annual Growth Rate

Types of online advertising

- ❑ **Retargeting** – Using cookies, track if a user left a webpage without making a purchase and retarget the user with ads from that site
- ❑ **Behavioral targeting** – Data related to user's online activity is collected from multiple websites, thus creating a detailed picture of the user's interests to deliver more targeted advertising
- ❑ **Contextual advertising** – Display ads related to the content of the webpage
- ❑ **Geo targeting** – Ads are presented based on the user's suspected geography

Online Advertising – The Players

Publishers:

- ❑ Earn revenue by displaying ads on their sites
- ❑ Google, Wall Street Journal, Twitter

Advertisers:

- ❑ Pay for their ads to be displayed on publisher sites
- ❑ Goal is to increase business via advertising

Matchmakers:

- ❑ Match publishers with advertisers
- ❑ Interaction with advertisers and publishers occurs real time



Revenue models from online advertising

- ❑ **CPM** (Cost-Per-Mille): is an inventory based pricing model. Is when the price is based on 1,000 impressions.
- ❑ **CPC** (Cost-Per-Click): Is a performance-based metric. This means the Publisher only gets paid when (and if) a user clicks on an ad
- ❑ **CPA** (Cost Per Action): Best deal of all for Advertisers in terms of risk because they only pay for media when it results in a sale

Click-through rate (CTR)

- ❑ Ratio of users who click on an ad to the number of total users who view the ad
- ❑ Used to measure the success of an online advertising campaign
- ❑ Very first online ad shown for AT&T on website HotWired has a 44% Click

-through rate

$$\text{CTR} = \frac{\text{Clicks}}{\text{Impressions}} \times 100 \%$$

- ❑ Today, typical click through rate is less than 1%
- ❑ In a pay-per-click setting, revenue can be maximized by choosing to display ads that have the maximum CTR, hence the problem of CTR Prediction.

Predict CTR – a Scalable Machine Learning success story

Predict conditional probability that the ad will be clicked by the user given the predictive features of ads

Predictive features are:

- Ad's historical performance
- Advertiser and ad content info
- Publisher info
- User Info (eg: search/ click history)

Dataset and features

About the dataset used

- ❑ Sample of the dataset used for the Display Advertising Challenge hosted by Kaggle:
<https://www.kaggle.com/c/criteo-display-ad-challenge/>
- ❑ Consists of a portion of Criteo's traffic over a period of 7 days.



Completed • \$16,000 • 718 teams

Display Advertising Challenge

Tue 24 Jun 2014 – Tue 23 Sep 2014 (15 months ago)

Data can be downloaded from Baidu cloud
链接: <https://pan.baidu.com/s/1qYVhaJq> 密码: 8fyn

Dataset Features

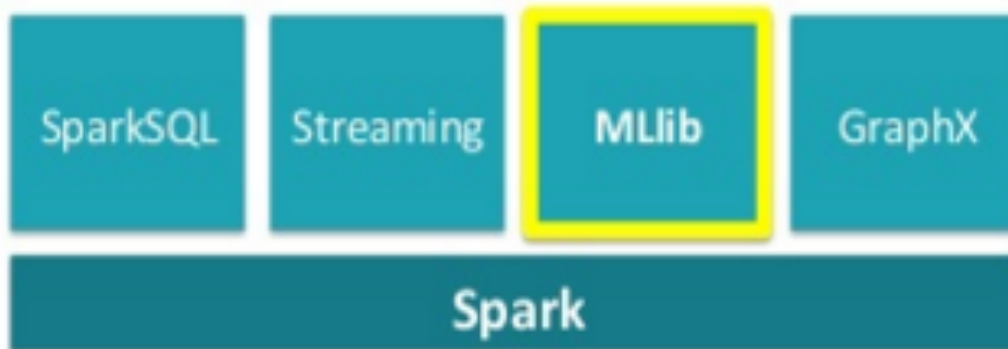
Data fields

- ❑ **Label** - Target variable that indicates if an ad was clicked (1) or not (0).
- ❑ **I1-I13** - A total of 13 columns of integer features (mostly count features).
- ❑ **C1-C26** - A total of 26 columns of categorical features. The values of these features have been hashed onto 32 bits for anonymization purposes.

Spark MLlib and Pipeline API

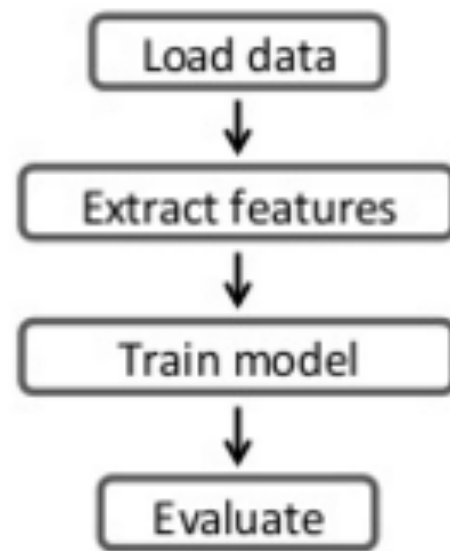
Spark MLlib

- Spark's machine learning (ML) library.
- Goal is to make ML scalable and easy
- Includes common learning algorithms and utilities, like classification, regression, clustering, collaborative filtering, dimensionality reduction
- Includes lower-level optimization primitives and higher level pipeline APIs
- Divided into two packages
 - Spark.mllib: original API built on top of RDDs
 - Spark.ml: provides higher level API built on top of DataFrames for constructing ML pipelines



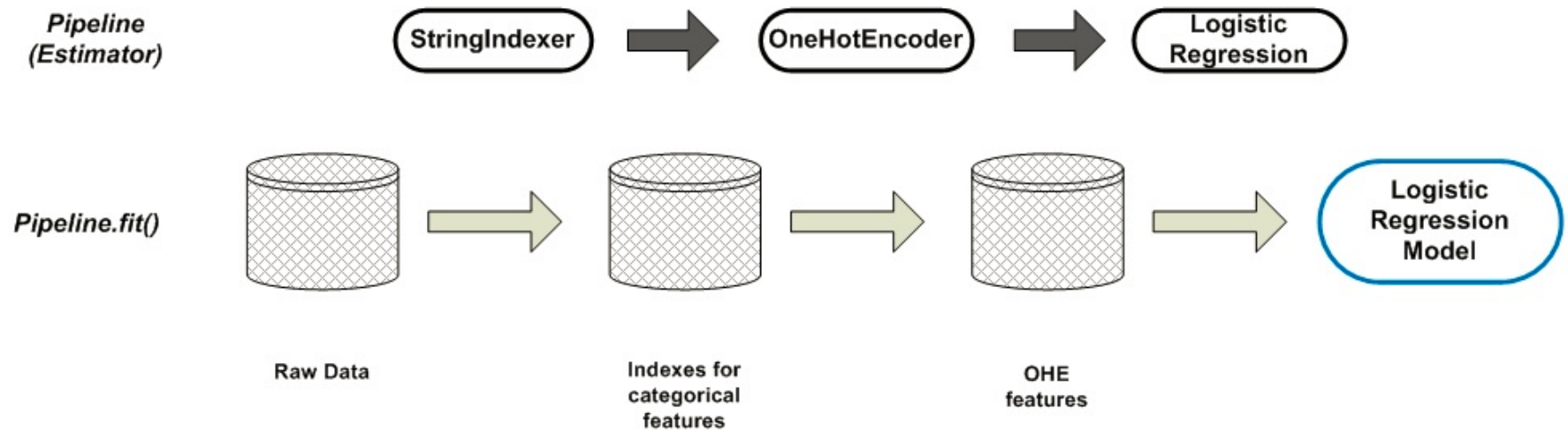
ML pipelines for complex ML workflows

- ❑ **DataFrame:** Equivalent to a relational table in Spark SQL
- ❑ **Transformer:** An algorithm that transforms one DataFrame into another
- ❑ **Estimator:** An algorithm that can be fit on a DataFrame to produce a Transformer
- ❑ **Pipeline:** Chains multiple transformers and estimators to produce a ML workflow
- ❑ **Parameter:** Common API for specifying parameters



MLlib Pipeline for Click Through Rate Prediction

Overview of ML pipeline for Click Through Rate Prediction



Step 1 - Parse Data into Spark SQL DataFrames

- ❑ Spark SQL converts a RDD of Row objects into a DataFrame
- ❑ Rows are constructed by passing key/value pairs where keys define the column names of the table
- ❑ The data types are inferred by looking at the data

```
def parseData(data, sqlContext):  
  
    #Split the csv file by comma and convert each line to a tuple.  
    parts = data.map(lambda l: l.split(",", -1))  
    features = parts.map(lambda p: Row(Label=(p[0]), IntFeature1=(p[1]), IntFeature2=(p[2]), IntFeature3=(p[3]),  
                                        IntFeature4=(p[4]), IntFeature5=(p[5]), IntFeature6=(p[6]), IntFeature7=(p[7]),  
                                        IntFeature8=(p[8]), IntFeature9=(p[9]), IntFeature10=(p[10]), IntFeature11=(p[11]),  
                                        IntFeature12=(p[12]), IntFeature13=(p[13]), CatFeature1=(p[14]), CatFeature2=(p[15]),  
                                        CatFeature3=(p[16]), CatFeature4=(p[17]), CatFeature5=(p[18]), CatFeature6=(p[19]),  
                                        CatFeature7=(p[20]), CatFeature8=(p[21]), CatFeature9=(p[22]), CatFeature10=(p[23]),  
                                        CatFeature11=(p[24]), CatFeature12=(p[25]), CatFeature13=(p[26])))  
  
    # Apply the schema to the RDD.  
    return sqlContext.createDataFrame(features)
```

Step 2 – Feature Transformer using StringIndexer

- ❑ Encodes a string column to a column of indices
- ❑ Indices are from 0 to max number of distinct string values, ordered by frequencies
- ❑ If column is numeric, cast it to string and index string values

id	category
0	a
1	b
2	c
3	a
4	a
5	c

id	category	categoryIndex
0	a	0.0
1	b	2.0
2	c	1.0
3	a	0.0
4	a	0.0
5	c	1.0

Step 3 - Feature Transformer using One Hot Encoding

- ❑ Maps a column of label indices to a column of binary vectors
- ❑ Allows algorithms which expect continuous features, to use categorical features

id	category	categoryIndex
0	a	0.0
1	b	2.0
2	c	1.0
3	a	0.0
4	a	0.0
5	c	1.0

id	category	categoryIndex	categoryVec
0	a	0.0	SparseVector(2, {0: 1.0})
1	b	2.0	SparseVector(2, {})
2	c	1.0	SparseVector(2, {1: 1.0})
3	a	0.0	SparseVector(2, {0: 1.0})
4	a	0.0	SparseVector(2, {0: 1.0})
5	c	1.0	SparseVector(2, {1: 1.0})

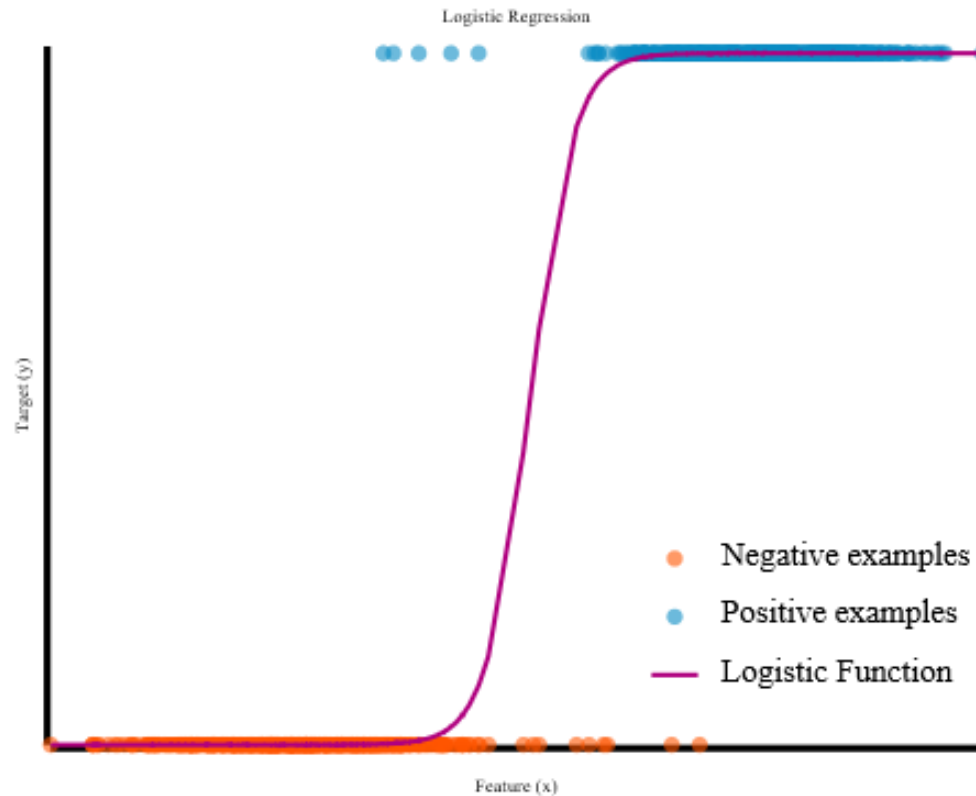
Step 4 – Feature Selector using Vector Assembler

- ❑ Transformer that combines a given list of columns into a single vector column
- ❑ Useful for combining raw features and features generated by different feature transformers into a single feature vector

id	hour	mobile	userFeatures	clicked
0	18	1.0	[0.0, 10.0, 0.5]	1.0

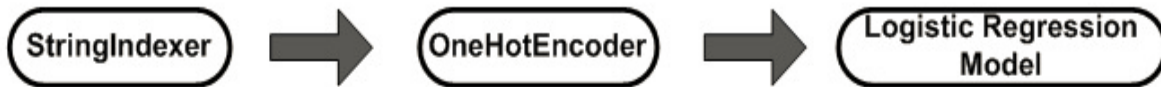
id	hour	mobile	userFeatures	clicked	features
0	18	1.0	[0.0, 10.0, 0.5]	1.0	[18.0, 1.0, 0.0, 10.0, 0.5]

Step 5 – Train a model using Estimator Logistic Regression

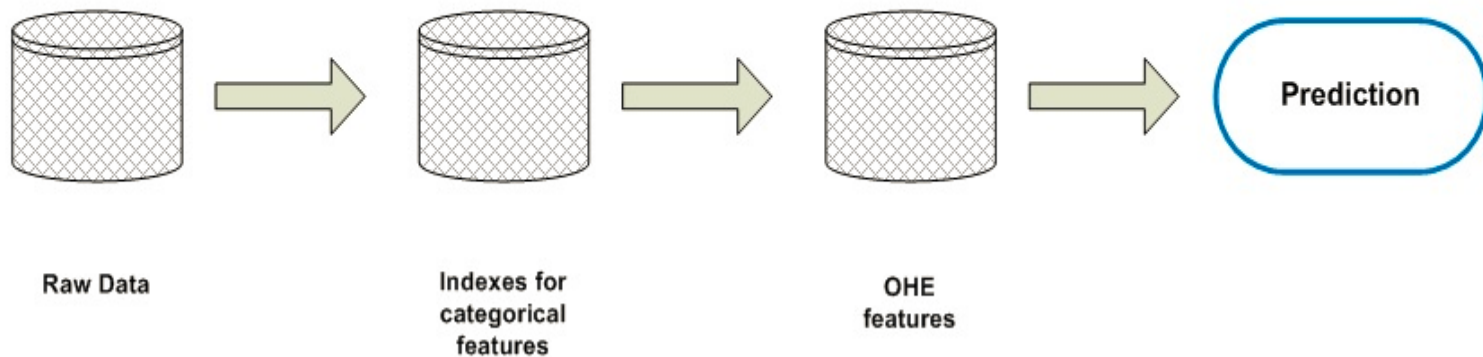


Apply the pipeline to make predictions

PipelineModel
(Transformer)



PipelineModel
.transform()



Demo

Parameter Tuning using CrossValidator or TrainValidationSplit

```
# Try Cross Validator
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator()
paramGrid = ParamGridBuilder().addGrid(lr.maxIter, [0, 1]).build()
numFolds=2

crossval = CrossValidator(
    estimator=pipeline,
    estimatorParamMaps=paramGrid,
    evaluator=evaluator,
    numFolds=numFolds)

cvModel = crossval.fit(dfTrain)
evaluator.evaluate(cvModel.transform(dfTest))
```

Some alternatives:

- ☐ Use Hashed features instead of OHE
- ☐ Use Log loss evaluation or ROC to evaluate Logistic Regression
- ☐ Perform feature selection
- ☐ Use Naïve Bayes or other binary classification algorithms

单机上的LR建模
请参见所给ipython notebook

Random Forest GBDT FM (factorization machine)

请参见比赛

<https://www.kaggle.com/c/avazu-ctr-prediction>

Rank 2nd Owen Zhang的解法

: <https://github.com/owenzhang/kaggle-avazu>

FFM (field-aware factorization machine)

参照课程所给文档与代码



Google Wide && Deep model

说明:

https://www.tensorflow.org/versions/r0.10/tutorials/wide_and_deep/index.html

代码:

https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/learn/wide_n_deep_tutorial.py

About Data

需要大规模数据做实验/学习的同学，可以在Cretio实验数据下载1TB的CTR预估所需数据。

<http://labs.criteo.com/downloads/download-terabyte-click-logs>

感谢大家！

恳请大家批评指正！