Using H2O R Models @ System1

Jeff Roach

01/23/2018

SYSTEM1

Campaign Optimization Problem

- Large dataset
  - >200M rows of session level data
  - Mixed effect models are typically used for granular data
- Train model in < 1 hour
  - Simple mixed effect model
    - 6 mins for 2M rows
    - Est. ~10 hours at 200M
- Need to easily expand model with more features
  - Time
  - User context
  - Other models

Benchmarking

- Check out Szilard's benchmarking
  - https://github.com/szilard/benchm-ml
  - Python, R, H2O, Spark, Vowpal Wabbit
- H2o is a clear winner in most categories
  - Training speed and memory usage
  - Especially with large data
- 10M row logistic regression model
  - R - 90s, 5GB
  - Python - crash/360s, 60/250GB
  - H2O - 5s, 3GB

R H2O Code

```
library(h2o)

h2o.init(max_mem_size="20g", nthreads=-1)  # Initialize h2o cluster locally using maximum number of threads

dx_train <- as.h2o(df_train)  # Make into h2o object
dx_test <- as.h2o(df_test)

# Train model
 model <- h2o.gbm(x = c("device_type", "country_code", "market", "partner", "category",
                        "day", "hour", "hourc", "hours", "hourcs", "hourcc", "hourss"),
                  y = "ctr",
                  training_frame = dx_train,
                  weights_column = 'sessions', distribution = "gaussian",
                  ntrees = 1000, max_depth = 9, learn_rate = 0.01,
                  min_rows = 1, sample_rate = 0.8, col_sample_rate = 1.0, nbins = 100)

h2o.performance(model, dx_test)  # Check performance on a holdout set

predictions <- h2o.predict(model, dx_test)  # Make predictions
```

H2O Pros/Cons

- Pros
  - Fast (utilizes multiprocessing)
  - Resource efficient (utilizes distributed technology)
  - Easy to use
  - Python and R
  - Easily save model (POJO/MOJO)
  - Can distribute to multiple boxes for increased speedups
- Cons
  - Less complex models (currently)
    - No mixed effect models (use random forest instead?)
  - Predicting straight from model in prod requires Java code
    - Put your model in a np.memmap lookup table