Łukasz Grala

lukasz@tidk.pl | lukasz.grala@cs.put.poznan.pl

Analiza sentymentu przy użyciu bibliotek Microsoft

Data Scientist as a Service

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

Łukasz Grala

- Senior architekt rozwiązań Platformy Danych & Business Intelligence & Zaawansowanej Analityki w TIDK
- Twórca „Data Scientist as as Service"
- Certyfikowany trener Microsoft i wykładowca na wyższych uczelniach
- Autor zaawansowanych szkoleń i warsztatów, oraz licznych publikacji i webcastów
- Od 2010 roku wyróżniany nagrodą Microsoft Data Platform MVP
- Doktorant Politechnika Poznańska – Wydział Informatyki (obszar bazy danych, eksploracja danych, uczenie maszynowe)
- Prelegent na licznych konferencjach w kraju i na świecie
- Posiada liczne certyfikaty (MCT, MCSE, MCSA, MCITP,...)
- Członek zarządu Polskiego Towarzystwa Informatycznego Oddział Wielkopolski
- Członek i lider Polish SQL Server User Group (PLSSUG)
- Pasjonat analizy, przechowywania i przetwarzania danych, miłośnik Jazzu

email lukasz@tidk.pl - lukasz.grala@cs.put.poznan.pl blog: **~~grala.it~~**

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

PTI POLSKIE TOWARZYSTWO INFORMATYCZNE ODDZIAŁ WIELKOPOLSKI

Parallelized Algorithms

Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

- Subsample (observations & variables)
- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression Linear, Logistic and GLM

Simulation

- Monte Carlo
- Parallel Random Number Generation

Cluster Analysis

- K-Means

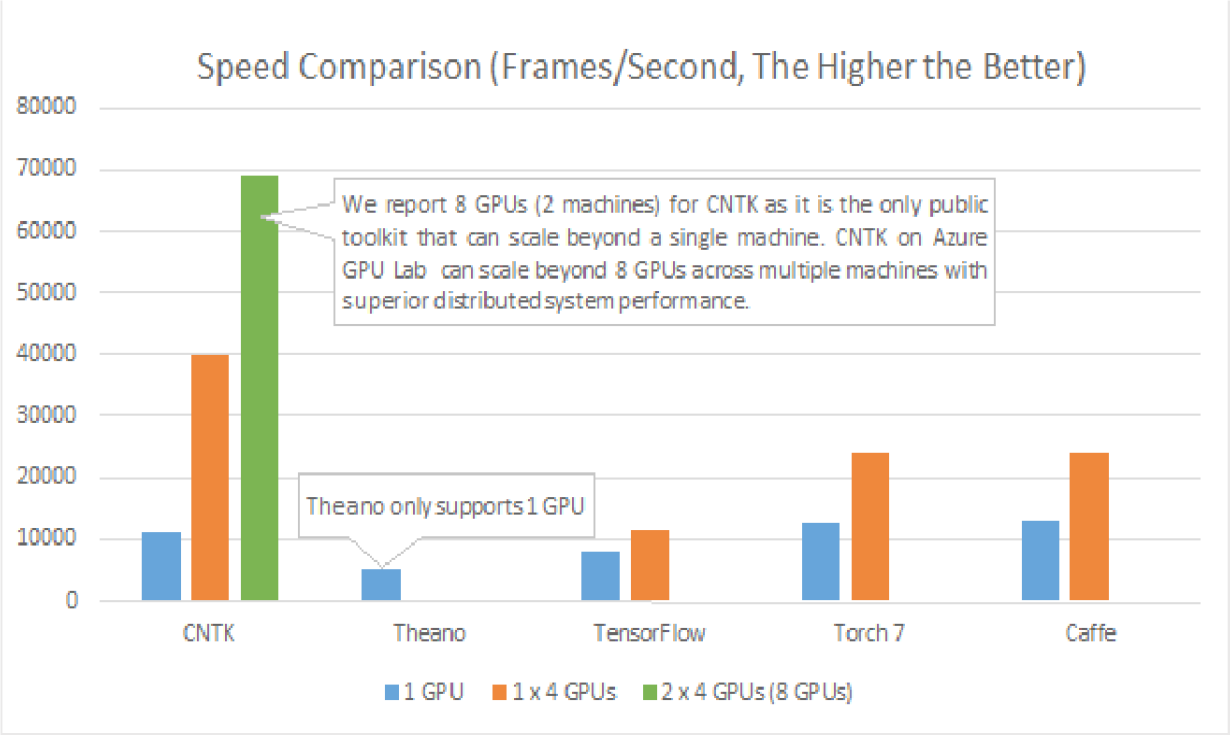
Classification

- Decision Trees
- Decision Forests
- Stochastic Gradient Boosted Decision Trees

Combination

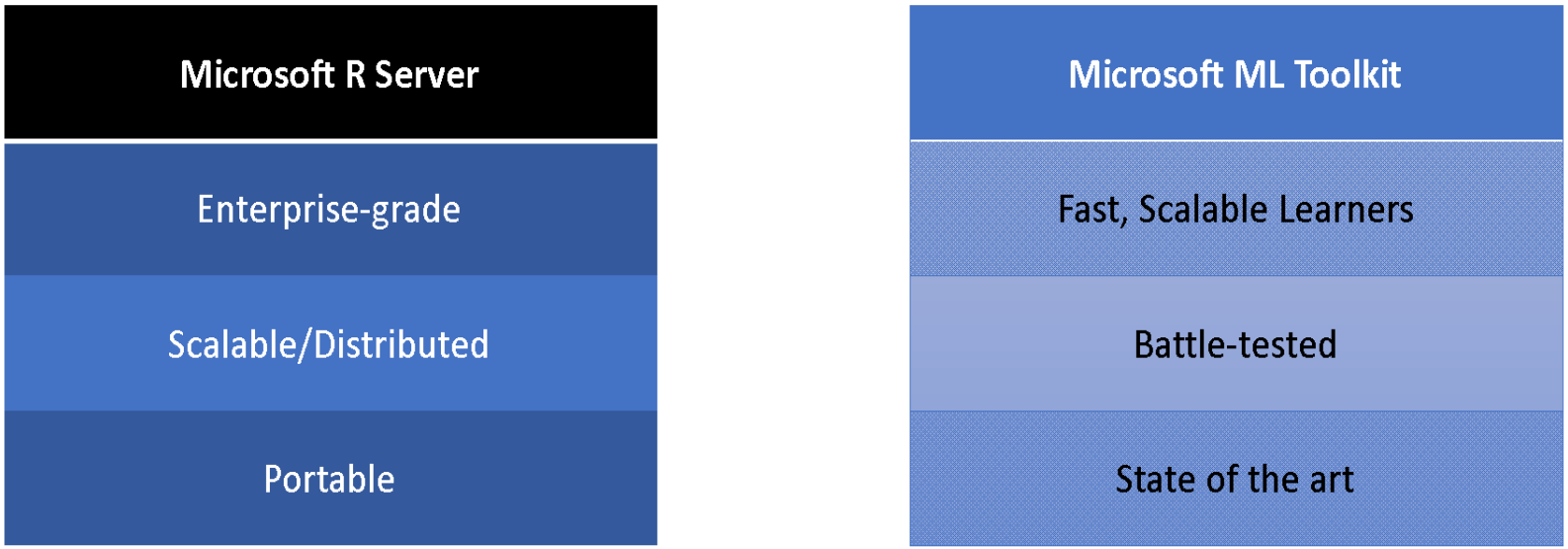
- Using Revolution rxDataStep and rxExec functions to combine open source R with Revolution R
- PEMA API

Benchmark CNTK



https://github.com/Alexey-Kamenev/Benchmarks
https://github.com/Microsoft/CNTK

MicrosoftML = Scalable R + World Class ML



Learners

Algorithms	Strengths
rxFastLinear	Fast, accurate linear learner with auto L1 & L2
rxLogisticRegression	Logistic Regression with L1 & L2
rxFastTree	Boosted Decision tree from Bing. Competitive wth XGBoost. Most accurate learner for most cases
rxFastForest	Random Forest
rxNeuralNet	GPU accelereted Net# DNNs with Convolutions
rxOneClassSvm	Anomaly or unbalanced binary classification

Algorithms and Transforms in MicrosoftML

Machine learning transforms

- concat: Transformation to create a single vector-valued column from multiple columns.
- categorical: Create indicator vector using categorical transform with dictionary.
- categoricalHash: Converts the categorical value into an indicator array by hashing.
- featurizeText: Produces a bag of counts of sequences of consecutive words, called n-grams, from a given corpus of text. It offers language detection, tokenization, stopwords removing, text normalization and feature generation.
- getSentiment: Scores natural language text and creates a column that contains probabilities that the sentiments in the text are positive.
- ngram: allows defining arguments for count-based and hash-based feature extraction.
- selectFeatures: Selects features from the specified variables using a specified mode.
- loadImage: Loads image data.
- resizeImage: Resizes an image to a specified dimension using a specified resizing method.
- extractPixels: Extracts the pixel values from an image.
- featurizeImage: Featurizes an image using a pre-trained deep neural network model.

Algorithms and Transforms in MicrosoftML

Machine learning algorithms

- rxFastTrees: An implementation of FastRank, an efficient implementation of the MART gradient boosting algorithm.
- rxFastForest: A random forest and Quantile regression forest implementation using rxFastTrees.
- rxLogisticRegression: Logistic regression using L-BFGS.
- rxOneClassSvm: One class support vector machines.
- rxNeuralNet: Binary, multi-class, and regression neural net.
- rxFastLinear: Stochastic dual coordinate ascent optimization for linear binary classification and regression.
- rxEnsemble: trains a number of models of various kinds to obtain better predictive performance than could be obtained from a single model.

Scoring and training and model summary

- rxPredict.mlModel: Runs the scoring library either from SQL Server, using the stored procedure, or from R code enabling real-time scoring to provide much faster prediction performance.
- rxFeaturize: Transforms data from an input data set to an output data set.
- mlModel Provides a summary of a Microsoft R Machine Learning model.

Algorithms and Transforms in MicrosoftML

Loss functions for classification and regression.

- expLoss: Specifications for exponential classification loss function.
- logLoss: Specifications for log classification loss function.
- hingeLoss: Specifications for hinge classification loss function.
- smoothHingeLoss: Specifications for smooth hinge classification loss function.
- poissonLoss: Specifications for poisson regression loss function.
- squaredLoss: Specifications for squared regression loss function.

Functions for feature selection.

- minCount: Specification for feature selection in count mode.
- mutualInformation: Specification for feature selection in mutual information mode.

Algorithms and Transforms in MicrosoftML

Functions for ensemble modeling.

- fastTrees: Creates a list containing the function name and arguments to train a Fast Tree model with rxEnsemble.
- fastForest: Creates a list containing the function name and arguments to train a Fast Forest model with rxEnsemble.
- fastLinear: Creates a list containing the function name and arguments to train a Fast Linear model with rxEnsemble.
- logisticRegression: Creates a list containing the function name and arguments to train a Logistic Regression model with rxEnsemble.
- oneClassSvmCreates a list containing the function name and arguments to train a OneClassSvm model with rxEnsemble.

Functions for neural networks.

- optimizer Specifies optimization algorithms for the rxNeuralNet machine learning algorithm.

Learners - Scalability

- Streaming (not RAM bound)
- Billions of features
- Multi-proc
- GPU acceleration for DNNs
- Distributed on Hadoop/Spark via Ensambling

```
rxSetComputeContext(RxSpark())

model <- rxFastLinear(
    Rating ~ Features, data = dataTrain, mlTransforms = mlTransforms,
    ensemble = ensembleControl(modelCount = 100)
)
```

Image Featurization

Convolutional DNNs with GPU

Pre-trained Models

- ResNet18
- ResNet 50
- ResNet 101
- AlexNet

Image Featurization

- Image similarity search:
  - Product Catalog search
- Classification
  - Plankton Monitoring
  - Galaxy Classification
  - Retina Pathology detection
- Anomaly Detection
  - Defects detection in manufacturing

Text Analytics Scenarios

Text Classification

- Email or support ticket routing
- Customer call triage
- Detect illegal trading activity

Sentiment Analysis

- Social media monitoring
- Call center support

Sentiment Analysis

- Pre-trained model
- Cognitive Service Parity
- Uses DNN Embedding
- Domain Adaptation

```
sentiment <- rxFeaturize(data = dataTrain
    mlTransforms = list(
        getSentiment(vars = "Text")
    )
)
```

Dataset

Large Movie Review Dataset

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.

Large Movie Review Dataset v1.0

When using this dataset, please cite our ACL 2011 paper [bib].

Contact

For comments or questions on the dataset please contact Andrew Maas. As you publish papers using the dataset please notify us so we can post a link on this page.

Publications Using the Dataset

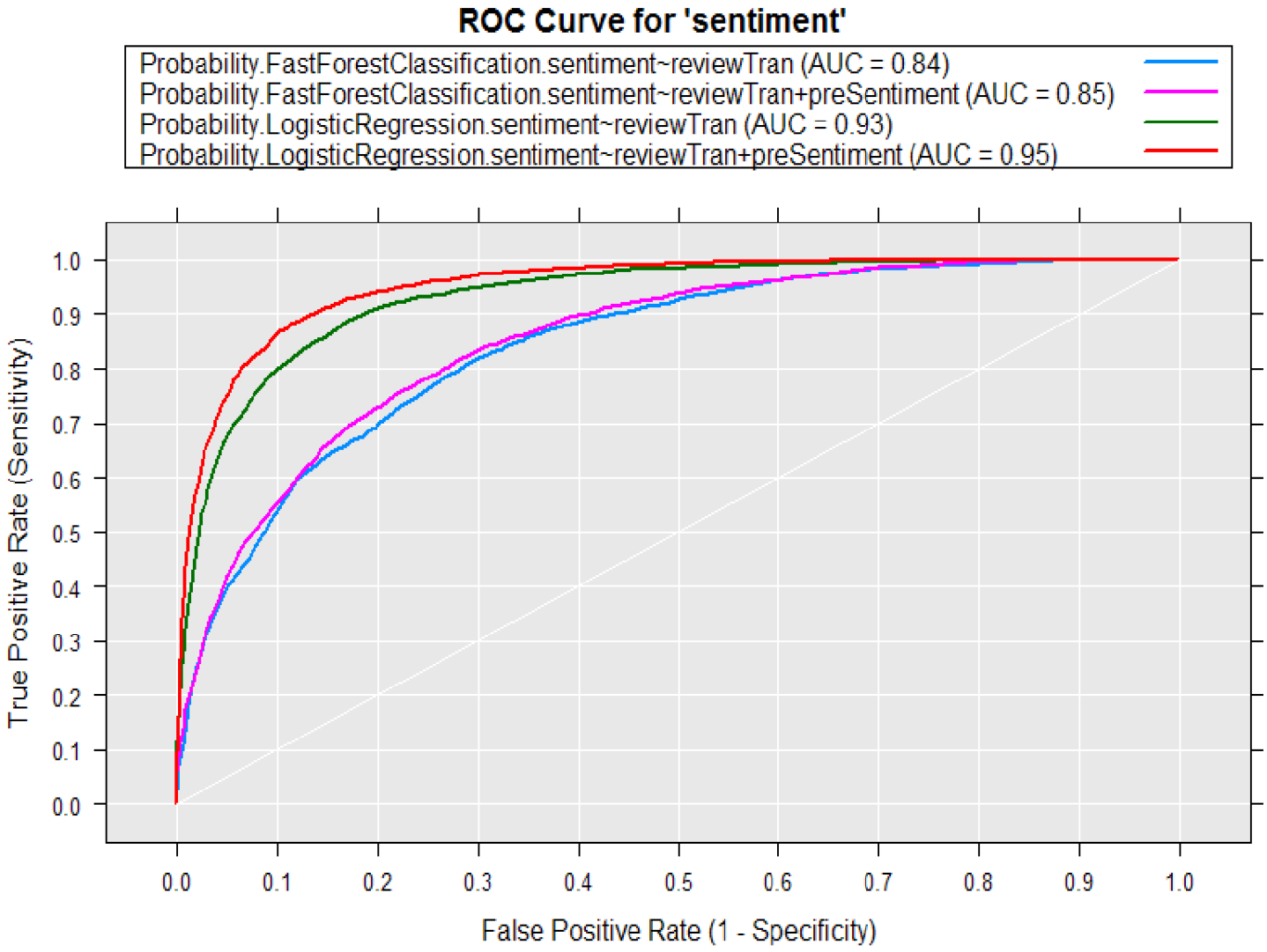
Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

http://ai.stanford.edu/~amaas/data/sentiment/

Experiment

- rxLogisticRegression
- rxLogisticRegression + getSentiment
- rxFastForest
- rxFastForest + getSentiment

ROC

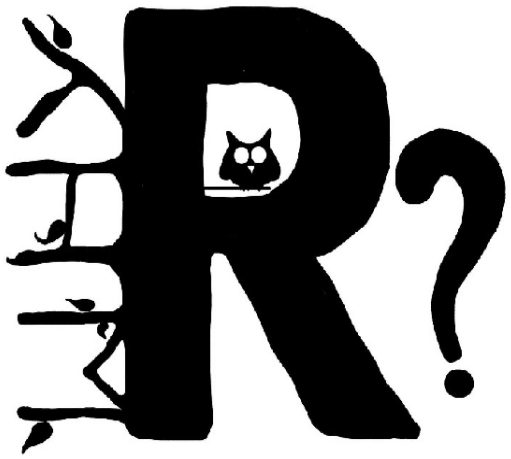

ROC Curve for 'sentiment'
Probability.FastForestClassification.sentiment~reviewTran (AUC = 0.84)
Probability.FastForestClassification.sentiment~reviewTran+preSentiment (AUC = 0.85)
Probability.LogisticRegression.sentiment~reviewTran (AUC = 0.93)
Probability.LogisticRegression.sentiment~reviewTran+preSentiment (AUC = 0.95)
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Compare features by product

	Microsoft R Open	Microsoft R Client	Microsoft R Server
Storage	Memory bound[1]	Memory bound[1] & operates on large volumes when connected to R Server.	Data chunking across multiple disks. Operates on bigger volumes & factors.
Speed of Analysis	Multithreaded via MKL[2] for non-RevoScaleR functions.	Multithreaded via MKL[2] for non-RevoScaleR functions, but only up to 2 threads for ScaleR functions with a local compute context.	Full parallel threading & processing for RevoScaleR functions as well as for non-RevoScaleR functions (via MKL[2]) in both local and remote compute contexts.
Analytic Breadth & Depth	Open source packages only.	Open source R packages plus proprietary packages.	Open source R packages plus proprietary packages with support for parallelization and distributed workloads.
Operationalization of R Analytics	Not available	Not available	Includes the instant deployment and easy consumption of R analytics, interactive remote code execution, speedy realtime scoring, scalability, and enterprise-grade security.

[1] Memory bound because product can only process datasets that fit into the available memory.

[2] Because the Intel Math Kernel Library (MKL) is included in MRO, the performance of a generic R solution is generally better. MKL replaces the standard R implementations of Basic Linear Algebra Subroutines (BLAS) and the LAPACK library with multithreaded versions. As a result, calls to those low-level routines tend to execute faster on Microsoft R than on a conventional installation of R.



Question?

lukasz@tidk.pl lukasz.grala@cs.put.poznan.pl

Data Scientist as a Service

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk