Machine Learning HW4

Chen Li

27607798

Chosen Software:

Specific-Domain: Word2Vec(gensim)

mathematical or domain oriented problem solving environment: R

API callable from a general purpose programing language: TensorFlow

1. Evaluate (in words) each of three machine learning software packages/systems from these points of view:

1) What functionalities within machine learning (e.g. book chapters in Bishop) does it provide? What does it not provide?

Tensor flow

TensorFlow mainly focus on machine learning and deep neural networks research. It supports popular deep learning techniques such as **feed-forward neural network (multilayer perceptron), convolutional neural network and recursive neural network**. Different from traditional machine learning library such as Scikit-learn, provides independent API for different machine learning algorithm such as Linear Regression or Logistic Regression, TensorFlow takes a more flexible approach. Users only need to specify the loss function and optimization approach they needed. This makes TensorFlow support **almost every supervised machine learning functionalities** as I know, because most of supervised machine learning algorithm is basically optimize the loss function. However, implement **unsupervised learning algorithm** such as **kmeans** is a little inconvenient for TensorFlow.

Word2Vec(gensim)

Word2vec is a **two-layer neural net** that processes text. It transfers text corpus to numeric vectors. It doesn’t provide traditional NLP algorithms or other machine learning functionalities.

R

R has a very large machine learning library with various packages. Almost every machine learning(perceptron, logistic regression, MLP etc) algorithms are covered in these packages, also deep learning algorithms (h2o).

2) How scalable is it?

Tensor flow

The scalability is a highlighted feature for TensorFlow. TensorFlow runs on CPUs or GPUs, and on desktop, server, or mobile computing platforms such as Docker. With first-class support for threads, queues, and asynchronous computation, TensorFlow allows you to make the most of your available hardware.

Word2Vec(gensim)

Gensim support distributed computing. In the context of *gensim*, computing nodes are computers identified by their IP address/port, and communication happens over TCP/IP. The whole collection of available machines is called a *cluster*. A physical machine is called a node.

R

The major distributed systems (e.g. Hadoop) has interface for R. That means R do support deal with large-scale data with the help of distributed system.

3) What other software is needed to make it functional in an ML project?

Tensor flow

No. TensorFlow is highly self-contained, you can write your own function on it using python or c++ if the provided functions don’t cover your needs as long as you have python or c++ on your computer.

Word2Vec(gensim)

Word2Vec has several implementations. Gensim is its python implementation(it also includes other topic model), so it needs python to be installed on your computer. If you want to use gensim in distributed environment, you should install the python package Pyro (PYthon Remote Objects).

R

R and its packages are sufficient for machine learning project. However, it is not very convenient to deal with text processing using R. But it needs general-purpose language to build machine learning application.

4) For what application domains would it be suitable? Unsuitable?

Tensor flow

Suitable for most of the machine learning algorithm, especially in deep learning. Actually it is the best machine learning framework I've ever seen, but it maybe a little bit confusing for users who get accustom to the traditional API but now they have to provide the loss function and optimize function themselves. Also, it maybe unsuitable if you want to use other optimizers rather than gradient approach in your algorithm because Tensor flow only covers limited optimization approaches.

Word2Vec(gensim)

Word2Vec is suitable for natural language processing especially in word perspective. It also works for information retrieval and sentiment analysis. The output of Word2Vec can be used in other machine learning algorithm. It may be unsuitable for Word2Vec be used in context other than NLP as I can see.

R

R is suitable for doing data analysis part of machine learning project, not suitable for the infrastructure part.

5) Explain how each is available (eg. free download site etc.) to anyone at UCI (including you, your group, the TAs, and the instructor).

Tensor flow

Tensor flow is open source under Apache 2.0 License. You can get it from **https://github.com/tensorflow/tensorflow**

Word2Vec(gensim)

Gensim is a free software, you can get it in https://radimrehurek.com/gensim/index.html

R

R language is a free software, you can get it in
https://www.r-project.org

2. List the documents you read for each of the three packages, in bibliography format.

Tensor flow

1. https://github.com/tensorflow/tensorflow
2. http://tensorflow.org
3. Mart´ın Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems
4. https://github.com/nlintz/TensorFlow-Tutorials

Word2Vec(gensim)

1. https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-2-word-vectors
2. https://radimrehurek.com/gensim/tutorial.html
3. http://deeplearning4j.org/word2vec.html

R

1. https://www.r-project.org
2. http://gekkoquant.com/2012/05/26/neural-networks-with-r-simple-example/
3. http://www.r-bloggers.com/things-to-try-after-user-part-1-deep-learning-with-h2o/
4. https://cran.r-project.org/doc/manuals/r-release/R-intro.html