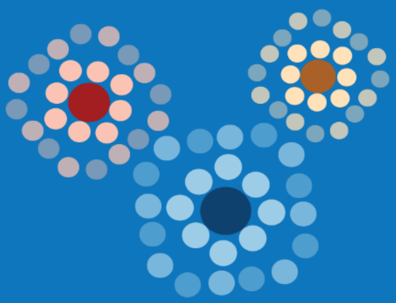
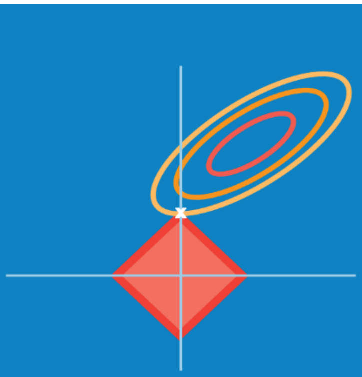
Distributed Machine Learning



MB-KMeans

Minibatch K-Means

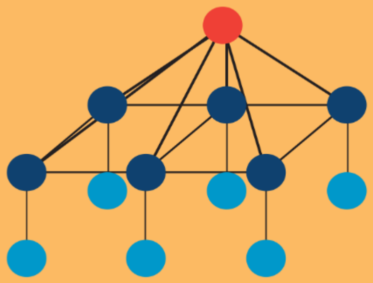


Lasso

Linear Regression with l1 Regularizer

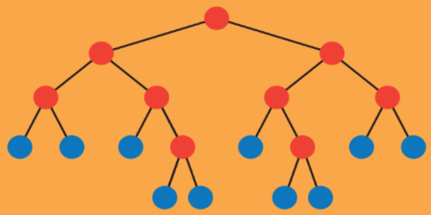
MLR

Multiclass Logistic Regression



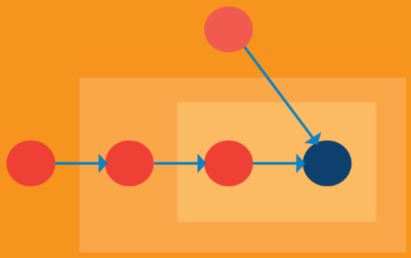
CRFs

Conditional Random Fields



RandomForest

Decision Tree + Ensemble

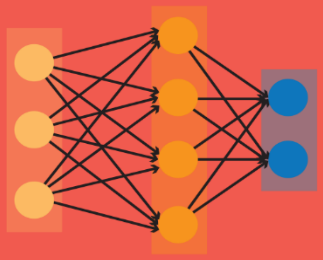


LDA

Latent Dirichlet Allocation

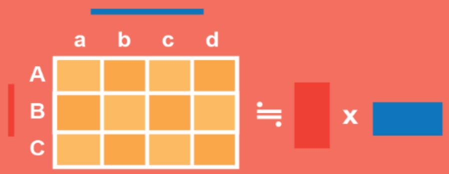
SVM

Support Vector Machine



NN

Neural Networks



FM

Factorization Machine

Guideline

Below are descriptions of term projects. Nine machine learning tasks are given, and all of them are popular algorithms that capable of dealing with large scale data analysis. You are encouraged to implement a distributed version among one of them on Harp and evaluate its performance.

To focus on the distributed computing aspects of the task, existing open source codes of these algorithms can be used in your project. You do not have to create new sequential algorithm itself. Related code may be found in reference papers, in popular machine learning libraries, such as sklearn, spark mlib, petuum, etc, or at github.

Algorithms

- MB-KMeans: minibatch K-Means
  - goal:Need a fast clustering tool suitable for large scale dataset.
  - paper reference: [Sculley 2010]
  - dataset: RCV1-v2

- Lasso: linear regression with l1 regularizer
  - goal:Need a high performance predictor for click-through data prediction.
  - paper reference: [McMahan 2013]
  - dataset: Criteo Click-through Data, Tecent Click-through Data

- MLR: Multiclass Logistic Regression
  - goal:Need a multiclass logistic regression classifier for large scale text classification problems.
  - paper reference: [Genkin 2007]
  - dataset: RCV1-v2, OHSUMED, LSHTC

- CRFs: conditional random fields
  - goal:Need a linear chain CRFs for POS tagger or NER task.
  - paper reference: [Lavergne 2010]
  - dataset: CoNLL03, Genia

- RandomForest: decision tree + ensemble
  - goal:Need a scalable decision tree ensemble for large nonlinear classification problem.
  - paper reference: [Genuer 2015] [Wright 2015]
  - dataset: Airline , PAMAP2

- LDA: latent dirichlet allocation
  - goal:Need a scalabe topic modeling tool for large scale text data.
  - paper reference: [Zhai 2012]
  - dataset: wikipedia, pubmed

- SVM: support vector machine
  - goal:Need a svm classifier for large image classification problem.
  - paper reference: [Lin 2011]
  - dataset: MNIST, PASCAL VOL, ImageNet

- NN: neural networks
  - goal:Need a multilayer neural network to get better digit recognition performance.
  - paper reference: [Ciresan 2010]
  - dataset: MNIST + deform(http://leon.bottou.org/projects/infimnist)

- FM: Factorization Machine(*included in Harp)
  - goal:Need a scalable matrix factorization tool for large collaborative filtering problem.
  - paper reference: [Rendle 2010]
  - dataset: YahooMusic, Netflix

Ratings of the Algorithms

The following ratings are an estimation of the effort on individual tasks.

Algorithm	**To understand the algorithm**	**To port existing code**	**Overall**
MB-KMeans	1	1	1
Lasso	2	2	2
MLR	2	2	2
RandomForest	3	2	2.5
FM	3	2	2.5
SVM	3	3	3
NN	4	4	4
CRFs	4	3	3.5
LDA	4	2(for Variational Bayesian)	3

Datasets:

Classification and regression

- RCV1-v2 [site]
- PAMAP2 [site]

- OHSUMED [site]
- Intrusion (KDDCUP2009) [site]
- Airline (Data Expo 09) [site]
- LSHTC [site]

Sequential labeling

- CoNLL03 [site]
- Genia [site]

Collaborative filtering

- YahooMusic (KDDCUP2011) [site]
- Netflix [site]

Click Through Rate (CTR) prediction

- Criteo Click-through Data (Kaggle Competition) [site]
- Tecent Click-through Data (KDDCUP2012) [site]

Image processing

- MNIST [site]
- PASCAL VOL [site]
- ImageNet [site]

Text mining

- Wikipedia [site]
- Pubmed [site]

References

[Sculley 2010] Sculley, D., 2010 Web-scale k-means clustering, in: Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 1177–1178.

[Rendle 2010] Rendle, S., 2010. Factorization machines, in: 2010 IEEE International Conference on Data Mining. IEEE, pp. 995-1000.

[Lavergne 2010] Lavergne, T., Cappé, O., Yvon, F., 2010. Practical very large scale CRFs, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 504–513.

[Ciresan 2010] Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J., 2010. Deep, big, simple neural nets for handwritten digit recognition. Neural computation 22, 3207–3220.

[Zhai 2012] Zhai, K., Boyd-Graber, J., Asadi, N., Alkhouja, M.L., 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce, in: Proceedings of the 21st International Conference on World Wide Web. ACM, pp. 879–888.

[Genkin 2007] Genkin, A., Lewis, D.D., Madigan, D., 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics 49, 291–304.

[McMahan 2013] McMahan, H.B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., others, 2013. Ad click prediction: a view from the trenches, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1222–1230.

[Lin 2011] Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., Huang, T., 2011. Large-scale image classification: fast feature extraction and svm training, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 1689–1696.

[Genuer 2015] Genuer, R., Poggi, J.-M., Tuleau-Malot, C., Villa-Vialaneix, N., 2015. Random Forests for Big Data. arXiv preprint arXiv:1511.08327.

[Wright 2015] Wright, M.N., Ziegler, A., 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Milestones

The term project is an opportunity for you to explore parallelizing an interesting machine learning problem in the context of a large scale data set. Projects can be done in project teams. Your project is worth 10 points of course grade with the following 4 deliverables as well as source code in github:

- ○ **Proposal**:1 page (10%) due Oct 9
- ○ **Midway Report**:3-4 pages (20%) due Nov 6
- ○ **Final Report**: 5-6 pages (40%) due Nov 27
- ○ **Presentation**: (30%) Dec 1

Project Proposal:

Read the list of available tasks below. You must turn in a brief project proposal (1-page maximum), including the following information:

- Project title (authors and affiliations)
- Task and data set
- Project idea. This should be approximately two paragraphs.
- Open source code and software you will use.
- Papers to read. You will read 1-3 relevant papers before submitting your proposal
- Milestone: What will you complete by Oct 9? Experimental results of some kind are expected. You should also specify what portion of the project each partner contributes.

Midway Report:

This should be a 3-4 pages short report, and it serves as a check-point. It should consist of the same sections as your final report (introduction, related work, method, experiment, conclusion), with a few sections `under construction'. Specifically, the introduction and related work sections should be in their final form; the section on the proposed method should be almost finished; the sections on experiments and conclusions can include the results you get, or `place-holders' for results you plan to obtain.

Grading for the project report:

- 70% for proposed method (should be almost finished)
- 25% for the design of upcoming experiments
- 5% for plan of activities (in an appendix, please show the old one and the revised one, along with the activities of each team member)

Final Report:

Your final report is expected to be 5-6 pages in length. It should follow the following format:

- Introduction - Motivation
- Problem definition
- Proposed method
    - Intuition - why should it work?
    - Description of your approach to distributed algorithms using Harp.
- Experiments
    - Description of your testing environment and questions your experiments are designed to answer
    - Details of the experiments; observations and figures
- Conclusions

Presentation:

We will have all teams present a poster in a poster session Dec 1. At least one project member should be presenting during the poster hours. The session will be open to everybody.