mlpack update

a quick update on things that have changed

August 28, 2019

https://www.ratml.org/misc/mlpack-meeting-slides.pdf

overview

- 74 pull requests merged since the last meeting (April 4, 2019)!
- I won't be able to cover all of them quickly, but it's really incredible how much is happening. I can't keep up!
- 10 GSoC projects this year; GSoC is just coming to a close now.
- 2750->**2900+** stars on Github, 97k->**113k** downloads (according to my server logs so that's an undercount), 142->**145** contributors and counting...
- Updates for today (since 4/4): administrivia, **neural networks**, **reinforcement learning**, **testing** , **documentation**, **bugfixes** **optimizations/speedups**, **new features**, **GSoC**, **ensmallen**. *(Not comprehensive.)*

administrivia

Some notes about project management and direction. Opinions and help are always welcomed!

- **Joining NumFOCUS**: *in progress*
- **mlpack 3.2.0 release**: *probably should have happened a while ago; let's make it happen soon!*
- **GSoC**: coding is over, evaluations due very soon. It's been a great summer and I hope everyone has enjoyed it as much as I have!
- **New website**: gmanlan has been hard at work putting together a new website and it's almost ready. I think it is a huge improvement and should help people find and be able to use the library.
- **Documentation and discoverability**: we add lots of new features—but sometimes these can be hard to find! The success of anything we add (as measured by number of users at least) is directly correlated with how easy it is to use. Basically everything we have could be improved. :)

neural networks

saksham189, *walragatver*, *ShikharJ*, *zoq*, *MuLx10*, *jeffin143*, and *cascala* have all been hard at work adding new support for neural networks and GANs.

- #1952: padding layer (`Padding<>`)
- #1978: support for more layer types
- #1956: bias visitor
- #1761: highway networks
- #1829: axis selection for `Concat<>`)
- #1800, #1865: accessors for parts of layers
- #1920: weight normalization layer
- #1932: added regularizers
- #1926: GAN virtual batch normalization
- #1843: concatenated ReLU
- #1823: GAN/FFN memory sharing
- models#29, models#30: LSTM univariate and multivariate time series models

neural networks (2)

Two of our GSoC projects this year were on *Implementing Essential Deep Learning Modules*, mentored by ShikharJ. walragatver (Toshal) and saksham189 (Saksham) worked hard to improve mlpack's GAN support.

Check out their blog posts for more details:

- Toshal's: `http://mlpack.org/blog/Toshal2019Summary.html`
- Saksham's: `http://mlpack.org/blog/SakshamBansalPage.html`

Reinforcement Learning

I'm not very knowledgeable about reinforcement learning but zoq, robotcator, favre49, and abhinavsagar are, and they have been improving mlpack's RL framework.

- #1931: pendulum action type
- #1614: Prioritized Experience Replay (PER)
- #1901: multiple pole balancing environment
- #1886: add accessors for the state and environment
- #1841: add decay rate to Greedy policy

zoq (Marcus) and manish7294 (Manish) mentored robotcator (Xiaohong) over the summer to implement PER and PPO. Xiaohong's summary blog post can be found here:

`http://mlpack.org/blog/Xiaohong2019Summary.html`

testing

Testing is hard! It's really difficult to keep all the pieces of our system running. **20 PRs** involving testing.

It might be straightforward how to reliably test deterministic code... but machine learning algorithms can be *much harder!*

rcurtin, zoq, Hsankesara, walragatver, saksham189, abhinavsagar, favre49, robotcator, Yashwants19

#1986, #1809, #1979, #1924, #1971, #1973, #1899, #1947, #1930, #1921, #1902, #1968, #1953, #1959, #1864, #1816, #1813, #1951, #1914, #1724

Any takers on #1741? (probably difficult to do right...)

documentation

It's awesome to see so many documentation improvements. Thank you zoq, birm, atulim, jeffin143, gmanlan, abhinavsagar, greatsharma, rcurtin, and favre49!

There were **17 PRs** about improving documentation or code quality:

#1977, #1976, #1945, #1925, #1917, #1907, #1898, #1897, #1794, #1893, #1889, #1890, #1882, #1832, #1831, #1828, #1941.

bugfixes

Always good to get things fixed. (Probably I should have made a release happen when these were each merged.)

- #1970: gcc 9 build fixes lozhnikov
- #1905: layer serialization fixes walragatver
- #1922: fix accuracy calculations Yashwants19
- #1891: fix random forest lack of randomness rcurtin
- #1847: fix `Predict()` in RNNs MuLx10

optimizations and speedups

There's always room for more speed!

- #1780: use `std::unordered_map` in `NormalizeLabels()` jeffin143
- #1855: refactor `ccov()` jeffin143
- #1860, #1666: wrap Armadillo to provide a `DiagonalGMM` class KimSangYeon-DGU
- #1943: `Backward()` computation optimization saksham189

new features

Through GSoC and other work, there is a lot that is new in the next release...

- `data::Load()` for images using `stb` (#1903 MuLx10)
- Monte Carlo sampling for faster KDE: use the `monte_carlo` option (#1934 robertohueso)
- Data scaling functionality: `mlpack_preprocess_scale`, `MaxAbsScaler`, `PCAScaler`, etc. (#1876 jeffin143)
- `max_depth` parameter for decision trees and random forests (#1916 Yashwants19)
- `ConfusionMatrix()` function (#1798 jeffin143)
- `OneHotEncoding()` function (#1784 jeffin143)

new features (2)

I mentored robertohueso (Roberto) this summer to implement improvements to kernel density estimation. This went really well; check out the final blog post here:

http://mlpack.org/blog/Roberto2019Summary.html

lozhnikov (Mikhail) mentored jeffin143 (Jeffin) for *String Processing Utilities*. This resulted in some really cool and helpful data science-type transformations and utilities. Check out the blog post here:

http://mlpack.org/blog/Jeffin2019Summary.html

gsoc

Some of the other projects were aimed at other parts of the mlpack project than just the main `mlpack` repository.

KimSangYeon-DGU (SangYeon) worked all summer under the guidance of sumedhghaisas (Sumedh) to explore Quantum GMMs further. This ended up being more implementation and research of QGMM itself; the writeup is very informative:

`https://github.com/KimSangYeon-DGU/GSoC-2019`

MuLx10 (Mehul) spent the summer implementing models for the `models` repository, with the help of zoq (Marcus). Lots of models were implemented and this should be really helpful both for benchmarking and documentation. See the writeup:

`http://mlpack.org/blog/Mehul2019Summary.html`

gsoc (2)

sreenikSS (Sreenik) was advised by akhandait (Atharva) and zoq (Marcus) to implement an mlpack-Torch translator. This code can be found at `https://github.com/sreenikSS/mlpack-Tensorflow-Translator`. The writeup is here:

`http://mlpack.org/blog/Sreenik2019Summary.html`

zoq (Marcus) mentored favre49 (Rahul), whose project was to implement NEAT, an optimization scheme for neural networks. This included some useful implementations of tasks for reinforcement learning environments, and NEAT is implemented in PR #1908. See the final report:

`http://mlpack.org/blog/Rahul2019Summary.html`

ensmallen

Some `ensmallen` updates:

- `ensmallen` is now packaged in Fedora, CentOS, and RHEL.
- Callbacks support is nearly ready (#119).
- Templated `Optimize()` is also ready and will be merged with callbacks. This will allow optimization with, e.g., `arma::fmat`.
- We'll submit `ensmallen` to the Journal of Machine Learning Research, and possibly a workshop paper on callbacks and templated `Optimize()` to the NeurIPS SysML workshop.
- Multi-objective optimization/NSGA-III (#120).

This summer, SuryodayBasak (Suryo) implemented Particle Swarm Optimization (PSO) with the help of zoq (Marcus) and rcurtin (Ryan). PR #85. See Suryo's writeup:

`https://medium.com/@suryodaybasak/a-tutorial-on-particle-swarm-opti`

I may have missed some things, but hopefully this was a useful update!

https://www.ratml.org/misc/mlpack-meeting-slides.pdf