

pomegranate: Fast and Flexible Probabilistic Modeling in Python

Jacob Schreiber

*Paul G. Allen School of Computer Science
University of Washington
Zheng-Chu Seattle, WA 98195-4322, USA*

JMSCHR@CS.WASHINGTON.EDU

Editor: Balazs Kegl

Abstract

We present pomegranate, an open source machine learning package for probabilistic modeling in Python. Probabilistic modeling encompasses a wide range of methods that explicitly describe uncertainty using probability distributions. Three widely used probabilistic models implemented in pomegranate are general mixture models, hidden Markov models, and Bayesian networks. A primary focus of pomegranate is to abstract away the complexities of training models from their definition. This allows users to focus on specifying the correct model for their application instead of being limited by their understanding of the underlying algorithms. An aspect of this focus involves the collection of additive sufficient statistics from data sets as a strategy for training models. This approach trivially enables many useful learning strategies, such as out-of-core learning, minibatch learning, and semi-supervised learning, without requiring the user to consider how to partition data or modify the algorithms to handle these tasks themselves. pomegranate is written in Cython to speed up calculations and releases the global interpreter lock to allow for built-in multithreaded parallelism, making it competitive with—or outperform—other implementations of similar algorithms. This paper presents an overview of the design choices in pomegranate, and how they have enabled complex features to be supported by simple code. The code is available at <https://github.com/jmschrei/pomegranate>

Keywords: probabilistic modeling, Python, Cython, machine learning, big data.

1. Introduction

The Python ecosystem is becoming increasingly popular for the processing and analysis of data. This popularity is in part due to easy-to-use libraries such as numpy (van der Walt et al., 2011), scipy (Jones et al., 2001), and matplotlib (Hunter, 2007) that aim to provide fast general purpose functionality. However, equally important are the libraries that are built on top of these to provide higher level functionality, such as pandas (McKinney, 2010) for data analysis, scikit-image (van der Walt et al., 2014) for computer vision, Theano (Theano Development Team, 2016) for efficient evaluation of mathematical expressions, gensim (Rehůřek and Sojka, 2010) for topic modeling in natural language processing, and countless others. Naturally, many machine learning packages have also been developed for Python, including those that implement classic machine learning algorithms, such as scikit-learn (Pedregosa et al., 2011), mlpy (Albanese et al., 2012), shogun (Sonnenburg et al., 2017), and xgboost (Chen and Guestrin, 2016).

pomegranate fills a gap in the Python ecosystem that encompasses building probabilistic machine learning models that utilize maximum likelihood estimates for parameter updates. There are several packages that implement certain probabilistic models in this style individually, such as hmmlearn for hidden Markov models, libpgm for Bayesian networks, and scikit-learn for Gaussian mixture models and naive Bayes models. However, pomegranate implements a wider range of probabilistic

models and does so in a more modular fashion than these other packages, having two main effects. The first is that the addition of a new probability distribution in pomegranate allows for all models to be built using that distribution immediately. The second is that improvements to one aspect of pomegranate immediately propagate to all models that would use that aspect. For example, when GPU support was added to multivariate Gaussian distributions, this immediately meant that all models with multivariate Gaussian emissions could be GPU accelerated without any additional code. pomegranate currently includes a library of basic probability distributions, naive Bayes classifiers, Bayes classifiers, general mixture models, hidden Markov models, Bayesian networks, Markov chains, as well as implementations of factor graphs and k-means++/|| that can be used individually but primarily serve as helpers to the primary models.

There are several already existing Python libraries that implement Bayesian methods for probabilistic modeling. These include, but are not limited to, PyMC3 (Salvatier et al., 2016), PyStan (Stan Development Team, 2016), Edward (Tran et al., 2016), pyro (Inc., 2017), and emcee (Foreman-Mackey et al., 2013). Bayesian approaches typically represent each model parameter as its own probability distribution, inherently capturing the uncertainty in that parameter, whereas maximum likelihood approaches typically represent each model parameter as a single value. An example of this distinction is that a mixture model can either be represented as a set of probability distributions and a vector of prior probabilities, or as a set of probability distributions that themselves have probability distributions over their respective parameters (such as the mean and standard deviation, should these distributions be normal distributions) and as a dirichlet distribution representing the prior probabilities. The first representation typically specifies models that are faster to both train and perform inference with, while the second is illustrative of the type of models one could build with packages that implement Bayesian methods, such as PyMC3. Both representations have strengths and weaknesses, but pomegranate implements models falling solely in the first representation.

pomegranate was designed to be easy to use while not sacrificing on computational efficiency. Models can either be specified by writing out each of the components individually if known beforehand, or learned directly from data if not. Key features, such as out-of-core learning and parallelization, can be toggled for each model independently of the definition or method calls, typically by simply passing in an optional parameter. The core computational bottlenecks are written in Cython and release the global interpreter lock (GIL), enabling multi-threaded parallelism that typically Python modules cannot take advantage of. Lastly, linear algebra operations such as matrix-matrix multiplications are implemented using BLAS with the ability to toggle a GPU if present.

All comparisons were run on a computational server with 24 Intel Xeon CPU E5-2650 cores with a clock speed of 2.2 GHz, a Tesla K40c GPU, and 256 GB of RAM running CentOS 6.9. The software used was pomegranate v0.8.1 and scikit-learn v0.19.0. pomegranate can be installed using `pip install pomegranate` or `conda install pomegranate` on all platforms. Pre-built wheels are available for Windows builds, removing the sometimes difficult requirement of a working compiler.

2. The API

pomegranate provides a simple and consistent API for all implemented models that mirrors the scikit-learn API as closely as possible. The most important methods are `fit`, `from_samples`, `predict` and `probability`. The `fit` method will use the given data and optional weights to update the parameters of an already initialized model, using either maximum-likelihood estimates (MLE) or expectation-maximization (EM) as appropriate. In contrast, the `from_samples` method will create a model directly from data in a manner similar to scikit-learn’s `fit` method. For simple models like single distributions this corresponds only to MLE on the input data, but for most other models this corresponds to an initialization step plus a call to `fit`. This initialization can range from using k-means for mixture models to structure learning for Bayesian networks. The `predict` method returns the posterior estimate $\arg\max_M P(M|D)$, identifying the most likely component of

the model for each sample. The `probability` method returns the likelihood of the data given the model $P(D|M)$. The other methods include `predict_proba` which returns the probability of each component for each sample $P(M|D)$, `predict_log_proba` which returns the log of the previous value, and `summarize` and `from_summaries` that jointly implement the learning strategies detailed below.

3. Key Features

pomegranate supports many learning strategies that can be employed during training, including out-of-core learning for massive data sets, semi-supervised learning for data sets with a mixture of labeled and unlabeled data, and minibatch learning. In addition, one can employ multithreaded parallelism or a GPU for data-parallel speedups. These features are made possible by separating out the collection of sufficient statistics from a data set (using the `summarize` method) from the actual parameter update step (using the `from_summaries` method).

Sufficient statistics are the smallest set of numbers needed to calculate some statistic on a data set. As an example, fitting a normal distribution to data involves the calculation of the mean and the variance. The sufficient statistics for the mean and the variance are the sum of the weights of the points seen so far $\left(\sum_{i=1}^n w_i\right)$, the sum of the weighted samples $\left(\sum_{i=1}^n w_i X_i\right)$, and the sum of the weighted samples squared $\left(\sum_{i=1}^n w_i X_i^2\right)$. The mean and variance can then be directly calculated from these three numbers using the following two equations:

$$\mu = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \quad \sigma^2 = \frac{\sum_{i=1}^n w_i X_i^2}{\sum_{i=1}^n w_i} - \left(\frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \right)^2 \quad (1)$$

Out-of-core Learning: The additive nature of the sufficient statistics means that if one were to summarize two batches of data successively and then add the sufficient statistics together, they would get the same sufficient statistics as if they were calculated from the full data set. This presents an intuitive way to handle data sets that are too large to fit in memory, by chunking the data set into batches that do fit in memory and summarizing them successively, adding the calculated sufficient statistics together afterwards. This can be done by passing in a `batch_size` parameter to your training method, for example `model.fit(X, batch_size=10000)` would train a pre-initialized model on more data than can fit in memory by successively summarizing batches of size 10,000 until the full data set has been seen. The `summarize` and `from_summaries` methods can also be used independently to implement custom out-of-core strategies.

Minibatch Learning: A natural extension of the out-of-core strategy is minibatch learning, where a parameter update is done after one or a few batches, instead of the full data set. This is in contrast to batch methods that calculate an update using the entire data set, and stochastic methods that typically update using only a single sample. Minibatching can be specified by passing values to both `batch_size` and `batches_per_epoch` parameters when using `fit` or `from_summaries`, where the `batches_per_epoch` is the number of batches to consider before making an update.

Semi-supervised Learning: Semi-supervised learning is the task of fitting a model to a mixture of both labeled and unlabeled data. Typically this arises in situations where labeled data is sparse, but unlabeled data is plentiful, and one would like to make use of both to learn an informed model.

pomegranate supports semi-supervised learning for `HiddenMarkovModel`, `BayesClassifier`, and `NaiveBayes` models as a combination of EM and MLE. Models are initialized using MLE on the labeled data. Next, a version of EM is used that combines the sufficient statistics calculated from the labeled data using MLE with the sufficient statistics calculated from the unlabeled data using EM at each iteration until convergence. This is automatically toggled whenever -1 is present in the label set, following scikit-learn conventions.

This EM-based approach compares favorably to scikit-learn. To demonstrate, we generate a data set of 100k samples in 10 dimensions from 2 overlapping Gaussian ellipses with means of 0 and 1 respectively and standard deviations of 2. It took pomegranate ~ 0.04 s to learn a Gaussian naive Bayes model with 10 iterations of EM, ~ 0.2 s to learn a multivariate Gaussian Bayes classifier with a full covariance matrix with 10 iterations of EM, whereas the scikit-learn label propagation model with a RBF kernel did not converge after ~ 220 s and 1000 iterations, and took ~ 2 s with a knn kernel with 7 neighbors. Both pomegranate models achieved validation accuracies over 0.75, whereas the scikit-learn models did no better than chance.

Parallelism: Another benefit of the use of additive sufficient statistics is that it presents a clear data-parallel way to parallelize model fitting. Simply, one would divide the data into several batches and calculate the sufficient statistics for each batch locally. These sufficient statistics can then be added together back on the main job and all parameters updated accordingly. This is implemented by dividing the data into batches and running `summarize` on each of them using separate threads and then running `from_summaries` after all threads finish. Typically, the global interpreter lock (GIL) in Python prevents multiple threads from running in parallel in the same python process. However, since the computationally intensive aspects are written in Cython the GIL can be released, allowing for multiple threads to run at once. On a synthetic data set with 3M samples with 1K dimensions it takes ~ 65 seconds to train a Gaussian naive Bayes classifier using pomegranate with 1 thread, but only ~ 17 seconds with 8 threads. For comparison, it takes ~ 53 seconds to train a Gaussian naive Bayes classifier using scikit-learn. On another synthetic data set with 2M samples and 150 dimensions it takes pomegranate ~ 470 s to learn a Gaussian mixture model with a full covariance matrix with 1 thread, ~ 135 s with 4 threads, ~ 57 s with 16 threads, and ~ 200 s using a GPU. Lastly, we compared the speed at which pomegranate and hmmlearn could train a 10 state dense Gaussian hidden Markov model with diagonal covariance matrices. On a synthetic data set of 100 sequences, each containing 1,000 10 dimensional observations, it took hmmlearn ~ 25 s to run five iterations of Baum-Welch training, while it only took pomegranate ~ 13 s with 1 thread, ~ 4 s with 4 threads, and ~ 2 s with 16 threads.

4. Discussion

pomegranate aims to fill a niche in the Python ecosystem that exists between classic machine learning methods and Bayesian methods by serving as an implementation of flexible probabilistic models. The design choices that were made early on while building pomegranate allowed for a great number of useful features to be added later on without significant effort.

A clear area of improvement in the future is the handling of missing values, because many probabilistic models can intuitively modify the EM algorithm to infer these missing values. For example, when trying to learn a Bayesian network over a data set with missing values, one can identify the best structure over the incomplete data set, infer the missing values, and relearn the structure, iterating until convergence. Given the prevalence of missing data in the real world, extending pomegranate to handle missing data efficiently is a priority.

Acknowledgments

We would like to first acknowledge all of the contributors and users of pomegranate, whom without this project would not be possible. We would also like to acknowledge Adam Novak, who wrote the first iteration of the hidden Markov model code. Lastly, we would also like to acknowledge Dr. William Noble for suggestions and guidance during development. This work was partially supported by NSF IGERT grant DGE-1258485.

References

- Davide Albanese, Roberto Visintainer, Stefano Merler, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. *mlpy: Machine learning python*, 2012.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL <http://arxiv.org/abs/1603.02754>.
- D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The mcmc hammer. *PASP*, 125:306–312, 2013. doi: 10.1086/670067.
- John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, pages 90–95, 2007.
- Uber Technologies Inc. pyro. <https://github.com/uber/pyro>, 2017.
- Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL <http://www.scipy.org/>.
- Wes McKinney. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Olivier Grisel, Bertrand Thirion, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesna. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, pages 2825–2830, 2011.
- Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. <http://is.muni.cz/publication/884893/en>.
- John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, April 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.55. URL <https://doi.org/10.7717/peerj-cs.55>.
- Soeren Sonnenburg, Heiko Strathmann, Sergey Lisitsyn, Viktor Gal, Fernando J. Iglesias García, Wu Lin, Soumyajit De, Chiyuan Zhang, frx, tklein23, Evgeniy Andreev, JonasBehr, sploving, Parijat Mazumdar, Christian Widmer, Pan Deng / Zora, Saurabh Mahindre, Abhijeet Kislay, Kevin Hughes, Roman Votyakov, khalednasr, Sanuj Sharma, Alesia Novik, Abinash Panda, Evangelos Anagnostopoulos, Liang Pang, Alex Binder, serialhex, Esben Sørig, and Björn Esser. shogun-toolbox/shogun: Shogun 6.0.0 - Baba Nobuharu, April 2017. URL <https://doi.org/10.5281/zenodo.556748>.
- Stan Development Team. Pystan: the python interface to stan, 2016.

- Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL <http://arxiv.org/abs/1605.02688>.
- Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.
- Stéfan van der Walt, Chris Colbert, and Gaël Varoquaux. The numpy array: A structure for efficient numerical computation. *Journal of Machine Learning Research*, pages 22–30, 2011.
- Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: Image processing in python. *PeerJ*, 2014.