Project topics for MATH-412

Fall 2020

Goal of the projects

The projects are to be done by groups of three students. Each group choses a topic and studies one or two papers on this topic. It will generally be useful to have read more than the main paper associated with a given topic in the provided list. The main evaluation criterion for the report that you will submit and the oral presentation is that you are able to demonstrate that you have understood how the proposed approach works and in which case it is applicable. In particular, you are able to use the concepts of the course like regularization, overfitting, validation to think about how the algorithm performs on a practical case, and you are able to make connections between the algorithms of the papers and the ones seen in class. Depending on the paper, the project does not have to cover the paper exhaustively, and can focus on the main algorithm. It is better to investigate well one method than several too superficially. In the context of an implementation make sure that you are making a comparison with the most related algorithms seen in class if it is comparable.

Oral Presentation

The oral presentation will take place on December 16th and December 17th respectively at the time of the class and the time of the exercise session,

Written report

The report is due on January 8th. The report should be 4 pages (A4 format) and in pdf. It should be submitted directly on Moodle. You may include an appendix of at most 4 pages that includes additional figures or additional experimental results). If you are writing in Latex (recommended) you can easily adjust the margin of the document by inserting `\usepackage{fullpage}` in the header of you latex document order to have standard margins.

Data

Finding data for your project can go from a relatively easy task to an absolutely daunting task if you would like to find the perfect dataset or if you are too ambitious. I would strong recommend to first test any algorithm that you are experimenting with on synthetic data that you would have generated yourself ; this is usually a very good way to understand the algorithm better, because you know what the algorithm should learn. After that, try and work with small real datasets first, and go for bigger later if relevant. A lot of data sources offer to download massive datasets that would only make your life extremely complicated for the projects of this course, so please avoid being drawned in data.

— Some papers in the list below propose experiments on datasets that they reference and that you can possibly still find.

If you would like to find a new dataset of you project, I would recommend to start with :

- The University of California at Irvine ML repository is one of the original repository listing datasets that can be used for projects in machine learning http://archive.ics.uci.edu/ml/ and, usually, the dataset is associated with an ML question (e.g. a classification or a regression task).
- Kaggle, the organizer of ML challenges, maintains a list of datasets https://www.kaggle.com/datasets which will often also be associated with specific questions or challenges.

Some other places on the internet where you can find a plethora of data (Be warned that many data sets are there for the sake of being indexed somewhere but are not a priori associated with a particular question or problem.) :

- EU Open Data Portal https://data.europa.eu/euodp/data/
- https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b
- https://opendatascience.com/25-excellent-machine-learning-open-datasets/
- Ideas of machine learning project with data sets
  https://data-flair.training/blogs/machine-learning-project-ideas/
  (you can use the datasets but for the projects in the list below, the projects proposed on that webpage are not projects that as of themselves can be picked as project for this course.)

List of topics

Gaussian processes

If kernel regression provides an elegant mathematical framework to a learn a non-parametric function, Gaussian processes are somehow their natural Bayesian counterparts. The same way that classical Bayesian method learn a distribution over parameters, the Bayesian models based on Gaussian processes are a way to learning posterior distributions over function spaces, with distributions that generalize the multivariate Gaussian distribution to distributions on functions. The goal of the project is to understand how Gaussian processes work, to apply them to a a non-linear regression problem and to compare them to kernel regression. One possible extension is to consider the particular case of Bayesian linear regression.

- A web site with many ressources :
  http://www.gaussianprocess.org/
- In particular the two first chapters of the book of Carl Rasmussen :
  http://www.gaussianprocess.org/gpml/chapters/

Non linear dimensionality reduction

Principal Component Analysis can be interpreted among other as a technique to approximate the data with a low dimensional representation via a linear transformation. It is suitable if the data lies close to a low dimensional subspace, and is somewhat limited by the fact that it is based on linear transformations. There exists a number of non-linear dimension reduction techniques. The most known are kernel PCA, MDS, Isomap, LLE, Laplacian Eigenmaps, U-MAPs, t-SNE and obviously auto-encoders. The goal of this project is to compare several of these methods on a well chosen dataset. I could possibly be the MNIST dataset (see link below).

- http://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction
- http://www.math.uwaterloo.ca/~aghodsib/courses/f06stat890/readings/tutorial_stat890.pdf
- Scholkopf, B., Smola, A., & Müller, K. R. (1999). Kernel principal component analysis. In Advances in kernel methods-support vector learning.
  http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.7613
- Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), 1373-1396.
  http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/Laplacian.pdf

— MNIST database http://yann.lecun.com/exdb/mnist/

Label propagation via diffusion on a graph

One of the simplest learning algorithm is the k-nearest neighbor predictor. However if a too small subset of the data is labelled, the Euclidian distance between labelled points (or any original distance in the original feature space) is not a good notion of distance on the data. When a significant amount of unlabelled data is available, it can be used to gain information about the geometry of the data. In particular, it is more relevant to use a graph connecting neighboring points, and try to let the information provided by labeled point diffuse on the graph. This is actually also relevant even if a larger fraction of the data is labelled. Harmonic functions on graph are related to heat diffusion equations on the graph, and also joint Gaussian models on variables whose correlation structure is specified by the graph. The goal of the project is to understand the method and to apply it to the data from the paper or to new data.

— Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In The 20th International Conference on Machine Learning (ICML), 2003. ICML 10-Year Classic Paper Prize.
http://pages.cs.wisc.edu/~jerryzhu/pub/zgl.pdf

Pairwise coupling and round-robin classification

In a number of settings, to solve multi-class classification problems, it is more efficient to combine multiple binary classifier than learning directly a multi-class logistic regression model, a multi-class decision tree, etc. One approach to this problem is to to use all-vs-all classification, which is also known as round-robin classification. Furthermore, assuming that each of the individual binary classifiers returns a probability which is calibrated, in the sense that it provides a proper estimate of the probability of the class label given the score information, pairwise coupling is one of the main techniques to compute compatible probability estimates for the original multi-class problem.

— Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. The Journal of Machine Learning Research, 5, 975-1005. http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf

— Fürnkranz, J. (2002). Round robin classification. The Journal of Machine Learning Research, 2, 721-747.

An incremental algorithm for large scale learning

A large number of learning problem are formulated as optimization problem. The simplest algorithm for the minimization of the empirical risk is gradient descent, which computes the gradient of the empirical risk at each iteration. But computing a gradient requires to go through the whole database at each iteration. An alternative is to use stochastic gradient descent, which only requires to consider a single datapoint at a time. but unfortunately the asymptotic convergence of stochastic gradient descent is slow and if it is useful to make many passes through the data, then the algorithm takes time. A number of incremental algorithms (which only consider a small number of datapoints at each iteration) also known as *stochastic algorithms with variance reduction* have been proposed in the last 10 years. The goal of this project is to understand one of the simplest ones SDCA (Stochastic Dual Coordinate Ascent), or SVRG (Stochastic Variance-Reduced Gradient descent) and to compare them in terms of speed with gradient descent and stochastic gradient descent.

— SDCA : Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning, 85(1-2), 41-75. http://www.csie.ntu.edu.tw/~cjlin/papers/maxent_dual.pdf

— In R https://www.rdocumentation.org/packages/gradDescent/versions/3.0/

— In R https://github.com/xinkai-zhou/Stochastic-Dual-Coordinate-Ascent

An extension of the project could consider other related algorithms like SVRG, SAG and SAGA. See the following paper (only the beginning) for a unified presentation of the different algorithms

- SAGA : A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, A. Defazio, F. Bach and S. Lacoste-Julien, Neural Information Processing Systems Conference (NIPS14), Montreal, Canada, December 2014.
- SAG and SAGA in Scikit-learn in Python
  https://towardsdatascience.com/dont-sweat-the-solver-stuff-aea7cddc3451
- https://contrib.scikit-learn.org/lightning/

Anomaly/novelty detection

In many application domains, machine learning techniques can be used to check whether the considered datapoints are "normal" in the sense that they follow the common distribution of the rest of the data. The focus of this project is to study techniques that allow to learn anomaly/novelty detection model from "normal" data only (and not as a simple binary classification problem which is only possible when sufficiently many anomalies exists and are labelled as such).

- Khan, S. S., & Madden, M. G. (2013). One-Class Classification : Taxonomy of Study and Review of Techniques. arXiv preprint arXiv :1312.0049.
  http://arxiv.org/abs/1312.0049
- Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (1999). Support Vector Method for Novelty Detection. In NIPS (Vol. 12, pp. 582-588).
  http://users.cecs.anu.edu.au/~williams/papers/P126.pdf
- Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471.
  http://research.microsoft.com/pubs/69731/tr-99-87.pdf

Classification without negative examples

Since it is often easier to prove that something is true than that something is false, in some classification problems, it is easy to produce positive examples amount a large number of unlabelled data, but it is difficult to say if the remaining points are positive or negative examples. For example, the problem of predicting whether there can be a protein-ligand interaction between a protein and a certain drug is an interesting machine learning problem that can be very useful to speed up drug discovery, but only a small number of existing interactions are known and for the others protein-ligand pairs, we don't know. This problem might seem similar to the previous one, except that in this setting we have many unlabelled examples.

- Khan, S. S., & Madden, M. G. (2013). One-Class Classification : Taxonomy of Study and Review of Techniques. arXiv preprint arXiv :1312.0049.
  http://arxiv.org/abs/1312.0049
- Blanchard, G., Lee, G., & Scott, C. (2010). Semi-supervised novelty detection. Journal of Machine Learning Research, 11, 2973-3009.
  http://jmlr.org/papers/volume11/blanchard10a/blanchard10a.pdf
- Thiran, J. P., Gass, V., Borgeaud, M., Tuia, D., & de Morsier, F. (2013). Semi-Supervised Novelty Detection using SVM entire solution path. IEEE Transactions on Geoscience and Remote Sensing, 51, 1939-1950.
  http://infoscience.epfl.ch/record/175357/files/SSNDNCSSVM_demorsier_infoscience.pdf&version=1

Non-negative matrix factorization & application to musical notes estimation

This project is about non-negative matrix factorization (NMF) a cousin of PCA in which both the *factors* and the *decomposition coefficients* are constrained to be non-negative. Different algorithms exist depending on which loss function is considered the square loss, the Kullback-Leibler divergence, or the Itakura-Saito divergence) The second article proposes an application for the estimation of individual musical note from a polyphonic recording.

- Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562).
- Févotte, C., Bertin, N., & Durrieu, J. L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. Neural computation, 21(3), 793-830.
- The data can be found here : https://www.irit.fr/~Cedric.Fevotte/index.html

Distance learning

Most learning problem start with a certain representation of the data points by a feature vector (which can be close to the raw data or obtained via feature engineering), but it is often difficult to specify ahead of time which features are most important and what is a good measure of distance between these feature vector. The Euclidean metric is often used by default and implicitly, just because we use the associated dot product. This is done for lack of a better metric. Distance learning techniques address in a way this type of question. In the deep learning community metric learning is often associated with the *triplet loss* (https://en.wikipedia.org/wiki/Triplet_loss)

- Kulis, B. (2012). **Metric learning : a survey**. Found. and Trends in Machine Learning, 5(4), 287-364.
- Bellet, A., Habrard, A., & Sebban, M. (2013). **A Survey on Metric Learning for Feature Vectors and Structured Data.** arXiv preprint arXiv :1306.6709.
  http://arxiv.org/abs/1306.6709

"Learning from the crowd"

In a certain number of learning problems, the "ground truth labels" are not really available, but instead, the data has been labelled by human experts that can make mistakes, and so the labels are "noisy" and not perfectly reliable. For example for tumor detection or classification in radiological images, the best experts do not necessarily agree on the correct labels. When several experts are available, and provide each they own label, one naive approach consists in assigning a consensus label or an average label, but it turns out that it is possible to do much better by learning at the same time the machine predictor and a collection of skill models for each of the experts, the idea being that experts who do not agree with the consensus are probably making mistakes and that this can inform the main learning algorithm. This type of technique is clearly useful if the data is labelled via "crowd sourcing".

- Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy ; 11(Apr) :1297 ?1322, 2010.
  http://jmlr.org/papers/volume11/raykar10a/raykar10a.pdf
- Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy Modeling Annotator Expertise : Learning when Everybody Knows a Bit of Something In Proc. International Conference on Artificial Intelligence and Statistics (AISTATS) 2010.
  http://people.csail.mit.edu/romer/papers/AIStatsAnnotExpertise.pdf

Fairness of decision algorithm in machine learning

Is it acceptable for an algorithm making a recommendation on whether a bank should grant a loan to an individual to know the gender of that individual ? To know their religious belief or their ethnicity ? Same question for an algorithm learning to rank automatically applicants to specific jobs on an "intelligent" human resource platform ? One might wish that an algorithm would not learn that it is better to be man to be a mechanic in spite of the statistical bias present in the data. Can an algorithm decide whether a detainee can be set free based on a model predicting his/her probability of committing new criminal offenses ? If so based on which data ? If the law forbids the use of some information (like gender, ethnicity), is it possible to build algorithms that guarantee that this information cannot be used indirectly by mistake because of the correlation with authorized variables ? Is possible to define fairness ? Or embed equal opportunity principles into algorithms ? The goal of this project is to study the concepts introduced in the following paper and to apply them to a binary classification problem.

— Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315-3323). https://arxiv.org/abs/1610.02413

Spectral Clustering

To solve unsupervised classification problems, algorithms such as k-means or the Gaussian mixture model assume that the clusters are round of have an ellipsoidal shape. In practice, the cluster forming the main lumps of the distribution can have much more irregular shapes. A method to decompose the support of the distribution into its main parts is spectral clustering, which is based on graph spectral theory.

— Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395-416. http://www.cyberneum.de/fileadmin/user_upload/files/publications/luxburg06_TR_v2_4139[1].pdf

Auto-encoders

Auto-encoders are neural networks that provide non-linear generalization of principal component analysis (PCA). The general principle is to build a compressed version of the data that would allow to reconstruct them as well as possible. The compressed version is obtained on the central hidden layer and the input and output are the data that needs to be "encoded". The model is thus trained to reconstruct the input data for example for the square loss. The goal of the project is to understand the similarities and difference between PCA and auto-encoders, to train an auto-encoder on real data, and possibly to compare the performance of the auto-encoder and of PCA on a denoising task.

— https://www.cs.toronto.edu/~hinton/science.pdf

— Chapitre 4.6 de http://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf

References on matrix factorization (and its relation to PCA)

— https://en.wikipedia.org/wiki/Low-rank_approximation

— A paper on the relation of auto-encoders with PCA :
http://ace.cs.ohiou.edu/~razvan/courses/dl6890/papers/bourlard-kamp88.pdf

— Relation between matrix factorization and PCA :
Sections 1 and 2.1 of https://people.csail.mit.edu/tommi/papers/SreJaa-aim03.pdf

— http://www.niss.org/sites/default/files/tr185.pdf

Covariate Shift and Domain adaptation

In a number of real world problems the data available for training does not follow exactly the same distribution as the test data for which we would like to learn a predictor. The training data is labelled and the test data is not (or it only contains a very small number of annotations). For example, you would like to learn a spam filter, and you have a database with annotated spam, but you would like to train a model which is going to be working well for somebody who is a pharmacist and you have access the emails of that person but they are not labelled. How can you train a model on the first set and make sure that it is well suited to the new data ? One key assumption is the *covariate shift* assumption : the distribution $p(y|x)$ has not changed but $p(x)$ has changed... The main techniques to address covariate shift is to use density-ratio estimation (DRE). The goal of this project is to understand *covariate shift* (based on the paper by Moreno-Torres and/or chapter 9.1 of the book of Sugiyama et al.) and one or two techniques for DRE and to apply them to address the *covariate shift* issue. As methods for DRE I would recommend to concentrate on techniques of chapter 4.2 (or 4.3) and on chapter 6 (LSIF).

— Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern recognition, 45(1), 521-530.
https://www.sciencedirect.com/science/article/pii/S0031320311002901

- A book on DRE : Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). Density ratio estimation in machine learning. Cambridge University Press.
https://1lib.eu/book/2927624/af5a34/
- Slides giving on many techniques of DRE
s http://yosinski.com/mlss12/MLSS-2012-Sugiyama-Density-Ratio-Estimation-in-Machine-Learning/
- Paper on least-squares importance fitting (LSIF)
https://www.jmlr.org/papers/volume10/kanamori09a/kanamori09a.pdf

Variable Importance

In a number of algorithms and models learned with different loss functions, it is possible to derive from the final predictor or from the algorithms, some measures of importance which aim at quantifying which variables contribute most the prediction. The goal of this project is to understand what are a few importance measures, what they measure exactly and what are the differences between different measures, what are pitfalls and best practice, and then to test some of these methods on a real dataset and assess the stability of these methods as well as compare what is obtained with each of them. The project would have to cover at least some of the main importance measures for random forest and then choose other measures based on relevance and interest.

- Variable importance measures in regression and classification methods
https://www.ifi.uzh.ch/dam/jcr:82fc9567-e690-40fa-baff-eb7a37aa00c0/MasterThesis.pdf

Ordinal Regression

Ordinal regression (also called "ordinal classification") is a type of regression analysis used for predicting an ordinal variable, i.e. a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant. It can be considered an intermediate problem between regression and classification. This would be for example the right method to use if you wanted to predict the rating of a movie directly from characteristics of the user and/or of the movie.

- To just get started https://en.wikipedia.org/wiki/Ordinal_regression
- To understand a few main ideas : https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/
- Once you understood the two previous ones : Gutiérrez, P. A. ; Pérez-Ortiz, M. ; Sánchez-Monedero, J. ; Fernández-Navarro, F. ; Hervás-Martínez, C. (January 2016). "Ordinal Regression Methods : Survey and Experimental Study". IEEE Transactions on Knowledge and Data Engineering. 28 (1) : 127–146
https://ieeexplore.ieee.org/document/7161338