Machine Learning notes

Soumitra Samanta

April 6, 2022

Contents

Topic 1

Motivation and Mathematics background: Some motivating ML practical examples, Linear algebra, Analysis and Probability theory

1.1 Suggested reading

Please go through the class slides.

1.2 ML datasets repository

You can find some datasets to evaluate yous ML models in *UCI Machine Learning Repository* (`https://archive.ics.uci.edu/ml/datasets.php`)

1.3 ML/AI top tier conference

- International Conference on Machine Learning (ICML) - `https://icml.cc/`

- Neural Information Processing Systems (NeurIPS) - https://neurips.cc/
- International Conference on Learning Representations (ICLR) - https://iclr.cc/
- Association for the Advancement of Artificial Intelligence (AAAI) - https://www.aaai.org/

1.4 ML top journals

- Journal of Machine Learning Research (JMLR) - https://www.jmlr.org/

1.5 For recent updates on ML you can follow the arXiv

You can go to Computer Science (CS) section in arXiv and under that you can find different branches of CS (like AI, ML, etc.).

- AI - https://arxiv.org/list/cs.AI/recent
- ML - https://arxiv.org/list/cs.LG/recent

Topic 2

Data: Representation/Featurization, Normalization (after some idea in classification/regression), Data partition (train, val and test)

2.1 Suggested reading

Please go through the class slides.

Topic 3

k- nearest neighbour (kNN) classifier

3.1 Suggested reading

For general discussion and algorithms, you can go through Duda et al.'s [6] book *Chapter 4, Section 4.5: The nearest-neighbour Rule*.

For deep theoretical development, you can look at Devroye et al.'s [5] book *Chapter 5*. You may find some helpful results in *Chapter 19* of Shalev-Shwartz and Ben-David's [17] book as well.

Topic 4

Bays decision theory

4.1 Suggested reading

For Bays decision theory, you can go through Duda et al.'s [6] book *Chapter 2*.

Topic 5

Regression: Linear, Ridge, LASSO

5.1 Suggested reading

You can find Linear and Ridge regression in any Machine Learning book.

For ordinary linear least square regression, you can go through *Chapter 9, Section 9.2: Linear Regression* of Shalev-Shwartz and Ben-David's book [17] or book by Mohri et al. [10] *Chapter 11, Section 11.3.1: Linear regression*.

Ridge regression was first invented by Tikhonov [21] in 1943 in the context of integral equations. Later Horel and Kennard [8] introduced in Statistics in 1970. As in this case we are taking L_2 norm of the parameters as a penalty/constraint, some people called it least squares with L_2 regularization. For book reference, you can go through Bishop's book [1] *Chapter 3, Section 3.1.4: Regularized least squares* and Murphy's book [11] *Chapter 7, Section 7.5: Ridge regression*. You can find some numerical tricks in *Sub-section 7,5.2*.

The *Least absolute shrinkage and selection operator*(LASSO) was invented by Tibshirani [19] in 1996 and later some more modifications in the literature [20]. For reference book by Hastie et al. [7] *Chapter 1 & 2*.

If you would like to correlate the different least square regression with statistical likelihoods and priors, then please have a look at Murphy's book [11] *Chapter 7 (Table 7.1 give you the summary)*.

Topic 6

Linear classifier (perceptron leaning algorithm)

6.1 Suggested reading

The perceptron learning algorithms were proposed by Frank Rosenblatt in 1958 [14]. An online version of the original paper can be found here. Also, he wrote a technical report [15] in detail about perceptron.

You can find perceptron learning algorithms in any Machine Learning or Pattern Recognition book. Here are some references:

You can find the algorithm in Shalev-Shwartz et al. book [17] *Chapter 9, Section 9.1.2: Perceptron for Half-spaces* and those of you interested in the convergence proof please go through the *Theorem 9.1* in [17].

Mohri et al. book [10] *Chapter 8, section 8.3.1: Perceptron algorithm* and for convergence proof *Theorem 8.8* [17].

For a brief history and original perceptron setup picture [1]

Topic 7

Support Vector Machine (SVM)

7.1 Suggested reading

Support Vector Machine is one of the most popular classification algorithms in Machine Learning. It was developed by Boser et al. [2] in 1992. The original paper can be found here. For rigorous theoretical results, You can go through Vapnik's (father of SVM) book [22]. Here are some book references we will follow for our course:

Mohri et al. book [10] *Chapter 5* and Shalev-Shwartz et al. book [17] *Chapter 15*.

Topic 8

Logistic regression

8.1 Suggested reading

Please go through the Murphy's book [11] *Chapter 8*. For our purpose *Sections 8.1, 8.2 and 8.3* are sufficient, but I am encouraging you to go through the whole chapter.

Topic 9

Decision tree

9.1 Suggested reading

Decision tree (DT) is one of the widely used algorithm in machine learning. There are many versions for the DT and here is a survey on *fifty years celebration on classification and regression tree* [9] Here are some popular algorithms based on different impurity measurements ID3 [12], C4.5 [13], CART [4]. For a summary, you can look into Shalev-Shwartz et al. book [17] *Chapter 18*. For detailed theory and related results, you can go through *Chapter 20* of Devroye et al. book [5].

Topic 10

Random forest

10.1 Suggested reading

The main paper of Random forest [3] and for more theoretical details you can look into *Chapter 15* of Hastie et al. book [7]. Here is some recent work on *When do random forests fail?* [18].

Topic 11

Boosting

11.1 Suggested reading

There is a whole book on Boosting by Schapire and Freund [16]. You can find the AdaBoost algorithm in *Algorithm 1.1* in [16], and for more details, you can go through the whole book. You can find the AdaBoost algorithm in Shalev-Shwartz et al. book [17] *Chapter 10, Section 10.2*, and I am encouraging you to go through the whole chapter of this book, particularly the *Theorem 10.2*.

Topic 12

Clustering

12.1 Suggested reading

For *Hierarchical clustering (bottom-up or agglomerative and top-down or divisive)*, you can go through the *Chapter 14, Section 14.3.12* of Hastie at al. book [7].

You can find *k-means* algorithm in any machine learning book. You can go through the Shalev-Shwartz and Ben-David's book *Chapter 22, Section 22.2* [17] and I am also encouraging you to go through the LEMMA 22.1.

Here is a survey on various clustering algorithms [23].

Bibliography

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2006.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir Naumovich Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5 – 32, 2001.

[4] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Chapman and Hall/CRC, 1st edition, 1984.

[5] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1st edition, 1996.

[6] Richard Duda, Peter Hart, and David Stork. *Pattern Classification.* Wiley, 2nd edition, 2007.

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 1st edition, 2015.

[8] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[9] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329 – 348, 2014.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

[11] Kevin Patrick Murphy. *Machine Learning: a Probabilistic Perspective*. The MIT Press, 1st edition, 2012.

[12] John Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[13] John Ross Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann, 1st edition, 1992.

[14] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[15] Frank Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Laboratory, 1961.

[16] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 1st edition, 2012.

[17] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1st edition, 2014.

[18] Cheng Tang, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B*, 73(3):273–282, 2011.

[21] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5):195–198, 1943.

[22] Vladimir Naumovich Vapnik. *Statistical Learning Theory*. Wiley, 1st edition, 1998.

[23] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.