

CHRISTIAN MICHELSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

A PHYSICIST'S
APPROACH TO
MACHINE LEARNING
—
UNDERSTANDING
THE BASIC BRICKS

SUPERVISOR:
TROELS PETERSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

Copyright © 2019

Christian Michelsen

`HTTPS://GITHUB.COM/CHRISTIANMICHELSEN`

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, December 2019

Contents

1	<i>Abstract</i>	1
2	<i>Introduction</i>	3
3	<i>Machine Learning Theory</i>	5
3.1	<i>Statistical Learning Theory</i>	5
3.2	<i>Supervised Learning</i>	6
3.3	<i>Generalization Bound</i>	7
3.3.1	<i>Generalization Bound for infinite hypotheses</i>	9
3.4	<i>Avoiding overfitting</i>	10
3.4.1	<i>Model Regularization</i>	10
3.4.2	<i>Cross Validation</i>	12
3.4.3	<i>Early Stopping</i>	14
3.5	<i>Loss functions</i>	14
3.5.1	<i>Evaluation Function</i>	16
3.6	<i>Decision Trees</i>	16
3.6.1	<i>Ensembles of Decision Trees</i>	17
3.7	<i>Hyperparameter Optimization</i>	19
3.7.1	<i>Grid Search</i>	20
3.7.2	<i>Random Search</i>	20
3.7.3	<i>Bayesian Optimization</i>	21
3.8	<i>Feature Importance</i>	23
4	<i>Danish Housing Prices</i>	27
4.1	<i>Data Preparation and Exploratory Data Analysis</i>	28
4.1.1	<i>Correlations</i>	30
4.1.2	<i>Validity of input variables</i>	32
4.1.3	<i>Cuts</i>	33

4.2	<i>Feature Augmentation</i>	33
4.2.1	<i>Time-Dependent Price Index</i>	34
4.3	<i>Evaluation Function</i>	35
4.4	<i>Initial Hyperparameter Optimization</i>	36
4.5	<i>Hyperparameter Optimization</i>	38
4.6	<i>Results</i>	40
4.7	<i>Model Inspection</i>	44
4.8	<i>Multiple Models</i>	46
4.9	<i>Discussion</i>	48
5	<i>Particle Physics and LEP</i>	51
6	<i>Quark Gluon Analysis</i>	53
6.1	<i>Sidenotes</i>	53
7	<i>Discussion and Outlook</i>	61
8	<i>Conclusion</i>	63
8.1	<i>Tufte-\LaTeX Website</i>	63
8.2	<i>Tufte-\LaTeX Mailing Lists</i>	63
8.3	<i>Getting Help</i>	63
A	<i>Housing Prices Appendix</i>	65
B	<i>Quarks vs. Gluons Appendix</i>	91
	<i>Index</i>	99

List of Figures

3.1	Overview of the learning problem.	6
3.2	Approximation-Estimation tradeoff	10
3.3	Regularization Effect	11
3.4	Regularization Effect of L_2	12
3.5	Regularization Effect of L_1	12
3.6	k -Fold Cross Validation	13
3.7	k -Fold Cross Validation for Time Series Data	13
3.8	Comparison of different objective functions.	16
3.9	Comparison of different objective functions zoom in.	16
3.10	Decision Tree Cuts In Feature Space	16
3.11	Decision Tree	17
3.12	Grid Search	20
3.13	Random Search	21
3.14	Bayesian Optimization	22
4.1	Danish Housing Price Index	27
4.2	Distributions for the housing price dataset	29
4.3	Distributions for the housing price dataset	30
4.4	Histogram of prices of houses and apartments sold in Denmark	30
4.5	Linear correlation between variables and price	31
4.6	MIC non-linear correlation.	31
4.7	Non-linear correlation between variables and price	32
4.8	Validity of input features	32
4.9	Validity Dendrogram	33
4.10	Prophet Forecast for apartments	34
4.11	Prophet Trends	35
4.12	XXX	37
4.13	Overview of initial hyperparameter optimization of the housing model for apartments	38
4.14	XXX	38
4.15	XXX	38
4.16	XXX	39
4.17	Hyperparameter optimization: random search results	40
4.18	Early Stopping results	40
4.19	Performance of XGB-model on apartment prices	41
4.20	2018 XGB Forecast	41
4.21	2018 XGB Forecast	43
4.22	SHAP Prediction Explanation for apartment	44
4.23	Feature importance of apartments prices using XGB	45

4.24 Feature importance of apartments prices using XGB XXX	46
4.25 Multiple Models XXX	47
4.26 SHAP plot villa TFIDF XXX	49
5.1 Feynman diagram for the jet production at LEP	51
6.1 Histograms of the vertex variables	53
6.2 UMAP vizualisation of vertex variables	54
6.3 b-tag scores in 3-jet events	54
6.4 ROC curve for b-tag in 4-jet events	54
6.5 g-tag scores in 4-jet events	55
6.6 g-tag scores in 4-jet events for signal and background	55
6.7 ROC curve for g-tag in 4-jet events	55
6.8 1D Sum Model Cuts for 4-jets	56
6.9 1D Sum Models Predictions and Signal Fraction for 4-jets	56
6.10 Hyperparameter Optimization of b- and g-tagging	56
6.11 Overview of Hyperparamaters of g-tagging for 3-jet shuffled events	56
6.12 SHAP Prediction Explanation for b-like jet	57
6.13 Monte Carlo – Data bias for b-tags and jet energy	57
6.14 b-Tagging Efficiency $\epsilon_b^{b\text{-sig}}$ as a function of jet energy	57
6.15 b-Tagging Efficiency $\epsilon_b^{g\text{-sig}}$ as a function of jet energy	57
6.16 b-Tagging Efficiency $\epsilon_g^{g\text{-sig}}$ as a function of jet energy	58
6.17 b-Tagging Efficiency $\epsilon_g^{b\text{-sig}}$ as a function of jet energy	58
6.18 g-Tagging proxy efficiency for $b\bar{b}g$ -events as function of the mean invariant mass	58
6.19 g-Tagging proxy efficiency for $b\bar{b}g$ -events as function of g-tag	58
6.20 g-Tagging efficiency for 4-jet events in MC as a function of normalized gluon gluon jet energy difference	59
6.21 Closure plot between MC Truth and the corrected g-tagging model in 4-jet events for the normalized gluon gluon jet energy difference	59
6.22 R kt CA overview XXX TODO!	59
6.23 R kt CA cut region A XXX TODO!	59
A.1 Validity Heatmap	65
A.2 Distributions for the housing price dataset	66
A.3 Distributions for the housing price dataset	67
A.4 Distributions for the housing price dataset	68
A.5 Distributions for the housing price dataset	69
A.6 Distributions for the housing price dataset	70
A.7 Distributions for the housing price dataset	71
A.8 Distributions for the housing price dataset	72
A.9 Distributions for the housing price dataset	73
A.10 Distributions for the housing price dataset	74
A.11 Distributions for the housing price dataset	75
A.12 Distributions for the housing price dataset	76
A.13 Distributions for the housing price dataset	77
A.14 Distributions for the housing price dataset	78
A.15 Distributions for the housing price dataset	79
A.16 Linear Correlations	81

A.17	MIC non-linear correlation	82
A.18	Prophet Forecast for apartments	82
A.19	Prophet Trends	82
A.20	Overview of initial hyperparameter optimization of the housing model for houses	86
A.21	XXX	87
A.22	XXX	87
A.23	XXX	87
A.24	XXX	88
A.25	XXX	88
A.26	XXX	88
A.27	Performance of XGB-model on apartment prices	89

List of Tables

4.1	XXX TODO!	29
4.2	XXX TODO!	33
4.3	XXX TODO!	33
4.4	XXX TODO!	33
4.5	XXX TODO!	33
4.6	train test split XXX TODO!	36
4.7	train test split tight XXX TODO!	36
4.8	Cauchy-ejerlejlighed.	37
4.9	Cauchy-villa.	37
4.10	XXX	39
4.11	XXX	41
4.12	XXX ejer	43
4.13	XXX villa	43

A.1	XXX TODO!	80
A.2	Rmse-ejerlejlighed-appendix.	83
A.3	Logcosh-ejerlejlighed-appendix.	83
A.4	Cauchy-ejerlejlighed-appendix.	83
A.5	Welsch-ejerlejlighed-appendix.	84
A.6	Fair-ejerlejlighed-appendix.	84
A.7	Rmse-villa-appendix.	84
A.8	Logcosh-villa-appendix.	84
A.9	Cauchy-villa-appendix.	85
A.10	Welsch-villa-appendix.	85
A.11	Fair-villa-appendix.	85
A.12	XXX ejer tight	90
A.13	XXX villa tight	90

1. Abstract

This sample book discusses the design of Edward Tufte's books and the use of the `tufte-book` and `tufte-handout` document classes.

5. Particle Physics and LEP

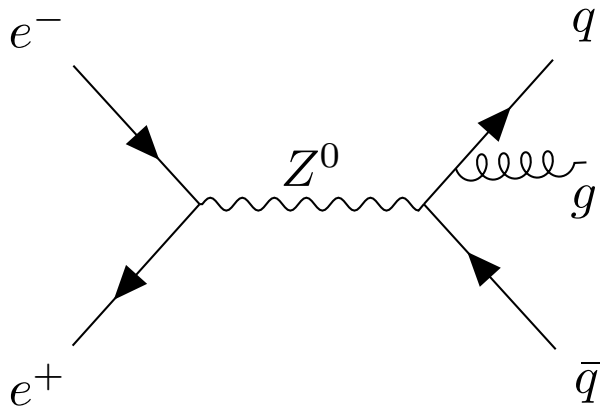


Figure 5.1: Feynman diagram showing the $e^+e^- \rightarrow Z^0$ production at LEP. The Z^0 has several decay modes where the $Z \rightarrow q\bar{q}g$ is shown here.

B. Quarks vs. Gluons Appendix

Bibliography

- [1] Advanced Topics in Machine Learning (ATML). URL <https://kurser.ku.dk/course/ndak15014u>.
- [2] Allstate Claims Severity - Fair Loss. URL <https://kaggle.com/c/allstate-claims-severity>.
- [3] Dmlc/xgboost. URL <https://github.com/dmlc/xgboost>.
- [4] HEP meets ML award | The Higgs Machine Learning Challenge. URL <https://higgsml.lal.in2p3.fr/prizes-and-award/award/>.
- [5] The Large Electron-Positron Collider | CERN. URL <https://home.cern/science/accelerators/large-electron-positron-collider>.
- [6] Datashader: Revealing the Structure of Genuinely Big Data. URL <https://github.com/holoviz/datashader>.
- [7] O. . Www.OIS.dk - Din genvej til ejendomsdata. URL <https://www.ois.dk/>.
- [8] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook. ISBN 978-1-60049-006-4.
- [9] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi. A practical tool for maximal information coefficient analysis. 7. ISSN 2047-217X. doi: 10.1093/gigascience/gyi032. URL <https://doi.org/10.1093/gigascience/gyi032>.
- [10] E. Anderson. The Species Problem in Iris. 23(3):457–509. ISSN 00266493. doi: 10.2307/2394164. URL www.jstor.org/stable/2394164.
- [11] M. Awad and R. Khanna. Support Vector Regression. In M. Awad and R. Khanna, editors, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 67–80. Apress. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4. URL https://doi.org/10.1007/978-1-4302-5990-9_4.
- [12] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. WileyBlackwell, reprint edition. ISBN

- o-471-92295-1. URL <http://www.amazon.co.uk/Statistics-Statistical-Physical-Sciences-Manchester/dp/0471922951%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0471922951>.
- [13] J. T. Barron. A General and Adaptive Robust Loss Function. URL <http://arxiv.org/abs/1701.03077>.
- [14] J. Bergstra and Y. Bengio. Random Search for Hyperparameter Optimization. 13:281–305. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2188385.2188395>.
- [15] Bolighed. Bolighed - usikkerhed i data-vurderingen. URL <https://bolighed.dk/om-bolighed/spoergsmaal-og-svar/#boligvaerdi>.
- [16] L. Breiman. Random Forests. 45(1):5–32. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- [17] E. Brochu, V. M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. URL <http://arxiv.org/abs/1012.2599>.
- [18] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794. doi: 10.1145/2939672.2939785. URL <http://arxiv.org/abs/1603.02754>.
- [19] S. D. DST. Price Index (EJ14) - Statistics Denmark. URL <https://www.dst.dk/en/Statistik/emner/priser-og-forbrug/ejendomme>.
- [20] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. 7(2):179–188. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>.
- [21] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In P. Vitányi, editor, *Computational Learning Theory*, pages 23–37. Springer Berlin Heidelberg. ISBN 978-3-540-49195-8. Adaboost.
- [22] A. E. Harvey and S. Peters. Estimation Procedures for Structural Time Series Models. doi: 10.1002/for.3980090203.
- [23] T. Hastie and R. Tibshirani. Generalized Additive Models: Some Applications. 82(398):371–386. ISSN 01621459. doi: 10.2307/2289439. URL www.jstor.org/stable/2289439.
- [24] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second*

- Edition*. Springer Series in Statistics. Springer-Verlag, 2 edition. ISBN 978-0-387-84857-0. URL [//www.springer.com/la/book/9780387848570](http://www.springer.com/la/book/9780387848570).
- [25] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley. ISBN 978-1-118-21033-8. URL https://books.google.dk/books?id=j10hquR_j88C.
 - [26] S. Hviid, Juul. Working Paper: A regional model of the Danish housing market. URL <http://www.nationalbanken.dk/en/publications/Pages/2017/11/Working-Paper-A-regional-model-of-the-Danish-housing-market.aspx>.
 - [27] F. James and M. Roos. Minuit – a system for function minimization and analysis of the parameter errors and correlations. 10:343–367. doi: 10.1016/0010-4655(75)90039-9.
 - [28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc. URL <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
 - [29] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49(4): 764–766. ISSN 0022-1031. doi: 10.1016/j.jesp.2013.03.013. URL <http://www.sciencedirect.com/science/article/pii/S0022103113000668>.
 - [30] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 4768–4777. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL <http://dl.acm.org/citation.cfm?id=3295222.3295230>.
 - [31] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles - SHAP. . URL <http://arxiv.org/abs/1802.03888>.
 - [32] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. . URL <http://arxiv.org/abs/1802.03888>.
 - [33] I. Mulalic, H. Rasmussen, J. Rouwendal, and H. H. Woltmann. The Financial Crisis and Diverging House Prices: Evidence from the Copenhagen Metropolitan Area. ISSN 1556-5068. doi:

- 10.2139/ssrn.3041272. URL <https://www.ssrn.com/abstract=3041272>.
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. 12:2825–2830.
- [35] E. Polley and M. van der Laan. Super Learner In Prediction. URL <https://biostats.bepress.com/ucbbiostat/paper266>.
- [36] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. 334(6062):1518–1524. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1205438. URL <http://science.sciencemag.org/content/334/6062/1518>.
- [37] P. J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. 88(424):1273–1283. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476408. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408>.
- [38] L. Shapley. A value for n-person games. In *The Shapley Value*, volume 28 of *Annals of Math Studies*, pages 307–317. doi: 10.1017/CBO9780511528446.003.
- [39] S. J. Taylor and B. Letham. Forecasting at scale. doi: 10.7287/peerj.preprints.3190v2. URL <https://peerj.com/preprints/3190>.
- [40] i. team. Iminuit – A python interface to minuit. URL <https://github.com/scikit-hep/minuit>.
- [41] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. 58(1):267–288. ISSN 0035-9246. URL www.jstor.org/stable/2346178.
- [42] A. Tikhonov. *On the Stability of Inverse Problems*, volume vol. 39 of *Doklady Akademii Nauk SSSR*.
- [43] d. L. M. J. van, E. C. Polley, and A. E. Hubbard. Super Learner. 6(1). ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL <https://www.degruyter.com/view/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml>.
- [44] F. van Veen. The Neural Network Zoo. URL <http://www.asimovinstitute.org/neural-network-zoo/>.
- [45] V. Vapnik. Principles of Risk Minimization for Learning Theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pages 831–838. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-222-9. URL <http://dl.acm.org/citation.cfm?id=2986916.2987018>.

- [46] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python. page arXiv:1907.10121. URL <https://ui.adsabs.harvard.edu/abs/2019arXiv190710121V/abstract>.
- [47] H. Wickham. Tidy data. 59(10):1–23. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL <https://www.jstatsoft.org/v059/i10>.

Index

license, [ii](#)