Kickoff Presentation Bachelor Thesis

Estimating hold-out-sample Performance for Active Learning

Florian Bethe

OvGU Magdeburg - Department of Computer Science

17th of September 2015

Table of Contents

Theoretical background for error estimation of clasifiers

- Rodríguez et al. [2013] defines classifier and error
- Hastie et al. [2009] gives a pretty good overview of everything except learning curves
- Dietterich [1995] examines overfitting
- Kohavi and Wolpert [1996] analyses bias-variance-decomposition for 0-1-loss
- Krogh and Vedelsby [1995] details the bias-variance-tradeoff

Classifier-dependent estimators

- AIC: introduced by Akaike [1998], explained and simplified by Bozdogan [1987] and Hastie et al. [2009]
- BIC: first presented by Schwarz [1978]. Further analysis provided by Weakliem [1999]
- VC-theory: published by Vapnik [1982], used by Brumen et al. [2004]

Cross-Validation

- LOO and K-fold: Kohavi [1995]
- Pahikkala et al. [2008] provides further analysis and testing
- Adaptive incremental k-fold described in Brumen et al. [2004]
- Error and bias-variance analysis in Airola et al. [2011], Rodríguez et al. [2013] and Efron and Tibshirani [1997]

Bootstrapping

- Plain bootstrap: Kohavi [1995]
- LOO: Borra and Di Ciaccio [2010]
- Efron [1983] and Efron and Tibshirani [1997] present .632 and .632+ and provide experiments, with Wood et al. [2007] presenting criticism

Learning curves

- Perlich et al. [2003] and Figueroa et al. [2012] for a general description of learning curves
- Figueroa et al. [2012] also presents algorithm for performance prediction and with Singh [2005] function families for model fitting
- Curve for optimism in Cortes et al. [1993]

Miscellaneous

- Model Retraining Improvement: Potential estimation framework Evans et al. [2015]
- Estimator definition as Monte-Carlo-Simulation in Roy and McCallum [2001]
- Kadie and Wilkins [1995] analyses Maximum-Likelihood-Estimation for finding prediction models

Adaptive Iterative K-fold Cross-Validation

Presented in Brumen et al. [2004]

- K-fold cross-validation modified for use with active learning
- $k = 5$ or 10 for an unbiased estimator Kohavi [1995], Airola et al. [2011]

.632+ bootstrap

Introduced by Efron and Tibshirani [1997]

- Estimates optimism of training error
- Builds on .632 estimator, but corrects bias for overfit classifiers
- Estimation of no-information error rate for binary classifiers given

Hold-out curve extrapolation

Description in Figueroa et al. [2012]

- Based on hold-out estimates for past iterations
- As estimation for accuracy reduction of cross-validation approach?

Estimating "old" accuracy

For each iteration:

- Use hold-k-out cross-validation with $k = 1...n - 1$
- Estimate beta distribution from the samples for each k
- Override old estimates in new iteration; only increased test set or also wider training spectrum?

Fit curve model

For each iteration:

- Compute expectation value for each beta distribution
- Fit log, exp and pow functions using non-linear regression
- Standard deviation of regression as intervals?
- Build beta distribution from old extrapolations using weighting?

Evaluation And Results

- Compare estimated accuracies to true hold-out estimates using m hold-out sets
- For proposed method: if beta distribution obtained, compute KL-divergence
- For comparison methods: obtain beta distribution by using multiple sets?

Questions

Questions?

A. Airola, T. Pahikkala, W. Waegeman, and B. De Baets. An experimental comparison of cross-validation techniques for estimating the area under the roc curve. *Computational Statistics & Data Analysis*, 55(4):1828–1844, April 2011.

H. Akaike. *Selected Papers of Hirotugu Akaike*, chapter Information Theory and an Extension of the Maximum Likelihood Principle, pages 199–213. Springer Series in Statistics. Springer New York, 1998. URL http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

S. Borra and A. Di Ciaccio. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, 54(12): 2976–2989, December 2010. URL http://www.sciencedirect.com/science/article/pii/S0167947310001064.

H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, September 1987. URL http://link.springer.com/article/10.1007/BF02294361.

B. Brumen, I. Golob, H. Jaakkola, T. Welzer, and I. Rozman. Early assessment of classification performance. In *ACSW Frontiers 2004, 2004 ACSW Workshops - the Australasian Information Security Workshop (AISW2004), the Australasian Workshop on Data Mining and Web Intelligence (DMWI2004), and the Australasian Workshop on Software Internationalisation (AWSI2004) . Dunedin, New Zealand, January 2004*, pages 91–96, 2004. URL http://crpit.com/confpapers/CRPITV32Brumen.pdf.

C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker. Learning curves: Asymptotic values and rate of convergence. In

Neural Information Processing Systems, 1993. URL
`http://papers.nips.cc/paper/803-learning-curves-asymptotic-values-and-rate-of-conver`
`pdf`.

T. Dietterich. Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 27(3):326–327, September 1995. URL `http://dl.acm.org/citation.cfm?id=212114`.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983. URL `http://www.cs.berkeley.edu/~jordan/sail/readings/archive/efron-improve_cv.pdf`.

B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, June 1997. URL http://www.stat.washington.edu/courses/stat527/s13/readings/EfronTibshirani_JASA_1997.pdf.

L. P. G. Evans, N. M. Adams, and C. Anagnostopoulous. Estimating optimal active learning via model retraining improvement. *Journal of Machine Learning Research*, 2015. URL http://arxiv.org/abs/1502.01664#.

R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 2012. URL http://www.biomedcentral.com/1472-6947/12/8.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2 edition, 2009. URL http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf.

C. M. Kadie and D. C. Wilkins. *Seer: Maximum Likelihood*

Regression for Learning-Speed Curves. PhD thesis, University of Illinois, 1995. URL http://research.microsoft.com/en-us/um/people/carlk/papers/kadiephdfull.htm.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, 1995. URL http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf.

R. Kohavi and D. H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996. URL ai.stanford.edu/~ronnyk/biasVar.pdf.

A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*. Massachusetts Institute of

Technology, 1995. URL http://papers.nips.cc/paper/1001-neural-network-ensembles-cross-validation-and-activ
pdf.

T. Pahikkala, A. Airola, J. Boberg, and T. Salakoski. Exact and efficient leave-pair-out cross validation for ranking rls. In *Proceedings of AKRR*, 2008.

C. Perlich, F. Provost, and J. S. Simonoff. Tree induction vs. logistic regression: a learning-curve analysis. *The Journal of Machine Learning Research*, 4:211–255, December 2003. URL http://dl.acm.org/citation.cfm?id=945375.

J. D. Rodríguez, A. Pérez, and J. A. Lozano. A general framework for the statistical analysis of the sources of variance for classification error estimators. *Pattern Recognition*, 46(3): 855–864, March 2013. URL http://www.sciencedirect.com/science/article/pii/S0031320312003998.

N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. In *Proceedings of the International Conference on Machine Learning*, 2001. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.137.6296.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. URL http://projecteuclid.org/euclid.aos/1176344136.

S. Singh. Modeling performance of different classification methods: Deviation from the power law. Project Report, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.5455&rep=rep1&type=pdf.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer New York, 1982. URL http://www.springer.com/us/book/9780387308654.

D. L. Weakliem. A critique of the bayesian information criterion for model selection. *Sociological Methods Research*, 27(3):359–397, February 1999. URL `smr.sagepub.com/content/27/3/359.short`.

I. A. Wood, P. M. Visscher, and K. L. Mengersen. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23(11):1363–1370, June 2007. URL `https://keppel.qimr.edu.au/contents/p/staff/CVPV127.pdf`.