

Understanding the Role of Statistics in Machine Learning and Data Science

Statistics provides an important foundation for machine learning (ML) and data science by supporting data analysis, feature selection, model development, interpretation, and evaluation. The key question is how statistical thinking contributes to building reliable predictive models and interpreting their results. The literature demonstrates that although ML often focuses on prediction, statistical principles remain important throughout the modelling process.

Feature selection is an important area where statistical methods contribute to ML. Yang, Liu and Wen investigate Bayesian optimization for feature selection and show that optimizing the parameters of feature-selection methods can improve their performance and, in their experiments, improve the accuracy of following prediction models [1]. This demonstrates how statistical optimization can help identify useful variables while reducing unnecessary information in high-dimensional datasets.

Statistical thinking is also important for understanding the different purposes of statistical modelling and machine learning. Breiman distinguishes between traditional statistical models, which make assumptions about the process generating the data, and algorithmic models, which focus on predictive accuracy without necessarily indicating that data generating process [2]. Bzdok, Altman and Krzywinski further explain that statistical interpretation and machine learning have different primary goals that is statistics can use samples to make population interpretations, whereas ML aims to identify predictive patterns that generalize to new observations [3]. These perspectives show that statistics and ML are closely related but can address different questions.

Model validation provides another important connection. Cross validation is widely used to estimate how well a model will perform on unseen data. However, Varma and Simon demonstrate that using cross validation both to tune a model and estimate its final prediction error can produce biased results [4]. They show that nested cross-validation can reduce this bias and highlight the importance of careful statistical design when evaluating ML models.

Overall, the literature shows that statistics highlight several stages of the ML workflow. Statistical optimisation supports feature selection, statistical reasoning helps distinguish assumption from prediction, and validation methods help determine whether model performance is reliable. Therefore, modern data science combines statistical thinking with computational and algorithmic approaches.

References

- [1] L. L. & Y. W. Kaixin Yang, "Scientific Reports," nature.com, 17 02 2024. [Online]. Available: <https://www.nature.com/articles/s41598-024-54515-w?utm>. [Accessed 21 09 2026].
- [2] L. Brieman, "Statistical Science," *Statistical Modeling: The Two Cultures*, vol. 15, no. 3, pp. 199-231, 2001.
- [3] D. A. N. & K. M. Bzdok, "Statistics versus machine learning," *Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.*, vol. 15, pp. 233-234, 2018.
- [4] S. S. R. B. Varma, "Soringer Nature Link," BMC Series, 23 02 2006. [Online]. Available: <https://link.springer.com/article/10.1186/1471-2105-7-91?utm.com#citeas>. [Accessed 28 04 2005].
- [5] R. T. ., J. F. Trevor Hastie, "Springer Naturelink," *The Elements of Statistical Learning*, vol. 2, p. 745, 26 08 2009.