Kjersti Aas

Short Bio:
Kjersti Aas received M.Sc. degree in Industrial Mathematics from the Norwegian Institute of Technology (NTH) in 1990 and Ph.D. degree in Statistics from the Norwegian University of Science and Technology (NTNU). She is currently Research Director at the Norwegian Computing Center and Adjunct Professor in Data Science at NTNU. Her current research interests include AI and Explainable AI for financial applications.

Topic:

EXPLAINABLE AI

Interpretability is crucial when a complex machine learning model is to be applied in areas such as medicine, fraud detection, or credit scoring. In many applications, complex hard-to-interpret machine learning models like deep neural networks, random forests and gradient boosting machines are currently outperforming the traditional, and to some extent interpretable, linear/logistic regression models. However, often, there is a clear trade-off between model complexity and model interpretability, meaning that it is often hard to understand why these sophisticated models perform so well. This lack of explanation constitutes a practical issue – can I trust the model? and a legal issue – those who develop the model can be required by law to explain what the model does to those who are exposed to automated decisions (the General Data Protection Regulation). In response, a new line of research has emerged that focuses on helping users interpret the predictions from advanced machine learning methods.

Existing work on explaining complex models can be divided into two main categories: global and local explanations. The former tries to describe the model as a whole in terms of which variables/features influenced the general model the most. On the other hand, local explanations try to identify how the different input variables/features influenced a specific prediction/output from the model and are often referred to as individual prediction explanation methods.

I will give three lectures centred around explainable AI. In the first lecture, I will provide a general introduction to this field and discuss some global explanation approaches. In the second lecture, I will introduce Shapley values, one of the most commonly used local explanation methods. Shapley values will also be the theme of the first part of the third lecture, while the second part will contain an introduction to counterfactual explanations, another well-known local explanation method.