# Bayesian Inference and Explainable AI

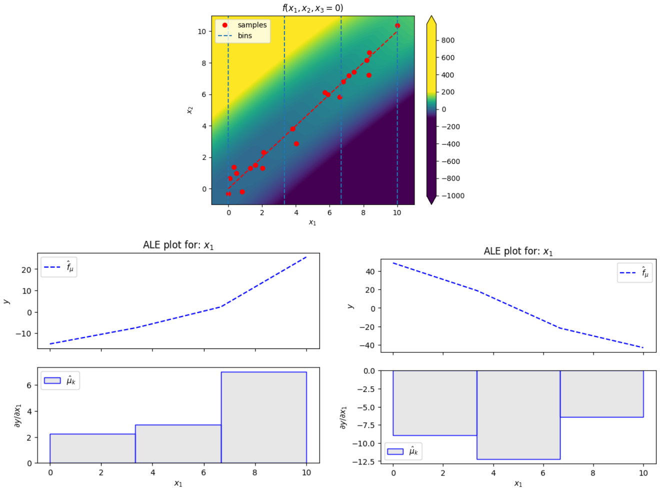
## Quantifying the uncertainty of the explanations

Vasilis Gkolemis[1]

[1]ATHENA Research and Innovation Center

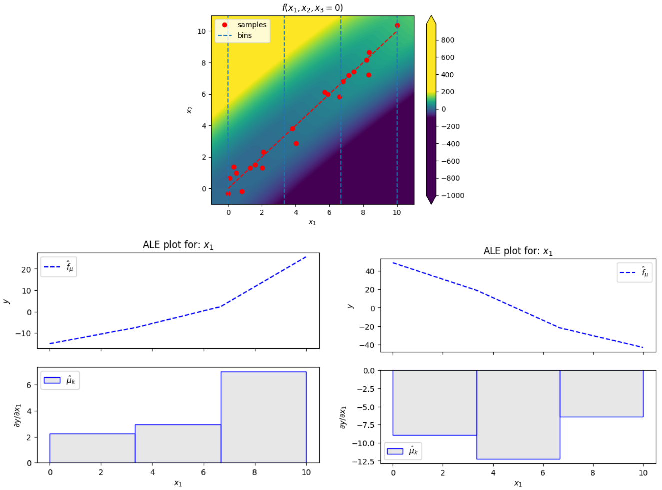
November 2021

# Feature Effect



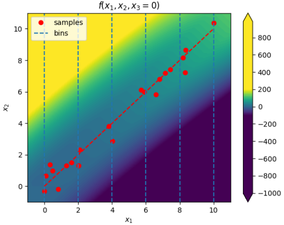

We search for a single configuration (point-estimate) $\hat{\theta}$

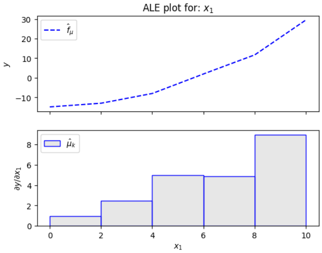
# DALE vs ALE - 3 Bins

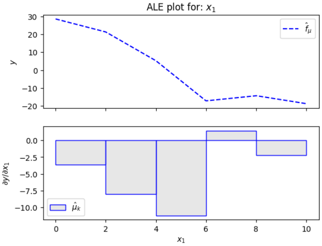



We search for a single configuration (point-estimate) $\hat{\theta}$

# DALE vs ALE - 5 Bins

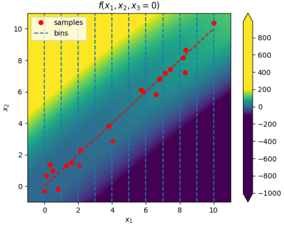


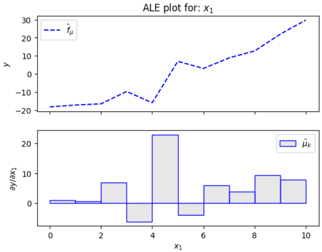



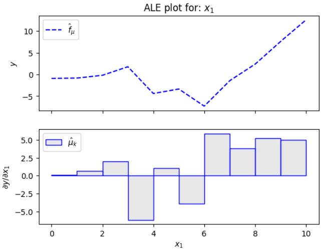




We search for a single configuration (point-estimate) $\hat{\theta}$

# DALE vs ALE - 10 Bins

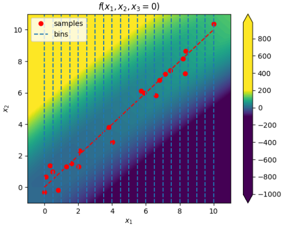


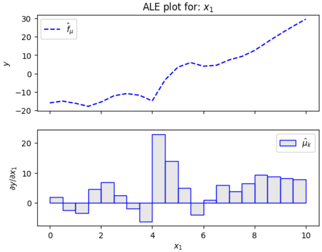



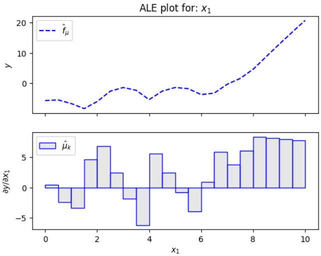




We search for a single configuration (point-estimate) $\hat{\theta}$

# DALE vs ALE - 20 Bins

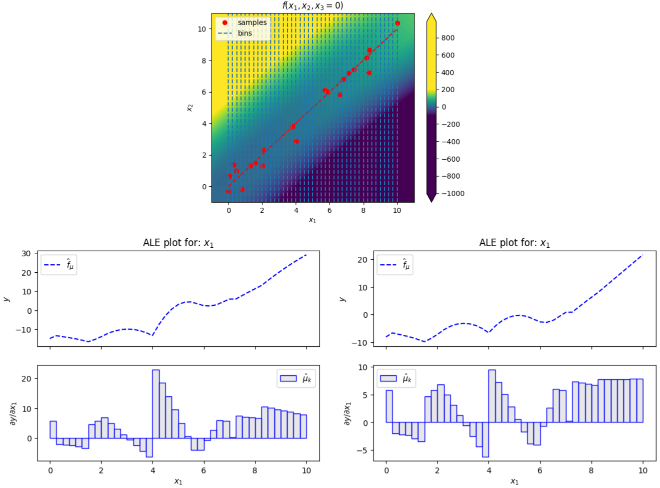



We search for a single configuration (point-estimate) $\hat{\theta}$

# DALE vs ALE - 40 Bins




We search for a single configuration (point-estimate) $\hat{\theta}$

# Traditional Machine Learning

- Parametric model $f_{\boldsymbol{\theta}} : \boldsymbol{x} \rightarrow y$
- Define a distance function $d(\cdot, \cdot)$ and measure the distance (loss) from observed data

$$L(\boldsymbol{\theta}) = \sum_{i}^{N} d(f_{\boldsymbol{\theta}}(\boldsymbol{x}^i), y^i) \tag{1}$$

- Search for the parameter set $\hat{\boldsymbol{\theta}}$ that reproduces the observed data best

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{2}$$

We search for a single configuration (point-estimate) $\hat{\boldsymbol{\theta}}$

# Bayesian Formulation

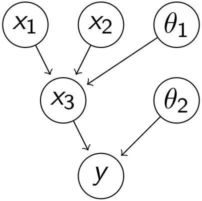
- On the modelling part:
  - we need the joint distribution $p(\boldsymbol{x}, y, \boldsymbol{\theta})$
  - to replace the parametric model $f_{\boldsymbol{\theta}} : \boldsymbol{x} \rightarrow y$
- Training part:
  - infer the posterior distribution $p(\boldsymbol{\theta}|D)$
  - to replace the optimal point estimate $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
- Prediction part:
  - infer the predictive distribution $p(y|\boldsymbol{x}, D)$
  - to replace the point-estimate prediction $y = f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$

---

We replace point estimates with distributions (uncertainty quantification)

# Modelling part

- joint distribution $p(\boldsymbol{x}, y, \boldsymbol{\theta}) = p(y|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x})p(\boldsymbol{\theta})$
- $p(\boldsymbol{\theta})$, our prior belief about the parameters of the model
- $p(y|\boldsymbol{x}, \boldsymbol{\theta})$, the likelihood of the model
- joint distribution can be defined as a DAG




- We need to model $p(\boldsymbol{\theta})$ and $p(y|\boldsymbol{x}, \boldsymbol{\theta})$

# Training part

- We use Bayes law to infer the posterior distribution

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} \propto \prod_i^N p(y^i|\mathbf{x}^i, \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3}$$

  where $D = \{\mathbf{x}^i, y^i\}_{i=\{1,\ldots,N\}}$, the observed data (training-set)
- In the extreme case where $p(\boldsymbol{\theta}|D) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, we get a point-estimate is in traditional ML

---

The 'training process' leads to many possible models, each one with different probability (uncertainty about the model)

# Inference part

- We need to solve/approximate the predictive distribution $p(y|\mathbf{x}, D) = \int_{\boldsymbol{\theta}} p(y|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D) \partial \boldsymbol{\theta}$
- We consider the posterior $p(\boldsymbol{\theta}|D)$ as known (computed exactly or approximated)
- In the extreme case where $p(\boldsymbol{\theta}|D) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, we get all the mass of the prediction $p(y|\mathbf{x}, D) = p(y|\mathbf{x}, \hat{\boldsymbol{\theta}})$ from a single model

---

The 'prediction process' gets one prediction per each plausible model (uncertainty about the model leads to uncertainty about the prediction)

# Bayesian Formulation - Disadvantages

What we lose

- On the modelling part
  - Time to think how the input features $x_i$ relate to each other i.e. building the DAG
- On the training-prediction (inference) part
  - Expressions difficult to approximate
  - $p(\boldsymbol{\theta}|D)$ - how to compute the posterior distribution?
  - $p(y|\mathbf{x}, D) = \int_{\boldsymbol{\theta}} p(y|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D) \partial \boldsymbol{\theta}$ - how to compute the predictive distribution?

---

Bayesian Formulation is difficult from both the mathematical and the computational point-of-view

# Bayesian Formulation - Advantages

What we get:

- On the modelling part
  - Specify who the features relate to each other
  - check Model-based Machine Learning - a new approach for ML model building
- On the inference part
  - Uncertainty estimation! Why we need it?
  - Most times the available data is not enough to reveal a single instance $\hat{\theta}$
  - Sometimes we want to predict on a new $\boldsymbol{x}$ that is very different from the training set

Let's be wise enough and be uncertain about our predictions.

# Bayesian Formulation - Example

- In areas without training points our uncertainty is bigger