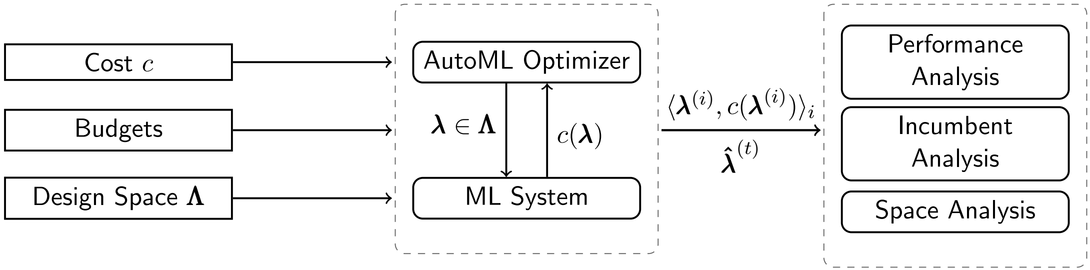
AutoML: Interpretability

Global Hyperparameter Importance

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Idea



$\rightsquigarrow$ focus on which hyperparameters are important across the entire search space

Importance Analysis of Surrogate Model

- Key Idea: Surrogate models ($\mathbf{\Lambda} \to \mathbb{R}$) learn from observations how to predict the performance of a hyperparameter configuration $\boldsymbol{\lambda} \in \mathbf{\Lambda}$

Importance Analysis of Surrogate Model

- Key Idea: Surrogate models ($\boldsymbol{\Lambda} \to \mathbb{R}$) learn from observations how to predict the performance of a hyperparameter configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$

$\rightsquigarrow$ These models can be used to figure out which hyperparameter was important

Importance Analysis of Surrogate Model

- Key Idea: Surrogate models ($\mathbf{\Lambda} \to \mathbb{R}$) learn from observations how to predict the performance of a hyperparameter configuration $\boldsymbol{\lambda} \in \mathbf{\Lambda}$

$\rightsquigarrow$ These models can be used to figure out which hyperparameter was important

- For example:
  - Use forward selection [Hutter et al. 2013]
  - Use automatic feature relevance determination of the model (e.g., of a surrogate model based on random forest)

Importance Analysis of Surrogate Model

- Key Idea: Surrogate models ($\boldsymbol{\Lambda} \to \mathbb{R}$) learn from observations how to predict the performance of a hyperparameter configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$

$\rightsquigarrow$ These models can be used to figure out which hyperparameter was important

- For example:
  - Use forward selection [Hutter et al. 2013]
  - Use automatic feature relevance determination of the model (e.g., of a surrogate model based on random forest)

- Advantages:
  - Very cheap to do, since we only have to query the surrogate model several times

Importance Analysis of Surrogate Model

- Key Idea: Surrogate models ($\boldsymbol{\Lambda} \to \mathbb{R}$) learn from observations how to predict the performance of a hyperparameter configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$

$\leadsto$ These models can be used to figure out which hyperparameter was important

- For example:
  - Use forward selection [Hutter et al. 2013]
  - Use automatic feature relevance determination of the model (e.g., of a surrogate model based on random forest)
- Advantages:
  - Very cheap to do, since we only have to query the surrogate model several times
- Potential drawback:
  - The surrogate model might overfit to different subsets of the hyperparameters (if we don't provide sufficient data)

Global Importance Analysis

- Key idea: What is the importance of a hyperparameter by marginalizing over all other hyperparameter effects?

Global Importance Analysis

- Key idea: What is the importance of a hyperparameter by marginalizing over all other hyperparameter effects?
- Key Insight: We can use a surrogate model to compute these effects

Global Importance Analysis

- Key idea: What is the importance of a hyperparameter by marginalizing over all other hyperparameter effects?
- Key Insight: We can use a surrogate model to compute these effects

fANOVA [Sobobl 1993]

Write performance predictions as a sum of components:

$$
\begin{aligned}
\hat{y}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n) = &\quad \hat{f}_0 + \textstyle\sum_{i=1}^{n} \hat{f}_i(\boldsymbol{\lambda}_i) + \sum_{i \neq j} \hat{f}_{ij}(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j) + \ldots \\
\hat{y}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n) = &\quad \text{average response} + \text{main effects} + \\
&\quad \text{2-D interaction effects} + \text{higher order effects}
\end{aligned}
$$

Global Importance Analysis

- Key idea: What is the importance of a hyperparameter by marginalizing over all other hyperparameter effects?
- Key Insight: We can use a surrogate model to compute these effects

fANOVA [Sobobl 1993]

Write performance predictions as a sum of components:

$$\begin{aligned}\hat{y}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n) &= \hat{f}_0 + \textstyle\sum_{i=1}^{n} \hat{f}_i(\boldsymbol{\lambda}_i) + \sum_{i \neq j} \hat{f}_{ij}(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j) + \ldots \\ \hat{y}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n) &= \text{average response} + \text{main effects} + \\ &\quad \text{2-D interaction effects} + \text{higher order effects}\end{aligned}$$

Variance Decomposition

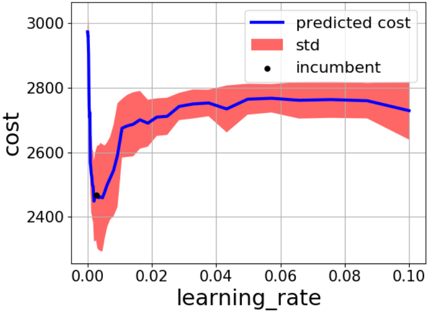
$$\mathrm{V} = \frac{1}{||\boldsymbol{\Lambda}||} \int_{\boldsymbol{\lambda}_1} \ldots \int_{\boldsymbol{\lambda}_n} [(\hat{y}(\boldsymbol{\lambda}) - \hat{f}_0)^2] d\boldsymbol{\lambda}_1 \ldots d\boldsymbol{\lambda}_n$$

fANOVA Analysis

- The fANOVA and variance decomposition can be done efficiently in linear time if the surrogate model is a random forest [Hutter et al. 2014]

fANOVA Analysis

- The fANOVA and variance decomposition can be done efficiently in linear time if the surrogate model is a random forest [Hutter et al. 2014]

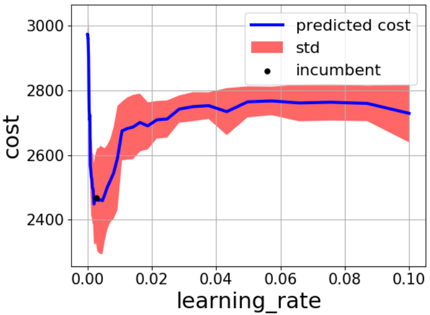


Source: [Lindauer et al. 2019]

- predicted cost is marginalized over all other hyperparameter effects

fANOVA Analysis

- The fANOVA and variance decomposition can be done efficiently in linear time if the surrogate model is a random forest [Hutter et al. 2014]



Source: [Lindauer et al. 2019]

- predicted cost is marginalized over all other hyperparameter effects
- Warning: The optimum on these curves does not have to be the global optimum across all hyperparameters

fANOVA Analysis

- How much of the variance can be explained by a hyperparameter (or combinations of hyperparamaters) marginalized over all other parameters?

Table: Exemplary analysis of PPO on cartpole

Hyperparameter	Explained Variance
Discount rate	19.3 %
Batch size	15.7 %
Learning rate	3.7 %
Likelihood ration clipping	3.4%
...	

fANOVA Analysis

- How much of the variance can be explained by a hyperparameter (or combinations of hyperparamaters) marginalized over all other parameters?

Table: Exemplary analysis of PPO on cartpole

Hyperparameter	Explained Variance
Discount rate	19.3 %
Batch size	15.7 %
Learning rate	3.7 %
Likelihood ration clipping	3.4%
...	
discount rate & batch size	10.4%
discount rate & likelihood ration clipping	4.4%
...	

Remarks on fANOVA

- Given compute higher-order interaction effects
  - Often too expensive for more than 2 or 3 dimensions

Remarks on fANOVA

- Given compute higher-order interaction effects
  - Often too expensive for more than 2 or 3 dimensions
- Implicit assumption: the surrogate model models the space fairly well

Remarks on fANOVA

- Given compute higher-order interaction effects
  - Often too expensive for more than 2 or 3 dimensions
- Implicit assumption: the surrogate model models the space fairly well
- Global analysis and local analysis of hyperparameter importance does not always agree [Biedenkapp et al. 2018]

Remarks on fANOVA

- Given compute higher-order interaction effects
  - Often too expensive for more than 2 or 3 dimensions
- Implicit assumption: the surrogate model models the space fairly well
- Global analysis and local analysis of hyperparameter importance does not always agree [Biedenkapp et al. 2018]

$\rightsquigarrow$ You should run both to get a good understanding of why an AutoML tool chose a configuration