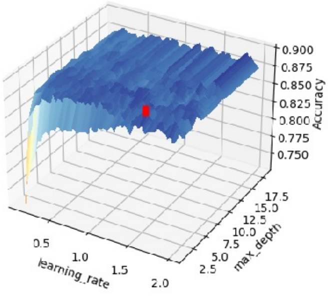
Optimization in Machine Learning

Optimization Problems

Other optimization problems

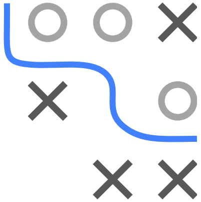


(a) vehicle

Learning goals

- Discrete / feature selection
- Black-box / hyperparameter optimization
- Noisy
- Multi-objective

OTHER CLASSES OF OPTIMIZATION PROBLEMS



So far: "nice" (un)constrained problems:

- Problem defined on continuous domain $\mathcal{S}$
- Analytical objectives (and constraints)

Other characteristics:

- Discrete domain $\mathcal{S}$
- f **black-box**: Objective not available in analytical form
  computer program to evaluate
- f **noisy**: Objective can be queried but evaluations are noisy
  $f(\mathbf{x}) = f_{\text{true}}(\mathbf{x}) + \epsilon, \quad \epsilon \sim F$
- f **expensive**: Single query takes time / resources
- f multi-objective: $f(\mathbf{x}) : \mathcal{S} \to \mathbb{R}^m$, $f(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_m(\mathbf{x}))$

These make the problem typically much harder to solve!

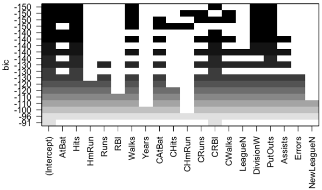
EXAMPLE 1: BEST SUBSET SELECTION

Let $\mathcal{D} = \left(\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right)_{1 \leq i \leq n}$, $\mathbf{x}^{(i)} \in \mathbb{R}^p$. Fit LM based on best feature subset.

$$\min_{\boldsymbol{\theta} \in \Theta} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}\right)^2, ||\boldsymbol{\theta}||_0 \leq k$$

Problem characteristics:

- White-box: Objective available in analytical form
- Discrete: $\mathcal{S}$ is mixed continuous and discrete
- Constrained

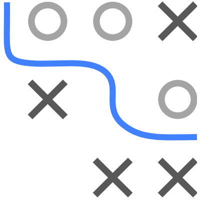


Click for source

The problem is even **NP-hard** (Bin et al., 1997, The Minimum Feature Subset Selection Problem)!

EXAMPLE 2: WRAPPER FEATURE SELECTION

Subset sel. can be generalized to any learner $\mathcal{I}$ using only features $\boldsymbol{s}$:



$$\min_{\mathbf{s}\in\{0,1\}^p} \widehat{\mathrm{GE}}(\mathcal{I}, \mathcal{J}, \rho, \boldsymbol{s}),$$

$\widehat{\mathrm{GE}}$ general. err. with metric ρ and estim. with resampling splits $\mathcal{J}$

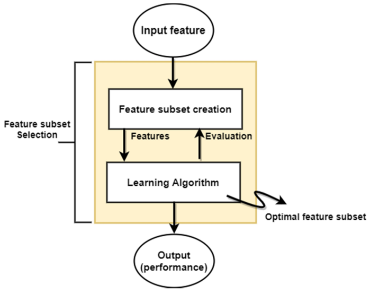
Problem characteristics:

- black box

  eval by program
- $\mathcal{S}$ is discrete / binary
- expensive

  1 eval: 1 or multiple ERM(s)!
- noisy

  uses data / resampling
- NB: Less features can be better in prediction (overfitting)



EXAMPLE 3: FEATURE SEL. (MULTIOBJECTIVE)

Feature selection is usually inherently multi-objective, with model sparsity as a 2nd trade-off target:

$$\min_{\mathbf{s}\in\{0,1\}^p} \left(\widehat{\mathrm{GE}}(\mathcal{I}, \mathcal{J}, \rho, \boldsymbol{s}), \sum_{i=1}^{p} s_i\right) .$$

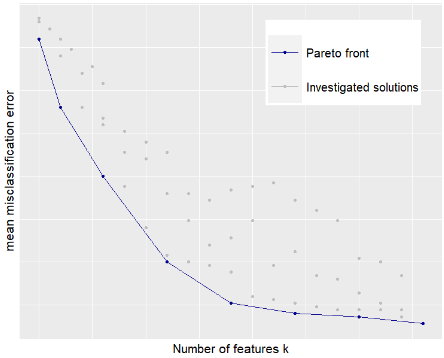
$\widehat{\mathrm{GE}}$ general. err. with metric ρ and estim. with resampling splits $\mathcal{J}$

- Multiobjective
- black box

  eval by program
- S is discrete / binary
- expensive

  1 eval: 1 or multiple ERM(s)!
- noisy

  uses data / resampling

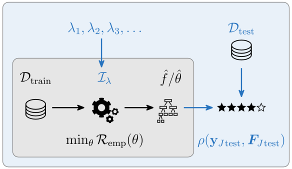


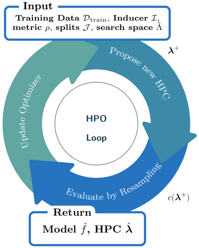
EXAMPLE 4: HYPERPARAMETER OPTIMIZATION

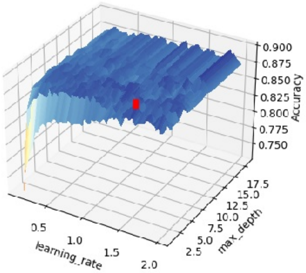
- Learner $\mathcal{I}$ usually configurable by hyperparameters $\boldsymbol{\lambda} \in \Lambda$
- Find best HP configuration $\boldsymbol{\lambda}^*$

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \Lambda} c(\boldsymbol{\lambda}) = \arg\min \widehat{\mathrm{GE}}(\mathcal{I}, \mathcal{J}, \rho, \boldsymbol{\lambda})$$

$\widehat{\mathrm{GE}}$ general. err. with metric ρ and estim. with resampling splits $\mathcal{J}$







(a) vehicle

▸ Click for source

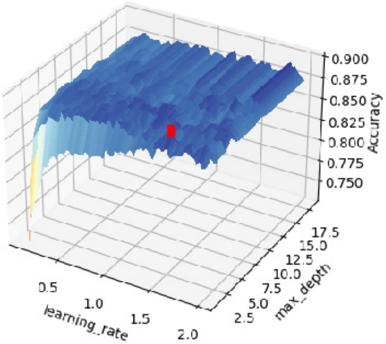
EXAMPLE 4: HYPERPARAMETER OPTIMIZATION

Solving

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda}\in\Lambda} c(\boldsymbol{\lambda})$$

is very challenging:

- c black box
  eval by progrmm
- expensive
  1 eval: 1 or multiple ERM(s)!
- noisy
  uses data / resampling
- the search space Λ might be mixed
  continuous, integer, categ. or hierarchical



(a) vehicle

▸ Click for source