AutoML
Automating the hyperparameter search
Lyle Ungar

Learning objectives
Auto-SKlearn
Ensemble selection
Auto-Sklearn

- 15 Classifiers
- 14 feature preprocessing methods
- 4 data preprocessing methods
  → 110 hyperparameters

Combined Algorithm Selection and Hyperparameter (CASH) Optimization
Preprocessing & Methods

- extreme. rand. trees prepr. feature selection
- fast ICA
- feature agglomeration
- kernel PCA
- rand. kitchen sinks (random projection)
- linear SVM prepr. L1 feature selection
- no preprocessing (random projection)
- nystroem sampler
- PCA
- polynomial
- random trees embed.
- select percentile
- select rates
- one-hot encoding
- imputation
- balancing
- rescaling
- AdaBoost (AB)
- Bernoulli naïve Bayes
- decision tree (DT)
- extreme. rand. trees
- Gaussian naïve Bayes
- gradient boosting (GB)
- kNN
- LDA
- Linear Discriminant analysis
- linear SVM
- kernel SVM
- multinomial naïve Bayes
- passive aggressive
- QDA
- Quadratic Discriminant analysis
- random forest (RF)
- Linear Class. (SGD)
Auto-Sklearn

- **Warmstart/Metalearning**: Start from hyperparameters that worked in the past for similar datasets.
  - Based on 38 metafeatures of 140 datasets

- **Uses Bayesian optimization**
  - Fit a random forest model predicting performance from hyperparameters and use it to find the optimum
    - speed up by discarding values that look bad on the first fold of 10-fold CV

- **Use Ensembles** of the 50 best classifiers considered
Ensemble selection

- Greedy (stagewise)
- start from an empty ensemble
- iteratively add the model that minimizes ensemble validation loss
  - with uniform weight, but allowing for repetitions
- Why not optimize the weights on each model?
Metafeatures

- **Number of features & observations**
  - With transformations
  - Number and percentage missing
  - Number real or categorical

- **Class probability stats**
  - Min, max, entropy…
Auto-Sklearn performance

- Performance (with limited CPU) was third among a large set of human competitors
AutoML using Metadata Language Embeddings

- Use text description of problems to pick hyperparameters
  - Use vector embeddings of dataset title, description and keywords
  - For each new dataset, find most similar prior dataset and use its hyperparameters
  - The similarity metric is learned (supervised)

Drori et al 2019
Conclusions

◆ **AutoML is close to best humans**
  - And less likely to overfit
  - Different ensembles for different problem types

◆ **To really avoid overfitting, do nested CV**
  - For each of ten folds, on the 90%
    - Do 10-fold CV to find the best method
  - Observe performance on the held-out 10%