Education presentation (20 min.)

Paolo Toccaceli

Centre for Reliable Machine Learning
Royal Holloway, University of London

https://cml.rhul.ac.uk/people/ptocca/HomePage/

March 2021

This slide set is available at:
https://ptocca.github.io/UoE-2021/Education_Presentation.pdf

Feature selection

- ML setting: data set with ℓ examples z_i, each a pair of an object $\mathbf{x}_i$ and a label $y_i \in Y$
  $\mathbf{x}_i$ is a vector of p features $x_{ij} \in X_j$, where X_j and Y can be continuous ($\mathbb{R}$, $\mathbb{Z}$, or a subset) or a generic discrete set
- **Objective**: building a model with satisfactory performance (e.g. accuracy, F1-score, etc.) on a subset of the existing features.
- Different from feature extraction, feature engineering, dimensionality reduction
- A wide variety of techniques are available, to cater for the wide variety of problems

Motivation

- Some features may be completely irrelevant
- Faster model convergence, fewer training examples needed to achieve the same error
- Curse of dimensionality
- Over-parameterized models may give the illusion of higher accuracy
- The cost of collecting data
- Potentially better interpretability

Types of feature selection methods

- Wrappers: generic combinatorial optimization approach
- Embedded: take advantage of efficient estimates of relevance offer by some ML methods
- Filters: model-independent preprocessing step, based on estimation of informational content and mutual dependence

Wrappers

- Method-agnostic: the ML method is treated as a black box
- FS is framed as a problem of combinatorial optimization: Identify the choice of variables resulting in the best metric value
- The brute force exploration of the space of combinations is unfeasible
  - In principle, FS is an NP-complete problem
- Many algorithms available from combinatorial optimization
- Greedy approach
  - Step-wise Forward selection:
    start from an initial selection and add feature(s) that improve the target metric the most
  - Step-wise Backward selection:
    start from the full set of features and remove feature(s) that degrade the target metric the least

Embedded

- Some methods provide by-products relevant for FS
- Random Forests and Regularized methods
    - Random Forests:
      Efficient estimates of Variable Importance
    - Regularized methods
      Feature weights shrink as regularization increases

FS with Random Forests

- Estimate of Variable Importance: Mean Decrease in Impurity
- MDI computes the average reduction in loss or impurity contributed by a given feature k.
  - For each tree:
    - Find all the nodes in which the split was on feature k.
    - Sum the decrease in impurity, weighted by the fraction of samples in the node.
  - Compute average across trees
- Interpretation can be tricky
  - Low values do correspond to low actual importance
  - High values do not necessarily reflect actual importance (*feature selection bias*)
    - Mitigated with judicious choice of maximum depth, minimum leaf size

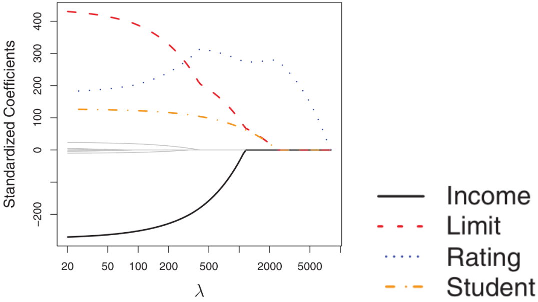
FS with LASSO

- Regularized methods: loss function contains a regularization term, a penalty for “model complexity”
- Let’s assume $f(\mathbf{x}_i) = w_0 + \sum_{j=1}^{p} w_i x_i$
- A regularized model chooses parameters $\mathbf{W}$ to minimize

$$\sum_{i=1}^{\ell} L(y_i, f(\mathbf{x}_i)) + \lambda \cdot \Omega(\mathbf{W})$$

- When the complexity penalty is an L1 metric, i.e. $\Omega(\mathbf{W}) = \sum_{j=0}^{p} |w_i|$, as in LASSO, some weights go to zero as the regularization coefficient λ increases.
- λ is usually chosen on the basis of *bias-variance* tradeoff (or balance between underfitting and overfitting)

FS with LASSO



Paths of the variable coefficients as λ varies in a LASSO model for the `credit` data set.

Chart taken from Fig 6.6 in James et al., An Introduction to Statistical Learning

- The variables with null coefficients can be removed from the model.
- With L2 metric, the weights “shrink” towards zero, but do not reach zero.
  - One can remove those below a threshold

Filters

- Rank and select feature based on estimates of their information content and dependence with the label; preprocessing independent of ML method
- **Variance thresholding**
  - Compute the (empirical) variance of each variable
  - Remove the variables with variance below a chosen threshold
- **Correlation**
  - Between features:
    One of two highly correlated features is redundant
  - With the label
    Features with high correlation with the label are desirable but...
    lack of correlation does not imply independence!

From Wikipedia entry "Correlation and dependence"

Filters

- **Mutual Information**
- The mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)}\right)$$

- An information theory notion, with solid theoretical backing
- Estimating densities can be challenging
- Looking at single variables can be misleading
  - Variables that have no predictive value on their own can become relevant in combination (XOR example)

Pragmatic considerations

- We showed only a few simple methods.
- First, use domain knowledge to
  - Identify irrelevant features
  - Investigate interdependence of features
  - Prioritize features to keep and to discard
- Then, use the FS techniques, possibly starting with the simplest

Recap

- Feature selection aims at building a model with satisfactory performance on a subset of the existing features
- Three main classes of FS techniques
  - **Wrappers**: generic combinatorial optimization approach
  - **Embedded**: take advantage of efficient estimates of relevance offer by some ML methods
  - **Filters**: model-independent preprocessing step, based on estimation of informational content and mutual dependence

References

- James, Witten, Hastie, Tibshirani
  An Introduction to Statistical Learning
  Springer, 2013
  Available online at: `https://www.statlearning.com/`
- Guyon, Elisseeff
  An Introduction to Variable and Feature Selection
  Journal of Machine Learning Research 3 (2003) 1157-1182
- Li, Wang, et al.
  A Debiased MDI Feature Importance Measure for Random Forests
  NeurIPS 2019, Vancouver, Canada