Interpreting Feature Selection

Ah Young (Amy) Kim

12/09/2018

Feature Selection Module: Table of Contents

What is Feature Selection?

Feature selection

- A sub-field in statistics and machine learning
- Variable selection and feature engineering used interchangeably
- Focuses on identifying the most important elements in a model
- Makes a model more interpretable
- The **most critical area** of CDSS where clinical input is heavily required

Why is Feature Selection so Important?

- We want to remove redundant and/or irrelevant elements to make a model the simplest as possible
- It can save time and/or money to train data with fewer elements in a model
- It can improve prediction accuracy and generalization

Since NLP scientists, statisticians, and computer scientists have minimal (or no) clinical training, clinical input must be required when developing models in a CDSS!

How Does Feature Selection Work?

- In general, an algorithm tests possible subsets of features (elements) and finds the one that minimizes the error rate
- However, it really depends on the algorithm, and different algorithms may choose a different set of features
- Again, clinical input is required in order to determine if the model makes sense, whether the selected features are useful, and what interactions are clinically insightful

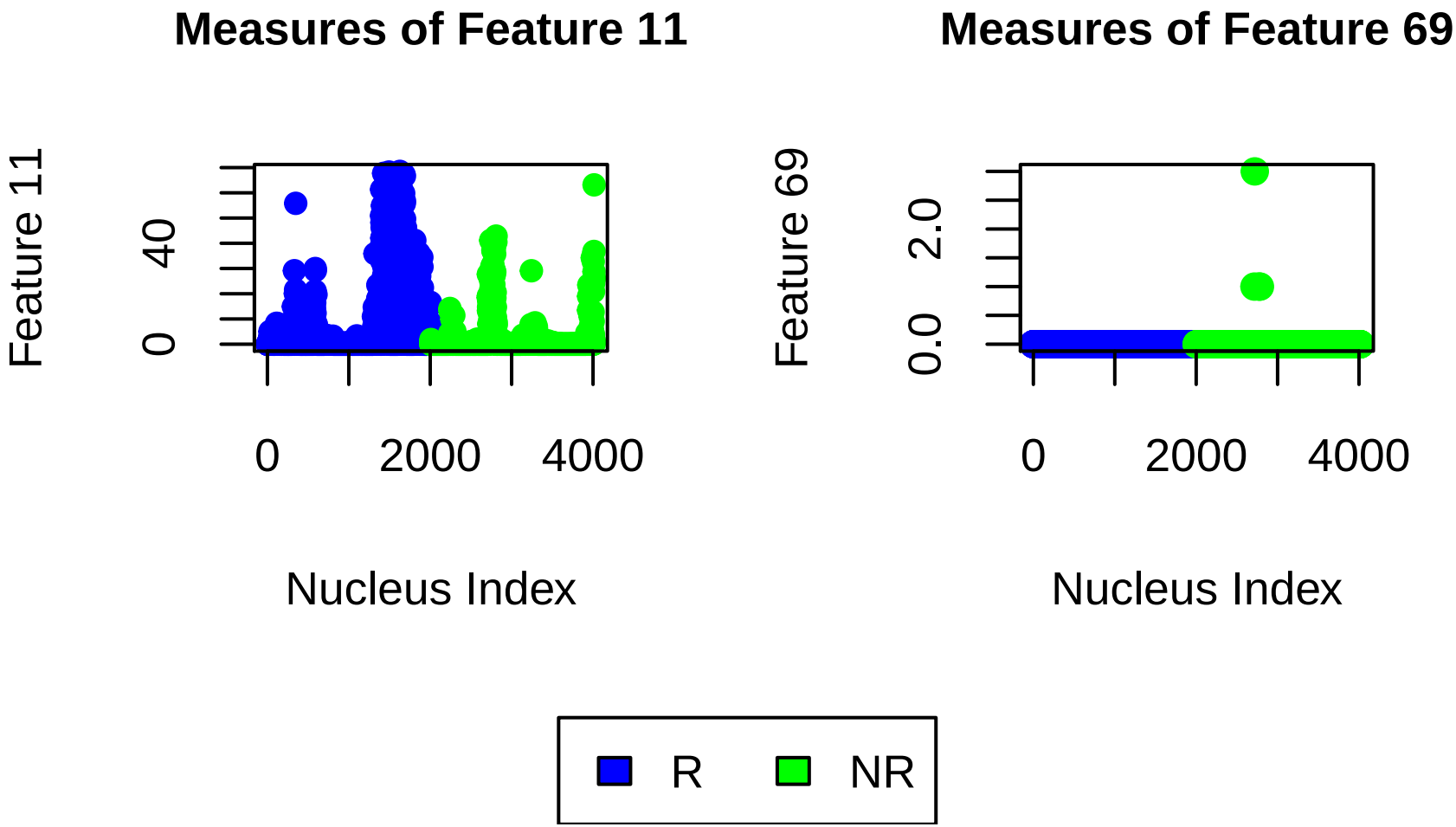
Bladder Cancer Data Example

Example Given 93 nuclear chromatin features of 40 patients who had recurrence or no recurrence of bladder cancer, which features can best predict whether a patient will have a recurrence or not?

- 20 patients are in recurrence group (R), and 20 patients are in no recurrence group (NR)
- For each patient, about 100 nuclei have measures for all 93 features
- Although there are 93 features, excluding features with all zero values, 72 features remain
- At the end, there are 4016 observations and 72 features

Discriminant Ability of Features

Illustration The plots below show that feature 11 measures are clearly different for the R and NR groups, but feature 69 measures are mostly zeros except for a few non-zero values for the NR group. Therefore, feature 11 may have a better discriminant ability



Feature Selection Embedded Learning Algorithms

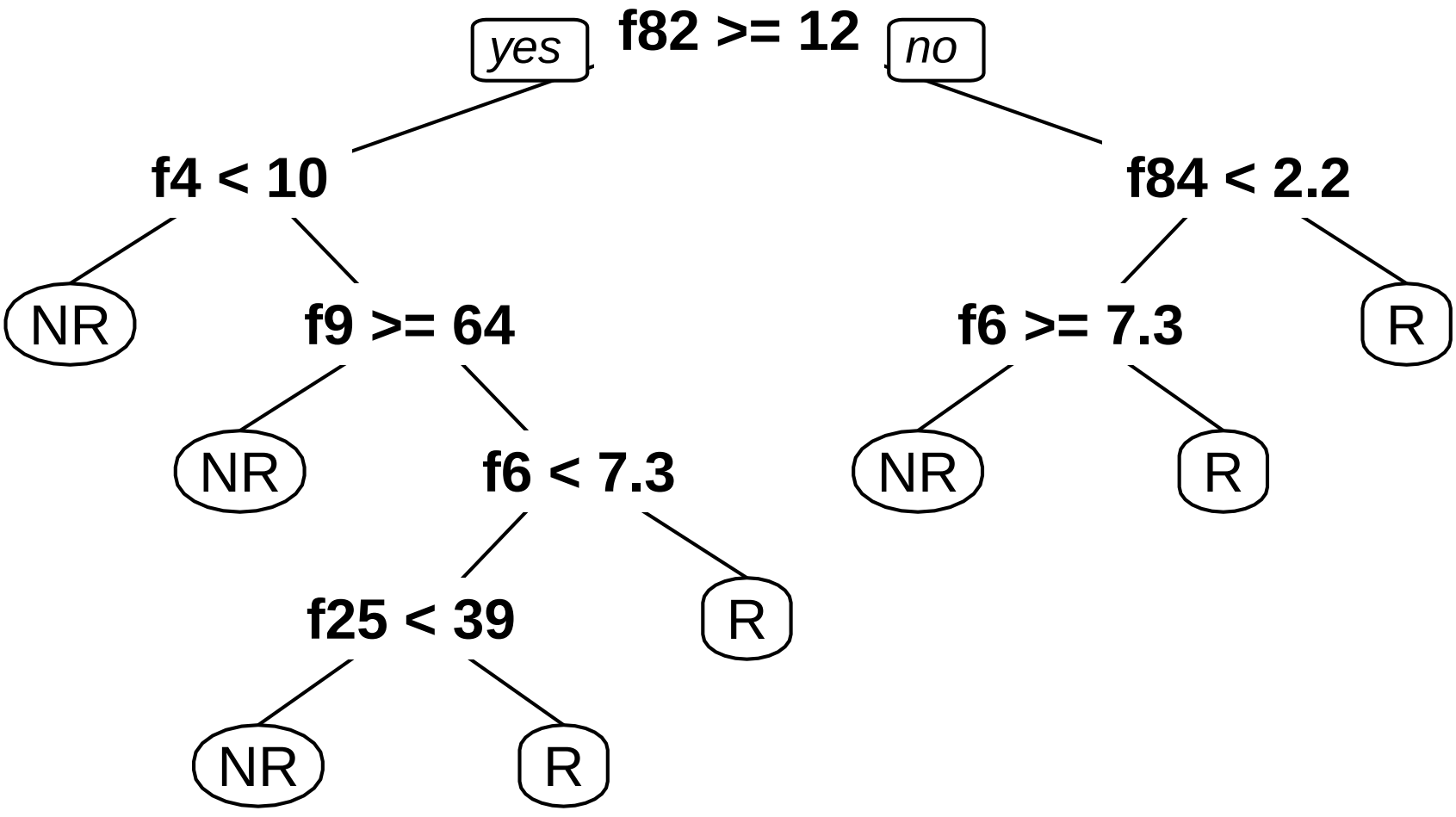
- The linear discriminant analysis (LDA), least absolute shrinkage and selection operator (Lasso), elastic net, support vector machine (SVM), trees, artificial neural network (ANN) are used to demonstrate feature selection using the bladder cancer data
- For some of these algorithms, feature selections are embedded, and different algorithms choose different features as important elements

Algorithms	# of Chosen Features
LDA	72
Lasso	45
Elastic Net	35
SVM	72
Tree	6
ANN	72

Feature Selection Example: Tree Method

- In a decision tree, a set of features that can partition the data is considered
- Given that the first feature that best splits the data is chosen, the tree finds another set of features that can partition the data again
- The very top node is the **Root node**, which represents the entire sample
- A **decision node** is any sub-node that splits into further sub-nodes
- A **terminal node** is any node at the very end of the tree, which does not split anymore
- In the bladder cancer data, each decision node represents one of the chosen features, and depending on the value of the feature at the decision node, a given data is classified as "NR" or "R"

Feature Selection Example: Tree Diagram



Bladder Cancer Example: Which Features Should Be Included in the Model?

- Assume each feature description is given
- The elastic net chooses features 3, 4, 6, 7, 9, 10, 24, 25, 26, 30, 31, 35, 39, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 83, 85, 89, 90, 91, and 92 as important elements to discriminate "R" and "NR"
- The tree algorithm only chooses features 4, 6, 9, 25, 82, and 84 as important elements
- Features 82 and 84 are not common in these two algorithms, but why?
- Developers will not know whether some chosen features in the elastic net are related, whether there are any features that are irrelevant but chosen by chance, or whether all features chosen by the tree method are enough to make accurate predictions Clinical input must be made in order to decide which features are really important and clinically relevant!

Prediction Performances of Learning Algorithms

- Depending on algorithms, prediction performances vary
- Prediction performances also vary depending on the given data set

Algorithm	Training Error	Test Error
LDA	0.1924	0.4459
Lasso	0.1900	0.4379
Elastic Net	0.1979	0.4419
SVM	0.0950	0.4564
Tree	0.1890	0.4848
Random Forest	0	0.4279
ANN	0.3138	0.3155

References

[1] Le, J. (2018). Decision Trees in R. DataCamp Community. https://www.datacamp.com/community/tutorials/decision-trees-R.

[2] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.