UNIVERSITY of WASHINGTON

LIMEaid

Local Interpretable Model-agnostic Explanations (LIME)

Data 515, Spring 2019
M.S. Data Science

Suman Bhagavathula
Patrick King
Javier Salido

Interpretability in Machine Learning

Some highly accurate models are not "explainable"

- Neural networks, random forests

Why is this a problem?

- Bias, not obvious
- High test set accuracy but poor results in the field
- Policy or law demands an explanation of any decision

Solution: model-agnostic local explanations

- Explain one instance, not entire model
- Fit a simple model to explain a small section of decision space

LIMEaid: A LIME solution for tabular data

LIMEaid explanations

Input

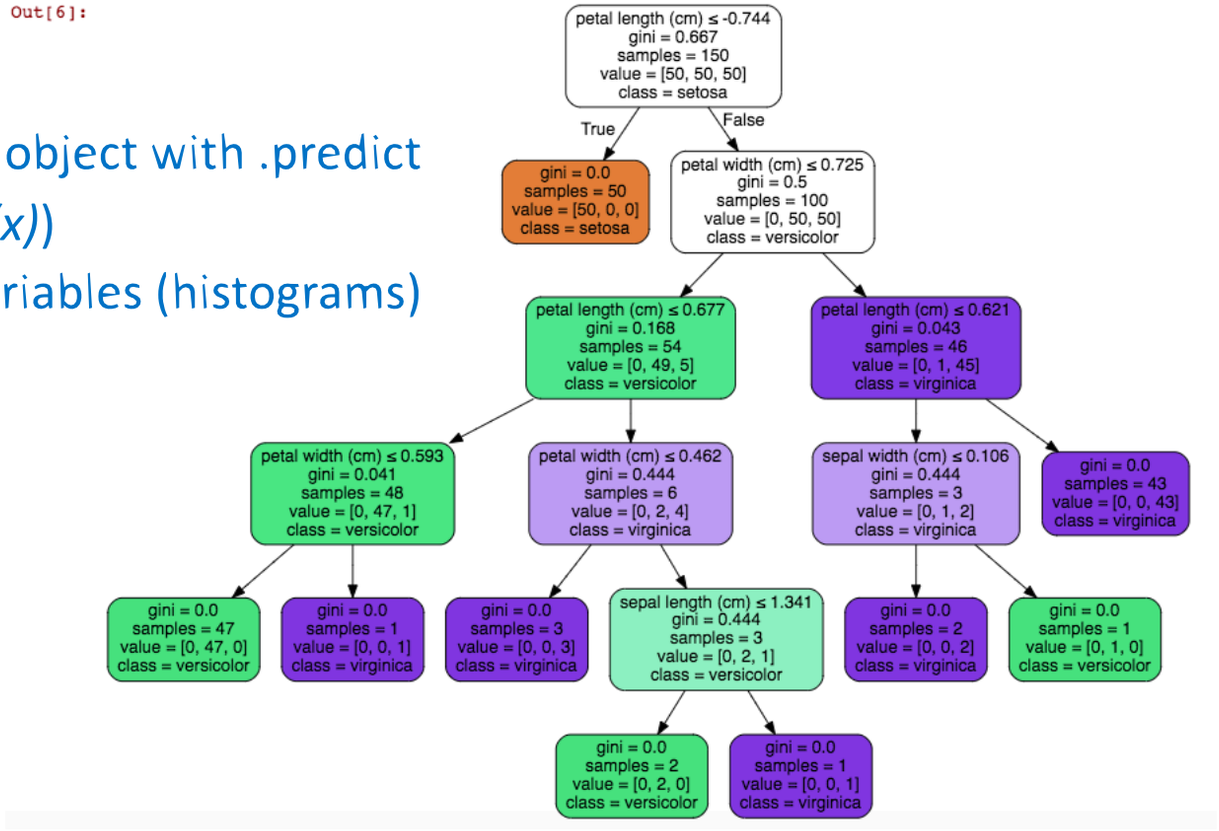
- A “complex” ML model, fit by sklearn classifier object with .predict
- An instance of data (*x*) and its model output (*f(x)*)
- Probability domain for normalized predictor variables (histograms)

Output

- Sparse linear models (few features), plottable
  List significant features

Analysis/verification

- Decision-tree comparison



Use Cases

User profile: data scientist with Python programming experience

1. **Model verification scenario**
   - User wants to preempt poor model performance “in the field”
   - Use LIMEaid to sample test dataset
   - Show most significant features for decision
   - Tune or replace model if spurious correlation or other issues

2. **Decision explanation scenario**
   - Classification has already been made by a model
   - Use LIMEaid to sample whole dataset
   - Produce easy-to-share “two-dimensional” plot of a linear correlation

Design

Model

lime_sample

lime_fit

lime_result

Data

Sources

- College Scorecard (data.gov): Annual report of schools and attributes (SAT scores, majors offered, region, cost, public/private/for profit, etc.)
- "Where it Pays to Attend College" (Kaggle.com) obtained from (*Wall Street Journal*), based on Payscale, Inc. (College Salary Report Methodology): Article reporting schools and salaries of graduates, salaries by major, etc.

Merge

- Significant cleaning, reformatting to match sets on college name
- String manipulation, removal of hyphens, abbreviations, region names, etc.

More

- Also tested with Sklearn's provided "Iris" data, to show comparison to Ribeiro's original LIME package ("LIME classic")

Models

Scikit-learn classifiers that predict probabilities (predict_proba implemented)

Multiclass logistic regression (sklearn.linear_model.LogisticRegression)

- 85% accuracy on College data

Random Forest (sklearn.ensemble.randomforestclassifier)

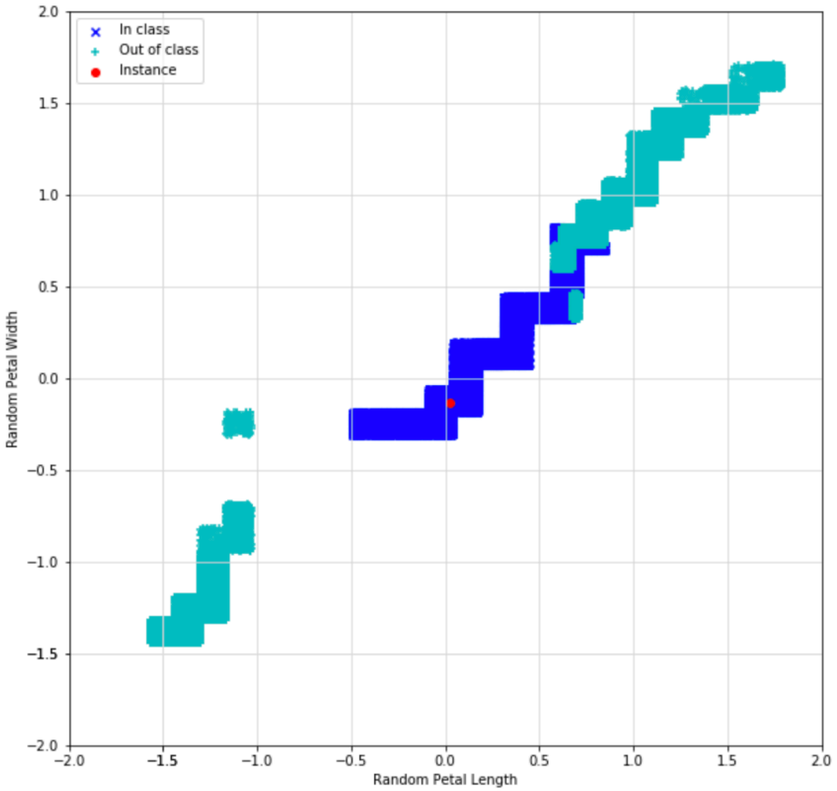
- 65% accuracy on College data

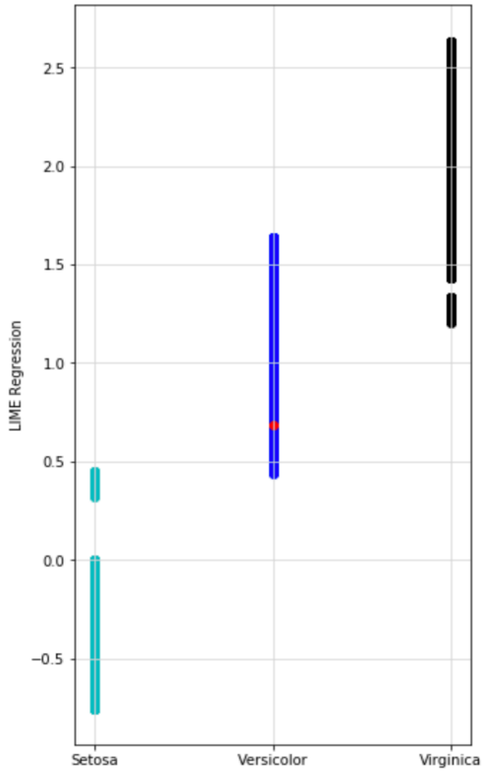
- Decision tree (sklearn.tree)
  65% accuracy on College data

Models not tuned for improved accuracy (default settings)

Demo

Iris notebook (LIME_Iris_ex_notebook.ipynb)





Project Structure

- LIMEaid Github Repo based on Shablona
- Used “M-V-C” architecture to codebase (Model-View-Controller)

Branch: master ▾ LIMEaid / LIMEaid / Create new file | Upload files | Find file | History

sumanbhagavathula Merge pull request #14 from PKing70/suman/testforversionpy ··· Latest commit a5bb620 23 hours ago

..		
controller	fix module import issue	yesterday
data	Renamed codebase folder to LIMEaid	yesterday
model	fix module import issue	yesterday
unittests	Merge pull request #14 from PKing70/suman/testforversionpy	23 hours ago
__init__.py	Renamed codebase folder to LIMEaid	yesterday
version.py	updated all references of codebase and DATA515-LIME to LIMEaid	yesterday

Future Work

- **API support** for data acquisition to support dynamic features:
  - College Scorecard (data.gov) currently published with new features and data dictionary yearly
- **Model tuning** for examples:
  - Currently using defaults, could improve accuracy > 65%
- **Modify penalty** for number of coefficients
- **More data type and model support:** image data, NLP support, support for model objects beyond sklearn classifiers

Questions?