VistaMilk

A World Leading SFI Research Centre

Science Foundation Ireland SFI For what's next

MLGIG Team – Tabular Approaches as Baselines

Georgiana Ifrim

Associate Professor

School of Computer Science, UCD

VistaMilk International Workshop on Spectroscopy and Chemometrics 2021

HOST INSTITUTION

Teagasc Agriculture and Food Development Authority

PARTNER INSTITUTIONS

DCU, UCD Dublin, NUI Galway OE Gaillimh, Waterford Institute of Technology, Tyndall National Institute, ICBF.com, UCC University College Cork, Ireland, TSSG, Insight

FUNDED BY:

An Roinn Talmhaíochta, Bia agus Mara, Department of Agriculture, Food and the Marine; Ireland's European Structural and Investment Funds Programmes 2014-2020; European Union, European Regional Development Fund; Science Foundation Ireland SFI For what's next

About Me

2020 - Now	Associate Professor, School of Computer Science, University College Dublin, Ireland co-Director ML-Labs CRT; Funded Investigator with Insight Centre for Data Analytics & VistaMilk SFI Research Centre
2015 - 2020	Assistant Professor, CS, UCD
2009 - 2015	Postdoc Researcher, Aarhus University, Denmark; UCC, UCD, Ireland
2005 - 2009	PhD, Computer Science (topic: Machine Learning), Max Planck Institute for Informatics, Germany
2003 - 2005	MSc, Computer Science, Max Planck Institute for Informatics, Germany
1999 - 2003	BSc, Computer Science, University of Bucharest, Romania

About MLGIG Team: Thach (postdoc VistaMilk), Bhaskar (PhD VistaMilk), Ashish (PhD Insight), Antonio (postdoc Insight).
We all have CS/ML background, and little knowledge about Spectroscopy.

Spectroscopy Challenge - Data Cleaning

- **Remove NaN rows before training:** only keep training samples with known targets
- **Remove outliers:** Based on Maria Frizzarin et al's paper, only keep training rows with **target <= mean + 3 std** for that target.

- Use **4-fold cross-validation** for comparison of results of different algorithms; also look at a **single train-test** split to check what the model learns, and where possible, what are the important features. Compare cv to single split results.
- Tuning models via cv: grid search. **Retrain best model on full labelled training** data and predict on unlabeled test set. Submit predictions, one column per target.

- Other things we tried, but did not work:
  - transforming the waves from transmittance to absorbance by log10(1/wave)
  - taking log of target

Modeling Strategies

Tabular Models:

- Treat each sample as a vector of independent features, without consideration to the ordering of the waves.
- Baseline for more sophisticated approaches that extract time series features (where the ordering of waves is considered).
- Select tabular models from **different categories of learning strategies**, e.g., from simple **linear models** (Linear Regression), to **regularized linear models** (Ridge Regression, Lasso Regression), to **non-linear models** (ensembles, neural networks).
- For linear models, mostly rely on regularisation to deal with noisy and correlated features (no fancier feature selection or dimensionality reduction methods).

Modeling Strategies

We evaluated the following tabular models:

Linear models:
LinearRegression(),
PLSRegression(),
Ridge(),
RidgeCV(),
RidgeCV(normalize=True),
RidgeCV(alphas=np.logspace(-2, 2, 10), normalize=True),
Lasso(),
Lasso(normalize=True),
ElasticNet()
SVR(kernel='linear'),

Ensembles:
Bagging:
RandomForestRegressor(n_estimators=100),

Boosting:
GradientBoostingRegressor(n_estimators=100)
A few other variants of gradient boosting, e.g., Xgboost and LightGBM, not better than GBR.

Not better than linear models.

Other:
KNeighborsRegressor(n_neighbors=1)
SVR(kernel='rbf')

MLPRegressor()
deep learning: FCN, Resnet

Not better than linear models.

Results

Cross-validation RMSE for best models. Standard deviation in brackets.

Method	Kappa-RMSECV	Micelle-RMSECV	Ph-RMSECV
LASSO(normalize=True)	1.5117 (0.04)	56.7817 (19.35)	0.1188 (0.02)
RidgeCV-tabular(normalize=True)	**1.1697** (0.06)	**57.1684** (20.21)	**0.0821** (0.01)
MiniROCKET	1.1863	60.4678	0.0811
EnsembleMiniROCKET	1.1740	58.3019	0.0768

LASSO just predicts the average. All feature weights are zero, and the intercept is the average of the target on the training samples.

Thach will present the TS models based on ROCKET.

- Best CV model often predicts the average (eg LASSO comes top wrt RMSECV for some targets, but it simply predicts the average, all feature coefficients are zero).
- RidgeCV top method across targets.
- It is important to look at what the model learns on a single train-test split and how do the predictions look like.

Take-away

- Linear model RidgeCV seems to be learning something useful, but it underestimates large targets.
- Good baseline for more complex time series models, extremely fast to learn and tune.
- Can look at the learned RidgeCV model and feature weights to focus on some of the waves (parts of the time series).

- We need to read more about this application domain, especially for appropriate data cleaning. Explanation methods: e.g., with LIME, SHAP.
- Our code is available, and our experiments are reproducible (Jupyter notebook).
- Looking forward to learn more from the other teams.
- Let’s have a look at the code!