Same vs. other cross-validation in supervsied machine learning

Toby Dylan Hocking
toby.hocking@nau.edu

April 11, 2024

Introduction to machine learning

Proposed same vs. other cross-validation

Results on real data sets

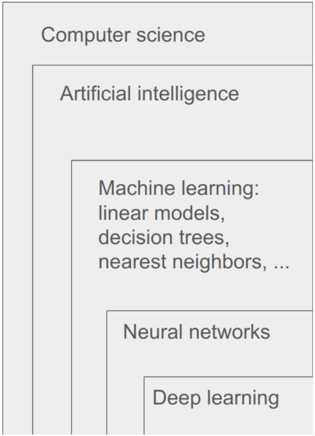
Results on machine learning benchmark data sets

Synthesis, Discussion and Conclusions

Supplementary slides

What is machine learning?

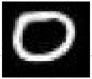
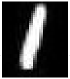
- Computer science: domain of study about efficient algorithms / computations.
- Artificial intelligence: sub-domain concerned with algorithms for accurate predictions/suggestions.
- Machine learning: sub-domain concerned with algorithms for large data.
- Machine learning is widely used in search engines, automatic translation, image analysis, ...



Machine learning intro: image classification example

ML is all about learning predictive functions $f(x) \approx y$, where

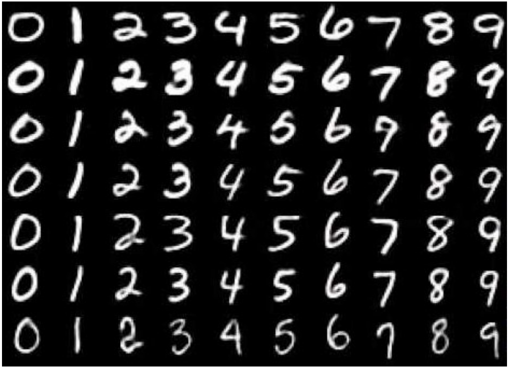
- Inputs/features x can be easily computed using traditional algorithms, e.g. matrix of pixel intensities in an image.
- Outputs/labels y are what we want to predict, easy to get by asking a human, but hard to compute using traditional algorithms, e.g. image class.
- Input $x =$ image of digit, output $y \in \{0, 1, \ldots, 9\}$,
  – this is a classification problem with 10 classes.

  $f($$) = 0$, $f($$) = 1$
- Traditional/unsupervised algorithm: I give you a pixel intensity matrix $x \in \mathbb{R}^{16 \times 16}$, you code a function f that returns one of the 10 possible digits. Q: how to do that?

Supervised machine learning algorithms

I give you a training data set with paired inputs/outputs, e.g.

$$y = 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9$$

$X =$ 

Your job is to code an algorithm that learns the function f from the training data. (you don't code f)

Source: github.com/cazala/mnist

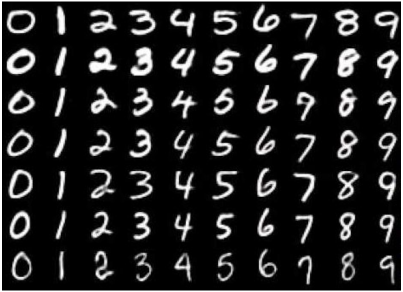
Supervised machine learning algorithms

Can be used whenever a knowledgeable/skilled human can easily/quickly/consistently create a large database of labels for training.

Should be used if it is not easy to code the function f for predicting the labels (using traditional/unsupervised techniques).

Accurate if the test data, on which you want to use f, is similar to the train data (input to learning algorithm).

Advantages of supervised machine learning

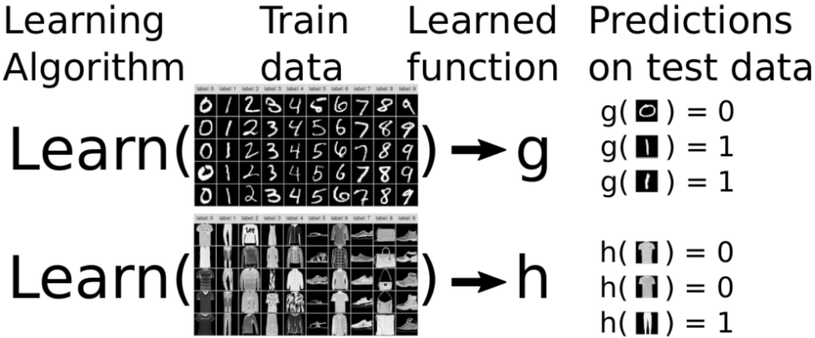


- Input $x \in \mathbb{R}^{16\times 16}$, output $y \in \{0, 1, \ldots, 9\}$ types the same!
- Can use same learning algorithm regardless of pattern.
- Pattern encoded in the labels (not the algorithm).
- Useful if there are many un-labeled data, but few labeled data (or getting labels is long/costly).
- State-of-the-art accuracy (if there is enough training data).

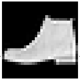
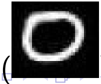
Sources: github.com/cazala/mnist, github.com/zalandoresearch/fashion-mnist

Learning two different functions using two data sets

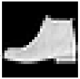
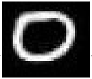
Figure from chapter by Hocking TD, *Introduction to machine learning and neural networks* for book *Land Carbon Cycle Modeling: Matrix Approach, Data Assimilation, and Ecological Forecasting* edited by Luo Y (Taylor and Francis, 2022).



Learn is a learning algorithm, which outputs g and h.

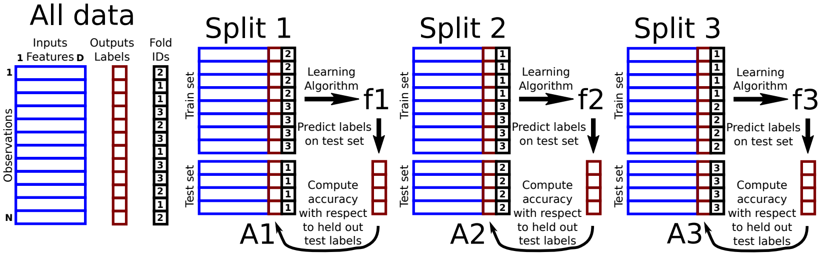
Q: what happens if you do $g($$)$, or $h($$)$?

Learning two different functions using two data sets

- What if you do $g($$)$, or $h($$)$?
- This is a question about **generalization**: how accurate is the learned function on a new/test data set?
- "Very accurate" if test data are similar enough to train data (best case is i.i.d. = independent and identically distributed)
- Predicting childhood autism (Lindly *et al.*), train on one year of surveys, test on another.
- Predicting carbon emissions (Aslam *et al.*), train on one city, test on another.
- Predicting presence of trees/fires in satellite imagery (Shenkin *et al.*, *Thibault et al.*), train on one geographic area/image, test on another.
- Predicting fish spawning habitat in sonar imagery (Bodine *et al.*), train on one river, test on another.
- But how do we check if "very accurate" in these situations?

K-fold cross-validation: a standard algorithm used to estimate the prediction accuracy in machine learning

- $K = 3$ folds shown in figure below, meaning three different models trained, and three different prediction/test accuracy rates computed.
- It is important to use several train/test splits, so we can see if there are statistically significant differences between algorithms.



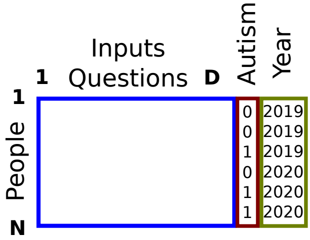
Hocking TD *Intro. to machine learning and neural networks* (2022).

Example data set: predicting childhood autism

- Collaboration with Lindly *et al.*
- Downloaded National Survey of Children's Health (NSCH) data, years 2019 and 2020, from `http://www2.census.gov/programs-surveys/nsch`
- One row per person, one column per survey question.
- Pre-processing to obtain common columns over the two years, remove missing values, one-hot/dummy variable encoding.
- Result is $N = 46,010$ rows and $D = 366$ columns.
- 18,202 rows for 2019; 27,808 rows for 2020.
- One column is diagnosis with Autism (binary classification, yes or no), can we predict it using the others?
- Can we combine data from different years?
- Can we train on one year, and accurately predict on another?

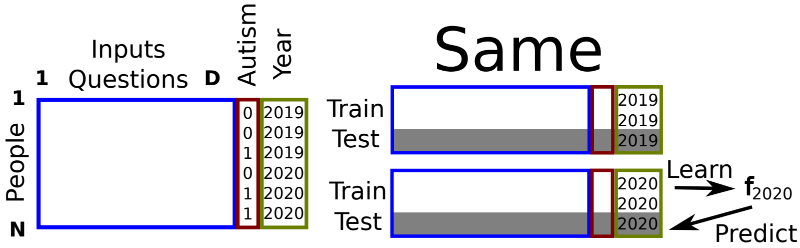
Proposed Same Other Cross-Validation

- Example: childhood autism prediction data set.



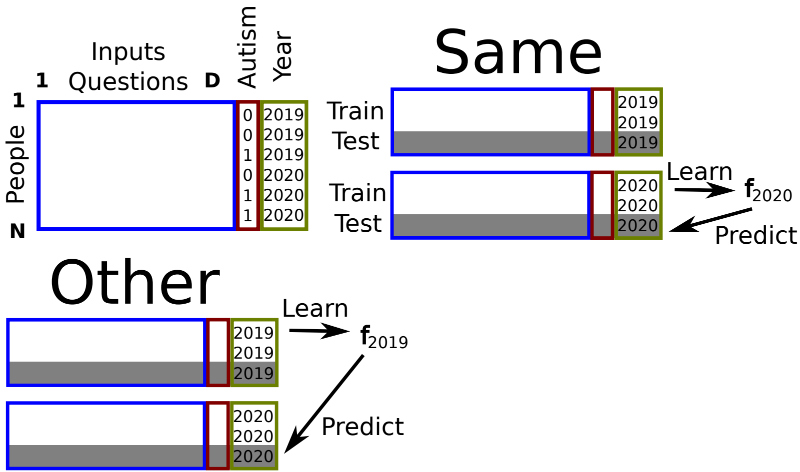
Proposed Same Other Cross-Validation

- Train group same as test (=regular K-fold CV on 2020).



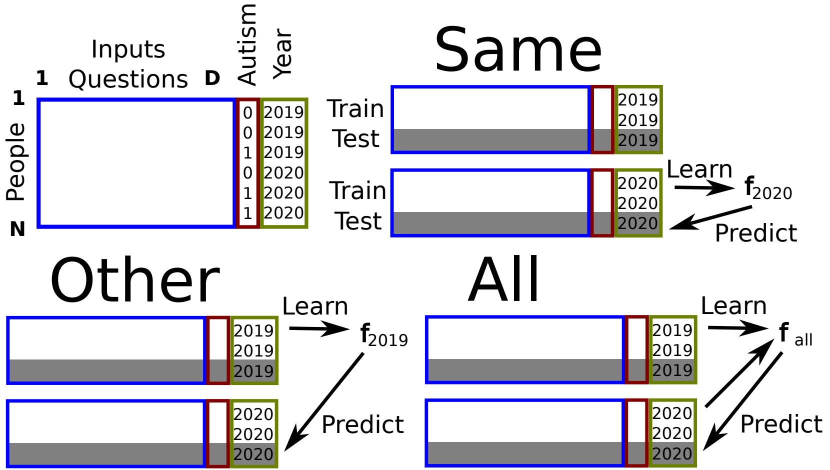
Proposed Same Other Cross-Validation

- Train group (2019) different from test (2020).



Proposed Same Other Cross-Validation

- Repeat for each of K folds, and each test group (2019,2020).



Proposed Same Other Cross-Validation

For a fixed test set from one group:
If train/test are similar/iid,

- **All** should be most accurate.
- **Same/Other** should be less accurate, because there is less data available (if other is larger than same, then other should be more accurate than same, etc).

If train/test are different (not iid),

- **Same** should be most accurate.
- **Other** should be substantially less accurate.
- **All** accuracy should be between same and other.

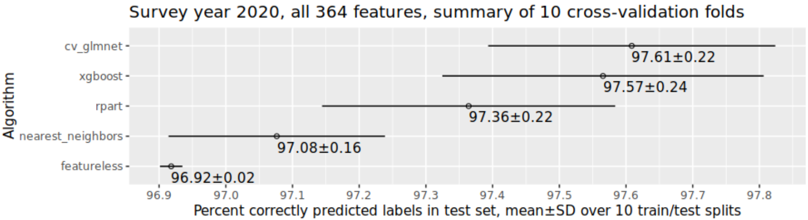
Learning algorithms we consider

We used the following learning algorithms:

- **cv_glmnet** L1-regularized linear model (feature selection). Friedman, *et al.* (2010).
- **xgboost** Extreme gradient boosting (non-linear). Chen and Guestrin (2016).
- **rpart** Recursive partitioning, decision tree (non-linear, feature selection). Therneau and Atkinson (2023).
- **nearest_neighbors** classic non-linear algorithm, as implemented in kknn R package. Schliep and Hechenbichler (2016).
- **featureless** un-informed baseline, ignores all inputs/features, and always predicts the most frequent label in train data. For example, Autism=No. Nomenclature from mlr3 R package, Lang, *et al.*, (2019).

Each learning algorithm has different properties (non-linear, feature selection, etc). For details see Hastie, *et al.* (2009) textbook.

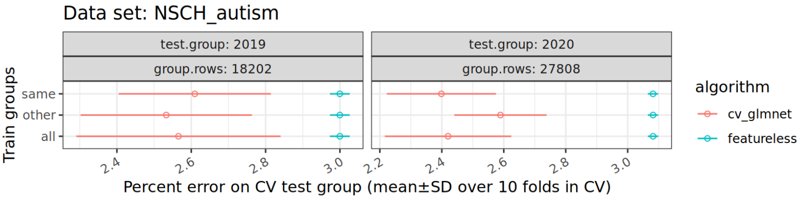
K-fold CV on NSCH data (predict autism), year 2020



Learning algorithms we consider:

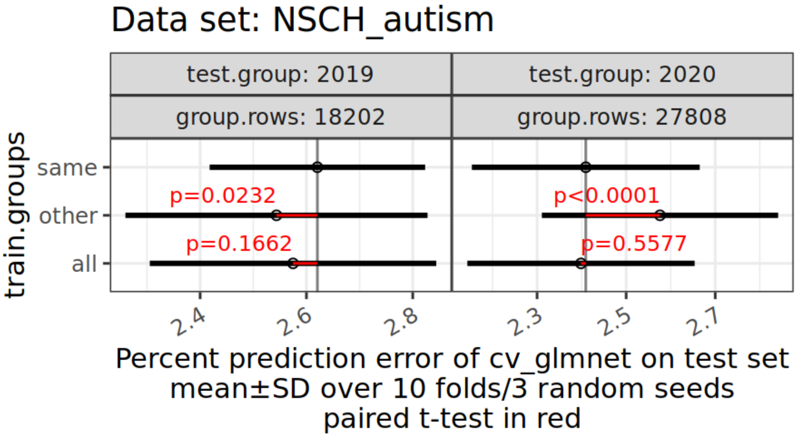
- cv_glmnet: L1-regularized linear model (feature selection).
- xgboost: Extreme gradient boosting (non-linear).
- rpart: Recursive partitioning, decision tree (non-linear, feature selection).
- nearest_neighbors: classic non-linear algorithm.
- featureless: un-informed baseline, ignores all inputs/features, and always predicts the most frequent label in train data (Autism=No in this case).

Same Other CV for Autism data



- Each cv_glmnet model has significantly less error than featureless, indicating that some non-trivial pattern has been learned.

Same Other CV for Autism data

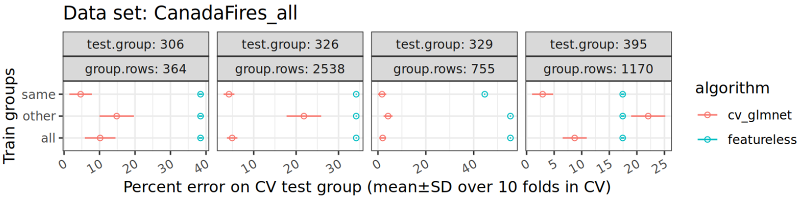


- All has slightly less error than same, which suggests the two years have similar patterns, and can be combined for learning a more accurate model.
- Other has either less error or more, suggesting that the error rate depends on the number of rows in the train set.

Example data 2: Canada fires

- Collaboration with Thibault *et al.*
- Satellite image data, $N = 4827$ rows/pixels, $D = 46$ features/spectral bands.
- Government land management project: oal is to predict whether the pixel has been burned (binary classification, yes or no).
- Four satellite images in different regions of the forest, numbered 306, 326, 329, 395.
- Can we train on one image, and accurately predict on another?

Same Other CV for Canada fires data

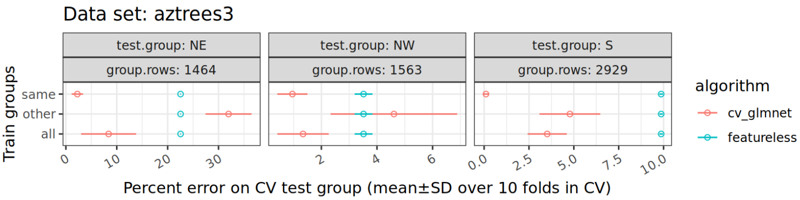


- Each cv_glmnet model has significantly less error than featureless, except train.groups=other for test.group=395 (must have a very different pattern than the other images).
- Training on all images is never as accurate as same, which suggests that images are substantially different, and we need labels from the same image to get optimal predictions.

Example data 3: AZ trees

- Collaboration with Shenkin *et al.*
- Satellite image data, $N = 5956$ rows/pixels, $D = 21$ features/spectral bands.
- Tree stress project: goal is to predict whether the pixel has a tree (binary classification, yes or no).
- Three regions around Flagstaff: NE, NW, S.
- Can we train in one region, and accurately predict on another?

Same Other CV for AZ trees data

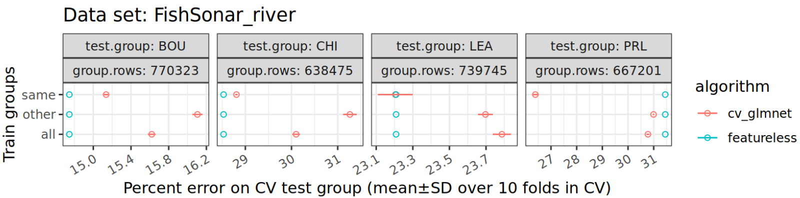


- Each cv_glmnet model has significantly less error than featureless, except train.groups=other for two test groups (must have a very different pattern than the other images).
- Training on all images is never as accurate as same, which suggests that images are substantially different, and we need labels from the same image to get optimal predictions.

Example data 4: fish sonar

- Collaboration with Bodine *et al.*
- Sonar image data, $N = 2,815,744$ rows/pixels, $D = 81$ features (mean pixel intesity in windows around target pixel).
- Conservation project funded by Department of Fish/Wildlife: goal is to predict whether the pixel has a hard bottom suitable for fish spawning (binary classification, yes or no).
- Four rivers in southeast USA: CHI, PRL, LEA, BOU.
- Can we train in one river, and accurately predict on another?

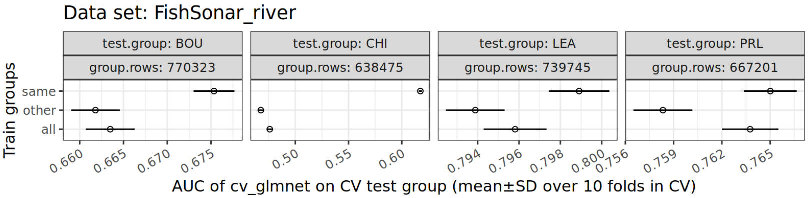
Same Other CV for fish sonar data



- When training on same group, cv_glmnet sometimes has larger test error than featureless, because of class imbalance (hard bottom suitable for fish spawning is rare).

```
       river
label       BOU    CHI    LEA    PRL
  hard   113592 182150 171684 209832
  other  656731 456325 568061 457369
```

Same Other CV for fish sonar data

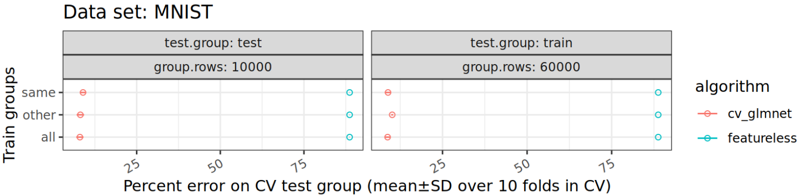


- Area Under the ROC Curve (AUC) is a good measure of accuracy for imbalanced binary classification problems (constant/featureless=0.5, best=1).
- Mostly test AUC is greater than 0.5, which means a non-trivial prediction function has been learned.
- For test.group=CHI with train.groups=all/other, test AUC$<$ 0.5, indicating a very different pattern in this river (opposite of the pattern in other rivers).
- Test AUC for all is never as large as same, indicating that you need data from the same river for optimal prediction accuracy.

Machine learning benchmark data sets

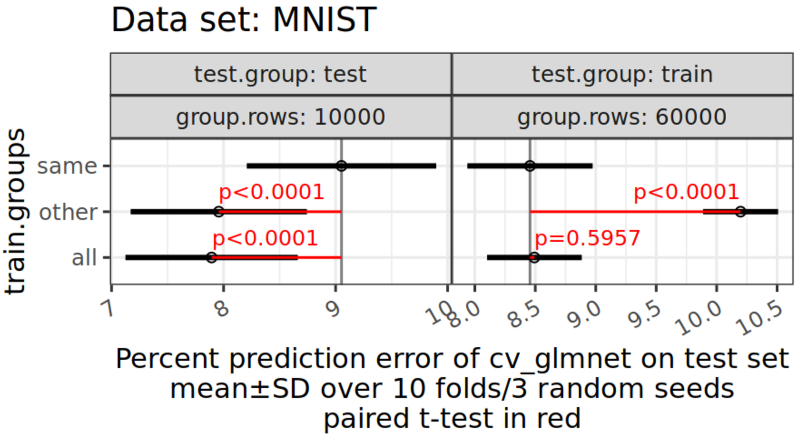
- Machine learning researchers evaluate new algorithms using benchmark data sets, which sometimes have pre-defined train/test splits.
- For example MNIST is a data set of images of handwritten digits (want to predict which digit, 0 to 9), with 60,000 train and 10,000 test images.
- spam is a data set of emails (want to predict spam or not, binary), with 3065 train and 1536 test emails.
- Are the patterns in the pre-defined train/test sets similar/iid?
- Or are they different?

Same Other CV for MNIST data (example 1)



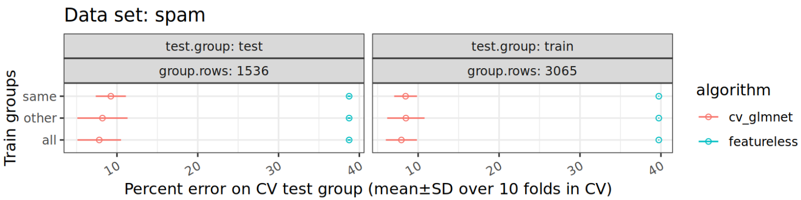
- MNIST data are images of handwritten digits (10 classes).
- Each linear model has much less error than featureless.

Same Other CV for MNIST data (example 1)



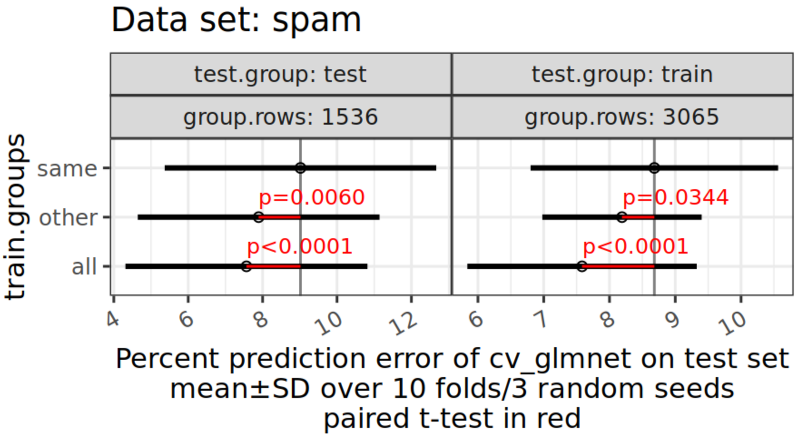
- When predicting on predefined test set, all has significantly lower test error than same, so it is beneficial to combine data (similar pattern, not enough data in small test set).

Same Other CV for spam data (example 2)



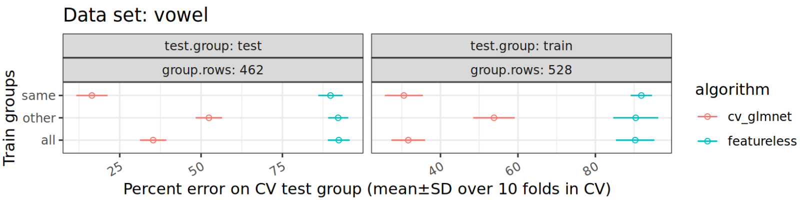
- spam data are emails (binary classification).
- Each linear model has much less error than featureless.

Same Other CV for spam data (example 2)



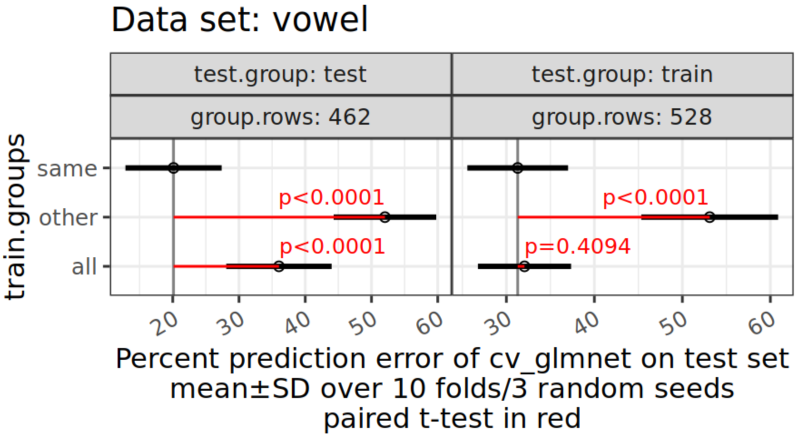
- train.groups=all has significantly lower test error than same, so it is beneficial to combine data (similar pattern, not enough data in either predefined set).

Same Other CV for vowel data (example 3)



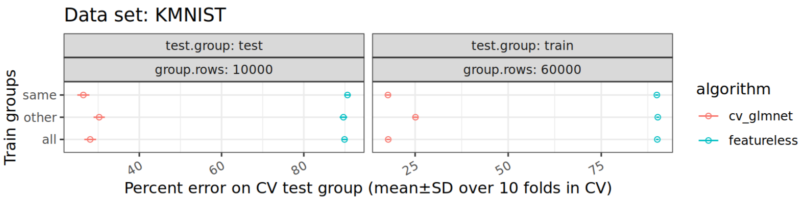
- vowel data are audio/speech recordings (11 classes/speakers).
- Each linear model has much less error than featureless.

Same Other CV for vowel data (example 3)



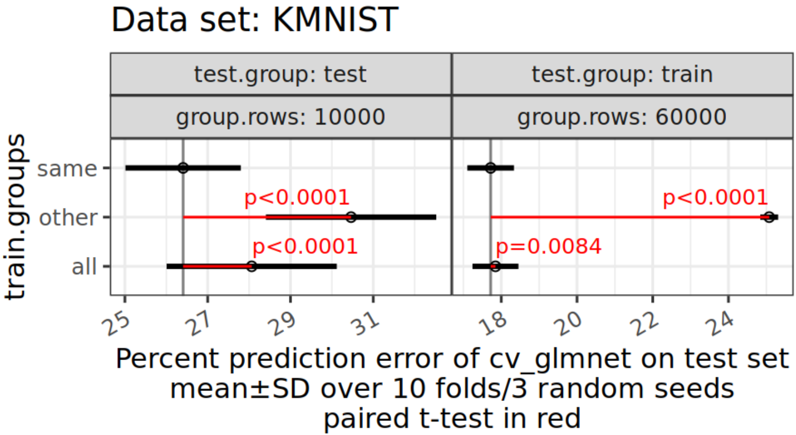
- train.groups=all has significantly larger test error than same, indicating that it is not optimal to combine the predefined sets (which have different patterns).

Same Other CV for KMNIST data (example 4)



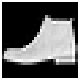
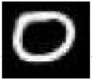
- KMNIST are images of handwritten Japanese (10 classes).
- Each linear model has much less error than featureless.

Same Other CV for KMNIST data (example 4)

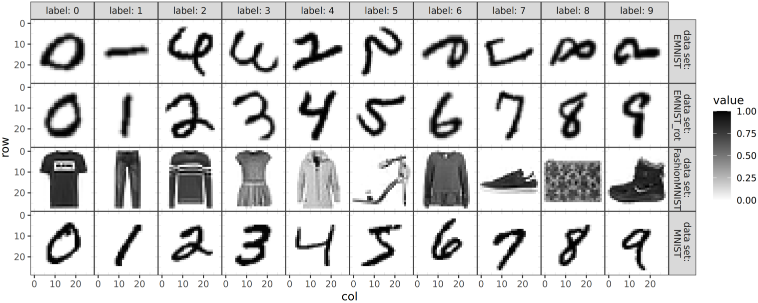


- train.groups=all has significantly larger test error than same, indicating that it is not optimal to combine the predefined sets (which have different patterns).

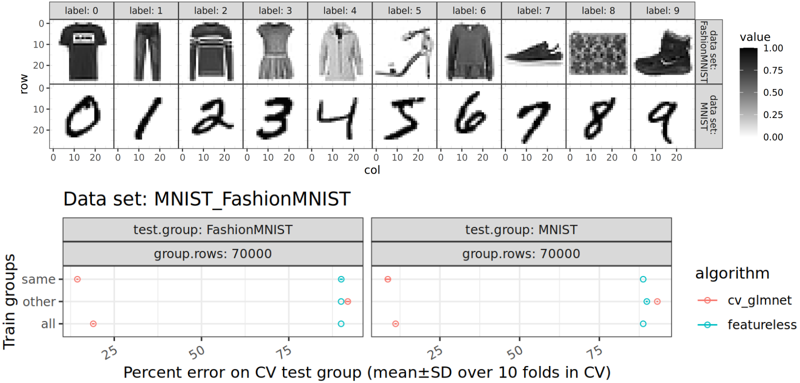
Train on MNIST and accurately predict on EMNIST?

Recall: what happens if you do $g($$)$, or $h($$)$?

- Boot image comes from FashionMNIST data, which were used to learn h.
- 0 image comes from MNIST data, which were used to learn g.

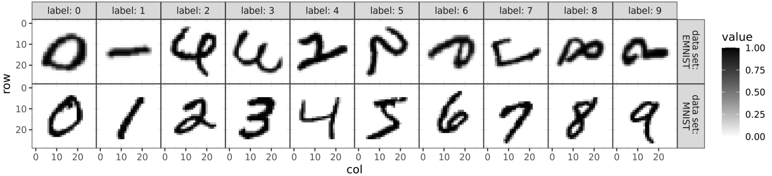


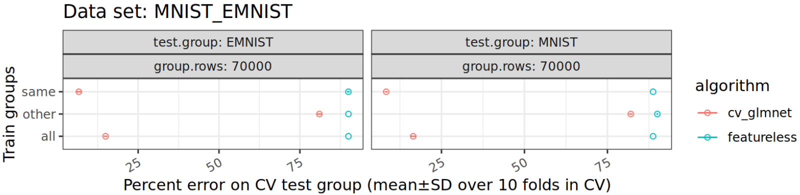
Same Other CV for MNIST+FashionMNIST data



- Other linear model has more test error than featureless, which indicates that the patterns are too different to learn anything at all.

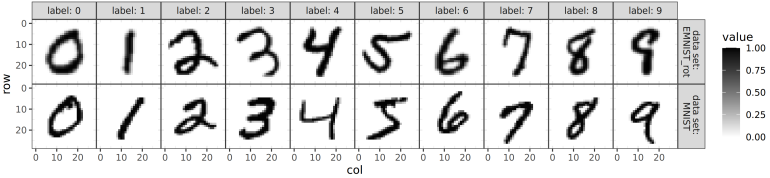
Same Other CV for MNIST+EMNIST data





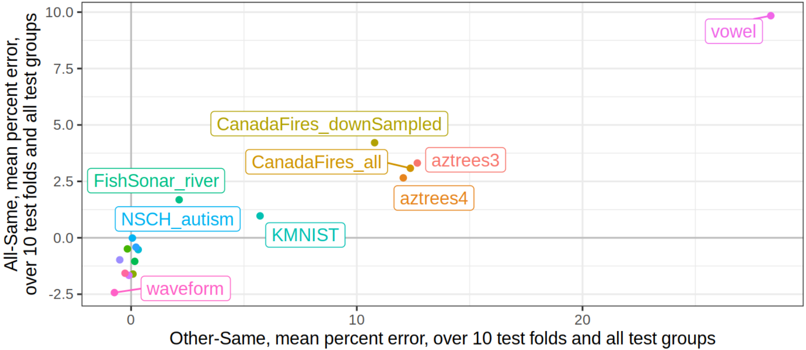
- Other has somewhat smaller test error than featureless, so something is learned/transferable between data sets, but it is still clear that the pattern is very different.

Same Other CV for MNIST+EMNIST_rot data



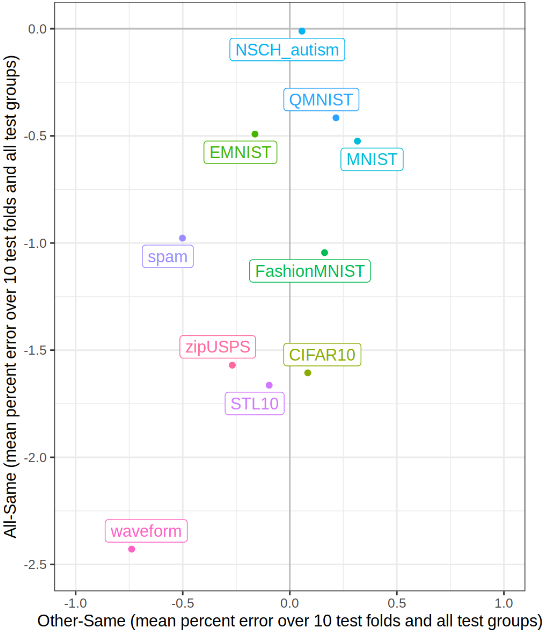
- TODO

Map of data set sub-group similarity



- Overview of how similar/different are sub-groups in each data set. (Data sets with similar sub-groups appear in lower left.)
- Mean difference in percent test error, all/other-same, is shown for each data set.

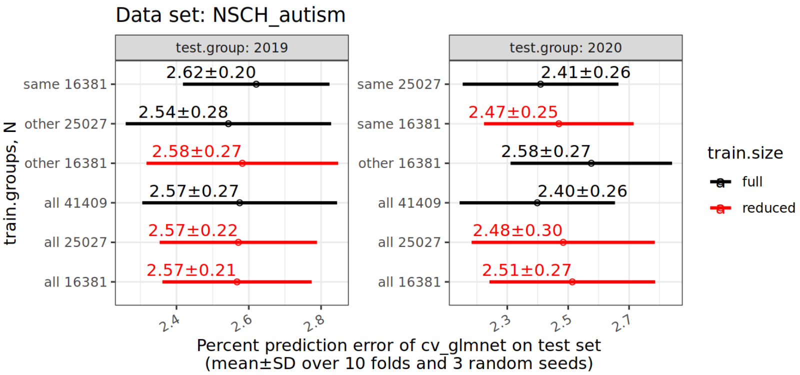
Zoom to most similar data sets



Discussion and Conclusions

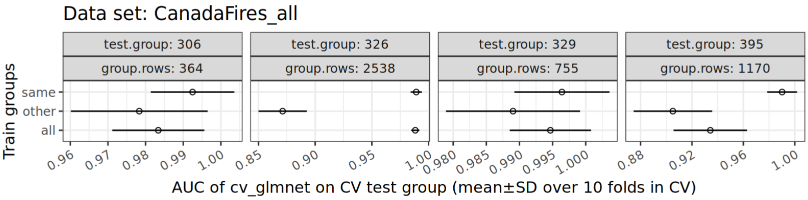
- Proposed Same Other Cross-Validation shows if data sets are similar enough to so that combining data is beneficial for learning (train on one group, test/predict on another).
- In Autism data, there was a slight benefit to combining years.
- In fires/trees/fish data, we observed significant differences between images/regions/rivers.
- Some pre-defined train/test splits in benchmark data sets are similar/iid (MNIST/spam), others are not (KMNIST/vowel).
- Free/open-source R package available: `https://github.com/tdhock/mlr3resampling`
- These slides are reproducible, using the code in `https://github.com/tdhock/cv-same-other-paper`
- Contact: toby.hocking@nau.edu, toby.hocking@r-project.org

Same Other CV for Autism data



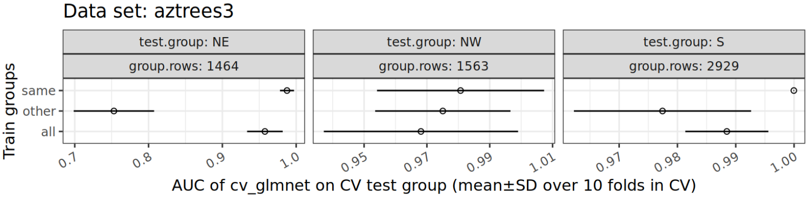
- Reduced sizes (red) are used to judge the effect of sample size.
- Sample size effect present for test group 2020, but not 2019.

Same Other CV for Canada fires data



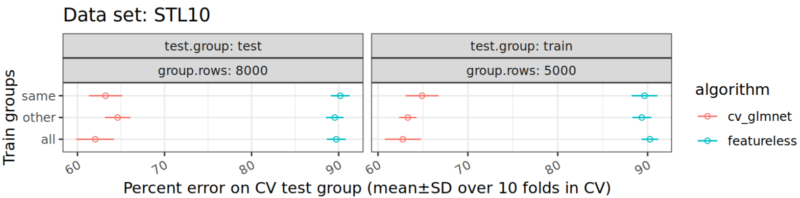
- Area Under the ROC Curve (AUC) is a good measure of accuracy for imbalanced binary classification problems (constant/featureless=0.5, best=1).
- Test AUC for all is never as large as same, indicating that you need data from the same river for optimal prediction accuracy.

Same Other CV for AZ trees data



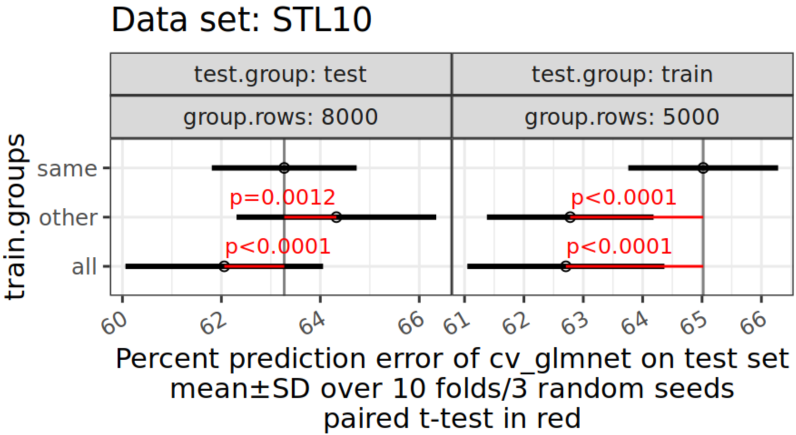
- Area Under the ROC Curve (AUC) is a good measure of accuracy for imbalanced binary classification problems (constant/featureless=0.5, best=1).
- Test AUC for all is never as large as same, indicating that you need data from the same river for optimal prediction accuracy.

Same Other CV for STL10 data



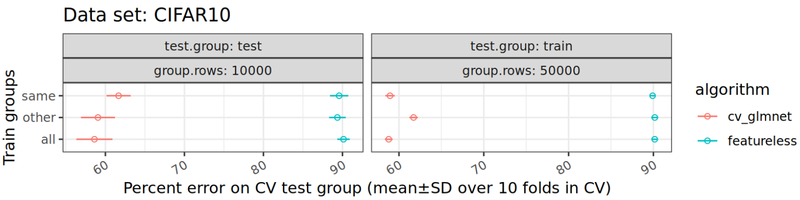
- Image classification data (10 different objects).
- Each linear model has much less error than featureless.

Same Other CV for STL10 data



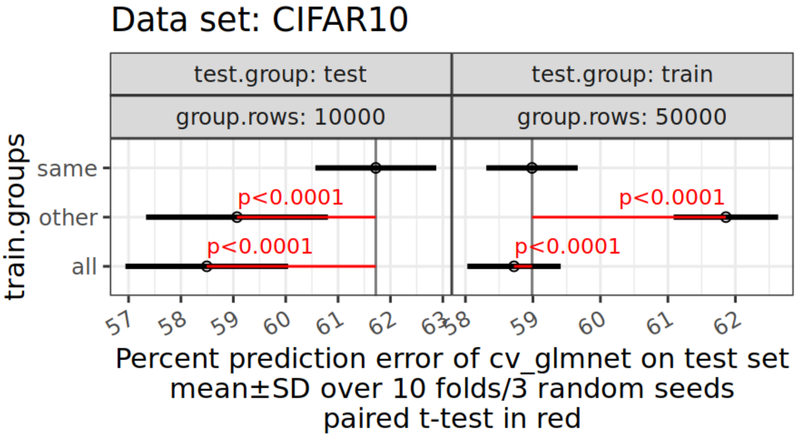
- train.groups=all has significantly lower test error than same, so it is beneficial to combine data (similar pattern, not enough data in predefined train set).

Same Other CV for CIFAR10 data



- Image classification data (10 different objects).
- Each linear model has much less error than featureless.

Same Other CV for CIFAR10 data



- train.groups=all has significantly lower test error than same, so it is beneficial to combine data (similar pattern, not enough data in predefined test set).