Cross-validation: what, how and which?

Pradeep Reddy Raamana

Statistics [from cross-validation] are like bikinis! What they reveal is suggestive, but what they conceal is vital!

Goals for Today
Goals for Today

• What is cross-validation?
Goals for Today

• What is cross-validation?

• How to perform it?
Goals for Today

• What is cross-validation?

• How to perform it?

• What are the effects of different CV choices?
Goals for Today

- What is cross-validation?
- How to perform it?
- What are the effects of different CV choices?

negative bias  unbiased  positive bias

negative bias  unbiased  positive bias

Training set  Test set

≈  ≈  ≈
What is generalizability?

available data (sample)
What is generalizability?

available data (sample)
What is generalizability?

available data (sample)

desired: accuracy on unseen data (population)
What is generalizability?

available data (sample)

desired: accuracy on unseen data (population)
What is generalizability?

- **available data (sample)**
- **desired**: accuracy on **unseen** data (population)
- **out-of-sample predictions**
What is generalizability?

available data (sample)

desired: accuracy on unseen data (population)

out-of-sample predictions

avoid overfitting
Why cross-validate?
Why cross-validate?

Training set

bigger training set

better **learning**

Test set
Why cross-validate?

- **Training set**: bigger training set → better **learning**
- **Test set**: bigger test set → better **validation**
Why cross-validate?

Key: training and test sets are disjoint.
Why cross-validate?

Key: training and test sets are disjoint. And the dataset or sample size is fixed.
Why cross-validate?

Key: training and test sets are disjoint. And the dataset or sample size is fixed. They grow at the expense of each other!
Why cross-validate?

Key: training and test sets are disjoint. And the dataset or sample size is fixed. They grow at the expense of each other!
Why cross-validate?

Key: training and test sets are disjoint. And the dataset or sample size is fixed. They grow at the expense of each other! Cross-validate to maximize both.
Use cases
Use cases

• “When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation (CV) is typically used”
Use cases

• “When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation (CV) is typically used”

• Use cases:
Use cases

• “When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation (CV) is typically used”

• Use cases:
  • to estimate generalizability (test accuracy)
Use cases

• “When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation (CV) is typically used”

• Use cases:
  • to estimate generalizability (test accuracy)
  • to pick optimal parameters (model selection)
Use cases

• “When setting aside data for parameter estimation and validation of results cannot be afforded, cross-validation (CV) is typically used”

• Use cases:
  • to estimate generalizability (test accuracy)
  • to pick optimal parameters (model selection)
  • to compare performance (model comparison).
Types of CV
Types of CV

1. **How you split** the dataset into train/test
Types of CV

1. **How you split** the dataset into train/test

- maximal independence between training and test sets is desired.
Types of CV

1. **How you split** the dataset into train/test

   • maximal independence between training and test sets is desired.
   • This split could be
     • over samples (e.g. indiv. diagnosis)
1. **How you split** the dataset into train/test

- maximal independence between training and test sets is desired.
- This split could be
  - over samples (e.g. indiv. diagnosis)
Types of CV

1. **How you split** the dataset into train/test

   • maximal independence between training and test sets is desired.

   • This split could be
     
     • over samples (e.g. indiv. diagnosis)
     
     • over time (for task prediction in fMRI)
Types of CV

1. **How you split** the dataset into train/test

   - maximal independence between training and test sets is desired.
   - This split could be
     - over samples (e.g. indiv. diagnosis)
     - over time (for task prediction in fMRI)
Types of CV

1. **How you split** the dataset into train/test
   - maximal independence between training and test sets is desired.
   - This split could be
     - over samples (e.g. indiv. diagnosis)
     - over time (for task prediction in fMRI)

2. **How often you repeat randomized splits?**
   - to expose classifier to full variability
   - As many as times as you can e.g. 100
Many other variations!

- **k-fold**, \(k = 2, 3, 5, 10, 20\)
- **hold-out**, train \%= 50, 63.2, 75, 80, 90
- **stratified**
  - across train/test
  - across classes
- **inverted**: very small training, large testing
- leave one sample / pair / tuple condition / task / block out

1. 2-fold cross-validation (kf2)
2. 3-fold cross-validation (kf3)
3. 5-fold cross-validation (kf5)
4. 10-fold cross-validation (kf10)
5. 2 times repeated 5-fold (2xkf5)
6. 2 times repeated 10-fold (2xkf10)
7. 5, 10, and 20 times repeated bootstrap (5xboot, 1
8. 80/20 hold-out (80/20) — a training set of size data, and test set of 20%, with similar proportion
9. resubstitution (resub), training and testing in the
10. inverted 5-fold (invkf5): learning on a single fold,
11. 20/20 hold out (20/20) — training and test sets c
12. 5 times repeated 20/20 hold out (5x20/20)
13. 20/10 holdout (20/10)
14. 10/10 hold out (10/10)
15. 5 times repeated 10/10 hold out (5x10/10)

Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing whole dataset
Hold-out CV

Set aside a fixed percentage (e.g., 30%) for testing.
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing

whole dataset

trial
1 Train Test
2
...

n
Hold-out CV

Set aside a fixed percentage (e.g., 30%) for testing

whole dataset

<table>
<thead>
<tr>
<th>trial</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing

Note: there could be overlap among the test sets!
i.e. test sets in different iterations could have common samples
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Note: different folds won’t be contiguous.
K-fold CV

Test sets in different trials are indeed mutually disjoint

Note: different folds won’t be contiguous.
Validation set

Training set
Validation set

Training set

goodness of fit of the model
Validation set

Training set

goodness of fit of the model

biased towards the training set
Validation set

Training set

goodness of fit
of the model

biased towards
the training set

Test set
Validation set

Training set

Test set

≈ \mathbb{N} \approx

Goodness of fit of the model

Biased towards the training set
Validation set

- Training set
  - Goodness of fit of the model
  - Biased towards the training set
- Test set
  - Optimize parameters

\[ \approx \approx \]

10
Validation set

Training set

- goodness of fit of the model
- biased towards the training set

Test set

- optimize parameters
- biased towards the test set
Validation set

- Training set
- Test set
- Validation set

- $\approx \aleph \approx 10$

- Goodness of fit of the model
  - Biased towards the training set

- Optimize parameters
  - Biased towards the test set
Validation set

- Training set
  - goodness of fit of the model
  - biased towards the training set

- Test set
  - optimize parameters
  - biased towards the test set

- Validation set
  - evaluate generalization
  - independent of training or test sets
Validation set

Whole dataset

Training set

Test set

Validation set

≈ \aleph \approx

goodness of fit
of the model

optimize
parameters

evaluate
generalization

biased towards
the training set

biased towards
the test set

independent of
training or test sets

- Validation set
- Training set
- Test set
- Whole dataset
Validation set

Whole dataset

Training set → Test set → Validation set

inner-loop

≈ \aleph

- goodness of fit of the model
- biased towards the training set

- optimize parameters
- biased towards the test set

- evaluate generalization
- independent of training or test sets
Validation set

Whole dataset

outer-loop

inner-loop

Training set

Test set

Validation set

≈ ℵ

goodness of fit of the model

optimize parameters

evaluate generalization

biased towards the training set

biased towards the test set

independent of training or test sets
Measuring bias in CV measurements
Measuring bias in CV measurements

Whole dataset

Training set

Test set

Inner-CV
Measuring bias in CV measurements

cross-validation accuracy!
Measuring bias in CV measurements

cross-validation accuracy!
Measuring bias in CV measurements

Whole dataset

- Training set
- Test set
- Validation set

Inner-CV

cross-validation accuracy!

validation accuracy!
Measuring bias in CV measurements

Whole dataset

Training set

Test set

Validation set

Inner-CV

cross-validation accuracy!

validation accuracy!
Measuring bias in CV measurements

Whole dataset

Training set

Inner-CV

Test set

Validation set

cross-validation accuracy!  \approx  validation accuracy!

positive bias  \quad \text{unbiased}  \quad \text{negative bias}
## Intra-subject datasets: Haxby

<table>
<thead>
<tr>
<th>Task</th>
<th># samples</th>
<th>#blocks</th>
<th>mean accuracy of SVM $l2$</th>
<th>mean accuracy of SVM $l1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bottle / scramble</td>
<td>209</td>
<td>12 secs</td>
<td>75%</td>
<td>86%</td>
</tr>
<tr>
<td>cat / bottle</td>
<td></td>
<td></td>
<td>62%</td>
<td>69%</td>
</tr>
<tr>
<td>cat / chair</td>
<td></td>
<td></td>
<td>69%</td>
<td>80%</td>
</tr>
<tr>
<td>cat / face</td>
<td></td>
<td></td>
<td>65%</td>
<td>72%</td>
</tr>
<tr>
<td>cat / house</td>
<td></td>
<td></td>
<td>86%</td>
<td>95%</td>
</tr>
<tr>
<td>cat / scramble</td>
<td></td>
<td></td>
<td>83%</td>
<td>92%</td>
</tr>
<tr>
<td>chair / scramble</td>
<td></td>
<td></td>
<td>77%</td>
<td>91%</td>
</tr>
<tr>
<td>chair / shoe</td>
<td></td>
<td></td>
<td>63%</td>
<td>70%</td>
</tr>
<tr>
<td>face / house</td>
<td></td>
<td></td>
<td>88%</td>
<td>96%</td>
</tr>
<tr>
<td>face / scissors</td>
<td></td>
<td></td>
<td>72%</td>
<td>83%</td>
</tr>
<tr>
<td>scissors / scramble</td>
<td></td>
<td></td>
<td>73%</td>
<td>87%</td>
</tr>
<tr>
<td>scissors / shoe</td>
<td></td>
<td></td>
<td>60%</td>
<td>64%</td>
</tr>
<tr>
<td>shoe / bottle</td>
<td></td>
<td></td>
<td>62%</td>
<td>69%</td>
</tr>
<tr>
<td>shoe / cat</td>
<td></td>
<td></td>
<td>72%</td>
<td>85%</td>
</tr>
<tr>
<td>shoe / scramble</td>
<td></td>
<td></td>
<td>78%</td>
<td>88%</td>
</tr>
</tbody>
</table>
Inter-subject fMRI datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th># samples</th>
<th># blocks (sess./subj.)</th>
<th>Task</th>
<th>SVM $\ell_2$</th>
<th>SVM $\ell_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duncan [9]</td>
<td>fMRI, across subjects</td>
<td>196</td>
<td>49 subj.</td>
<td>consonant / scramble</td>
<td>92%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>consonant / word</td>
<td>92%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>objects / consonant</td>
<td>90%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>objects / scramble</td>
<td>91%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>objects / words</td>
<td>74%</td>
<td>71%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>words / scramble</td>
<td>91%</td>
<td>89%</td>
</tr>
<tr>
<td>Wager [53]</td>
<td>fMRI</td>
<td>390</td>
<td>34 subj.</td>
<td>negative cue / neutral cue</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td>across subjects</td>
<td></td>
<td></td>
<td>negative rating / neutral rating</td>
<td>54%</td>
<td>53%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>negative stim / neutral stim</td>
<td>77%</td>
<td>73%</td>
</tr>
<tr>
<td>Cohen (ds009)</td>
<td>fMRI</td>
<td>80</td>
<td>24 subj.</td>
<td>successful / unsuccessful stop</td>
<td>67%</td>
<td>63%</td>
</tr>
<tr>
<td>Moran [34]</td>
<td>fMRI</td>
<td>138</td>
<td>36 subj.</td>
<td>false picture / false belief</td>
<td>72%</td>
<td>71%</td>
</tr>
<tr>
<td>Henson [19]</td>
<td>fMRI</td>
<td>286</td>
<td>16 subj.</td>
<td>famous / scramble</td>
<td>77%</td>
<td>74%</td>
</tr>
<tr>
<td></td>
<td>across subjects</td>
<td></td>
<td></td>
<td>famous / unfamiliar</td>
<td>54%</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>scramble / unfamiliar</td>
<td>73%</td>
<td>70%</td>
</tr>
<tr>
<td>Knops [23]</td>
<td>fMRI, across subjects</td>
<td>114</td>
<td>19 subj.</td>
<td>right field / left field</td>
<td>79%</td>
<td>73%</td>
</tr>
</tbody>
</table>

Results: hold-out (10 trials)
Results: hold-out (10 trials)
Results: hold-out (10 trials)
Results: hold-out (10 trials)

Classifier accuracy on validation set via cross-validation

Classifier accuracy

Intra subject
Inter subject

unbiased

Classifier accuracy on validation set
Results: hold-out (10 trials)

Classifier accuracy on validation set via cross-validation

Classifier accuracy via cross-validation

Intra subject
Inter subject

unbiased
negatively biased
Results: hold-out (10 trials)

Classifier accuracy on validation set

Classifier accuracy via cross-validation

- **Unbiased**
- **Negatively biased**
- **Positively biased**
CV vs. Validation: real data

Leave one sample out
CV vs. Validation: real data

negative bias  unbiased  positive bias

Leave one sample out

+3%  +43%
CV vs. Validation: real data

- Negative bias
- Unbiased
- Positive bias

Leave one sample out

-21%  +3%  +43%

Leave one subject/session

-21%  +17%
CV vs. Validation: real data

- **Negative bias**
  - Leave one sample out: +3%
  - Leave one subject/session: -21%
  - 20% left-out, 3 splits: -24%

- **Unbiased**
  - Leave one sample out: +3%
  - Leave one subject/session: +17%
  - 20% left-out, 3 splits: +16%

- **Positive bias**
  - Leave one sample out: +43%
  - Leave one subject/session: +17%
  - 20% left-out, 3 splits: +16%
CV vs. Validation: real data
Simulations: known ground truth
Simulations: known ground truth
Simulations: known ground truth
CV vs. Validation

Leave one sample out  negative bias  unbiased  positive bias

+4%  +33%
CV vs. Validation

- Negative bias
- Unbiased
- Positive bias

Leave one sample out

Leave one block out

-8% +8%
+4% +33%
CV vs. Validation

- **Leave one sample out**
  - Negative bias: -8%
  - Unbiased: +8%
  - Positive bias: +10%

- **Leave one block out**
  - Negative bias: -8%
  - Unbiased: +8%
  - Positive bias: +11%

- **20% left-out, 3 splits**
  - Negative bias: -8%
  - Unbiased: +8%
  - Positive bias: +11%
CV vs. Validation

- Leave one sample out
  - negative bias
  - unbiased
  - positive bias

- Leave one block out
  - -8% +8%

- 20% left-out, 3 splits
  - -10% +11%

- 20% left-out, 10 splits
  - -8% +8%

- 20% left-out, 50 splits
  - -7% +7%

Simulations
Aggregation across folds

- It’s not enough to properly split each fold, and accurately evaluate classifier performance!

- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

- It’s not enough to properly split each fold, and accurately evaluate classifier performance!

- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

- It’s not enough to properly split each fold, and accurately evaluate classifier performance!

- Not all measures across folds are commensurate!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

• It’s not enough to properly split each fold, and accurately evaluate classifier performance!

• Not all measures across folds are commensurate!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

- It’s not enough to properly split each fold, and accurately evaluate classifier performance!

- Not all measures across folds are commensurate!
  
  - e.g. decision scores from SVM (reference plane and zero are different!)

  - hence they can not be pooled across folds to construct an ROC!

  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

- It's not enough to properly split each fold, and accurately evaluate classifier performance!

- Not all measures across folds are commensurate!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Aggregation across folds

• It’s not enough to properly split each fold, and accurately evaluate classifier performance!

• Not all measures across folds are *commensurate*!
  
  • e.g. decision scores from SVM (reference plane and zero are different!)
  
  • hence they can not be pooled across folds to construct an ROC!

  • Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!
Conclusions

- Avoid leave-one-out cross-validation
  - esp. when correlations are present in your data
  - produces optimistic estimates with high variance
- Use repeated-holdout (10-50% for testing)
  - respecting sample/dependency structure
  - maximizing independence between train & test sets
In God we trust, but all others must cross-validate!

- Results could vary drastically with a different CV scheme
- CV results have variance (>10%)
- Document CV scheme in detail:
  - type of split
  - number of repetitions
  - Full distribution of estimates
- Proper splitting is not enough, proper pooling is needed too.

Reviewer 2 is watching!
References


Acknowledgements