Model Comparison

Cole Weis 1/18/22

Questions to answer as a group

1. What is included in our definition of a "model"?
2. Should we use validation testing?
3. Scoring of accuracy vs. spread
4. Pseudodata generation

Interface for method

```python
class CFFExtractionModel:
    def run(self, input_df) -> np.array:
        """Train model and output results"""
        pass


class myModel(CFFExtractionModel):
    def run(self, input_df):
        result = np.array([1, 2, 3, 4, 4])
        return result
```

What is included in our definition of a model?

What is included in our definition of a model?

a) Network architecture.

b) Loss function and optimizer.

What is included in our definition of a model?

a) Network architecture.

b) Loss function and optimizer.

c) Epochs, batch size, hyperparameters?

What is included in our definition of a model?

a) Network architecture.

b) Loss function and optimizer.

c) Epochs, batch size, hyperparameters?

d) Bootstrapping method?

    a) # of replicas?

What is included in our definition of a model?

a) Network architecture.
b) Loss function and optimizer.
c) Epochs, batch size, hyperparameters?
d) Bootstrapping method?
    a) # of replicas?

-> Cole's suggestion: ignore all interior components of a model except for a histogram of CFF predictions, score only the predictions.

Validation/Testing Set

- DNN almost always make use of validation sets to prevent overfitting

- Standard methods of model comparison use testing sets (k-fold CV)

- Current method for CFF extraction uses no validation sets, as the target of extraction is not the metric being scored.
  - i.e. CFF's can be scored without worrying about overfitting since they were never given
  - On the other hand fits to F should only be compared using validation sets

Validation/Testing Set

- DNN almost always make use of validation sets to prevent overfitting

- Standard methods of model comparison use testing sets (k-fold CV)

- Current method for CFF extraction uses no validation sets, as the target of extraction is not the metric being scored.
  - i.e. CFF's can be scored without worrying about overfitting since they were never given
  - On the other hand fits to F should only be compared using validation sets

- Do we still want to use k-fold or other CV methods for CFF comparison?

Scoring of Accuracy vs. Spread

- Should we use one metric or always consider both?
  - If both, std. dev will be sufficient for spread, but what is the best measure of accuracy?
- How can we account for different # of replicas between models (different number of predictions)?
- Ideas for a single metric:
  - Mean square-root error less errF, mean squared error less errF
  - Mean error of all predictions, Q1 error of predictions
  - General idea: reward larger spread when far from true value and smaller spread when on true value

Psuedodata Generation

- Should we consider coding an automated process of pseudodata generation for testing and possibly training purposes?
  - Decrease bias in testing