Protein Prediction

Exercise

22.06.2017
Hi, I'm Michael Bernhofer.

PhD student @ Rostlab

Email: bernhoferm@rostlab.org
Reminder: \( k \)-fold cross-validation

- Split data set into \( k \) folds (equal size):
  - Train on first \( k-2 \) folds (train)
  - Optimize on fold \( k-1 \) (cross-train)
  - Evaluate on fold \( k \) (test)
  - Rotate & repeat \( k \) times

- Never make decisions on the test set!

- Better: make one split you never use until final evaluation
  - Only viable if enough data
Reminder: Redundancy reduction

• Make sure your data is non-redundant
  • No identical/similar sequences
  • Use CD-Hit (if necessary) and Uniqueprot (HVAL 0)

• Not enough training data?
  • May allow similar sequences in train set
  • But not in cross-train/test
  • Check redundancy between train & cross-train/test!
Feature selection I

• Why?
  • Help the algorithm find the best features
  • Some features might be redundant/noisy/random
  • Reduce complexity of model (takes less space/time)

• Use cross-train set to select, never test set!

• How?
  • Manual selection or automatic methods
  • Due to time constraints: use Scikit learn or WEKA
Feature selection II: Scikit learn

- Methods:
  - Remove low variance features
  - Univariate feature selection
  - Recursive feature elimination (uses external model)
  - Feature importance (after training)

- Online documentation:
Feature selection III: WEKA

- Go to "Select attributes" tab or use "AttributeSelectedClassifier"
Parameter tuning I

- Neural networks
  - # of hidden units
  - Learning rate
  - Momentum

- Random forest
  - # of trees
  - # of randomly selected attributes
  - Size of random sub-sample
Parameter tuning II

• Again: do not optimize on test set!

• Scikit learn (online documentation):

• WEKA:
  • Meta-Classifier: "CVParameterSelection"
Performance evaluation I

- Accuracy (how many predictions are correct)
  - \( a = \frac{\text{correct predictions}}{\text{total predictions}} \)

- Which one is better?
  - Method A: 65% accuracy
  - Method B: 67% accuracy

- What’s the estimated error?
  - Method A: 65\( \pm 5 \)% accuracy
  - Method B: 67\( \pm 5 \)% accuracy
Performance evaluation II

• How to estimate error?

• Use bootstrapping:
  • Create random subset of your test samples
    • Either: \( \sim 60\% \) of test samples (without replacement)
    • Or: 100\% of test samples (with replacement)
  • Compute performance statistic \( x \)
  • Repeat \( N \) times
  • Calculate standard deviation:

\[
s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}
\]
Performance evaluation III

- Other statistics?

- Precision (how many “positive” predictions are correct)
  
  \[ p = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \]

- Recall (how many “positive” cases did we find)
  
  \[ r = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]

- Positives: e.g. “metal-binding residue“
Performance evaluation IV

- Method A:
  - Recall = 80%
  - Precision 90%

- Method B:
  - Recall = 80%
  - Precision 60%

- Which one is better (ignoring error)?
Performance evaluation V

• Look at confusion matrix!

<table>
<thead>
<tr>
<th></th>
<th>Method A</th>
<th></th>
<th>Method B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted as 1</td>
<td>Predicted as 2</td>
<td>Predicted as 1</td>
</tr>
<tr>
<td>Real class 1</td>
<td>80</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>Real class 2</td>
<td>9</td>
<td>1</td>
<td>53</td>
</tr>
</tbody>
</table>

False positive rate (FPR)
• Method A: 90%
• Method B: 30%
Performance evaluation VI

- Lesson to learn:
  - Use multiple measurements
  - Compute error estimate
  - Look at confusion matrix
  - Compare to random/baseline
Putting it all together

• Combine your \( k \) models into one predictor
  • Use either majority vote or average prediction score

• If you have a hold-out set
  • Combine your features & parameters
    • All features used for the different models
    • Average parameter values
  • Train on all CV folds
  • Test final model on hold-out set
Task for next week

• Build and optimize your models
  • Which algorithm (neural network, random forest, etc.)
  • What are the best features

• Evaluate your model’s performance

• If you have the time:
  • Can you see some trends?
  • What proteins/binding-types work best/worst?