Cross validation, training and evaluation of data driven prediction methods

Morten Nielsen
Department of Health Technology, DTU
Data driven method training

• A prediction method contains a very large set of parameters
  - A matrix for predicting binding for 9meric peptides has 9x20=180 weights
• Over fitting is a problem
Evaluation of predictive performance

- Train PSSM on raw data
  - No pseudo counts, No sequence weighting
  - Fit 9*20 (=180) parameters to 9 (*10 = 90) data points
- Evaluate on training data
  - PCC = 0.97
  - AUC = 1.0
- Close to a perfect prediction method
Evaluation of predictive performance

• Train PSSM on **Permuted** (random) data
  - No pseudo counts, No sequence weighting
  - Fit 9*20 parameters to 9*10 data points

• Evaluate on training data
  - PCC = 0.97
  - AUC = 1.0

• Close to a perfect prediction method AND
• Same performance as on the original data
Repeat on large training data
(229 ligands)
When is overfitting a problem?

**FLAFFSNGV**

**FLAFFSNGV**

<table>
<thead>
<tr>
<th>FLAFFSNGV</th>
<th>FLAFFSNGV</th>
<th>VLMEAQOGI</th>
<th>ILLLDQVLV</th>
<th>KMEYEVFKG</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLMRDPAQQ</td>
<td>WLHKQVFQWL</td>
<td>ALAPSTMK</td>
<td>MLLTFI</td>
<td></td>
</tr>
<tr>
<td>FLIVSLCP</td>
<td>ITWQVPFSV</td>
<td>RMPAVTDLV</td>
<td>ALSYASAK</td>
<td></td>
</tr>
<tr>
<td>YFLRRLALV</td>
<td>PLLDYEGETL</td>
<td>FLITGFDI</td>
<td>LLVLC</td>
<td></td>
</tr>
<tr>
<td>MTSELAALI</td>
<td>MLLHVG1PL</td>
<td>GLIIISIFL</td>
<td>IVYGRSNAI</td>
<td></td>
</tr>
<tr>
<td>GLEYAIEEC</td>
<td>SLSHYFTLV</td>
<td>GLYLYTTEV</td>
<td>AQSDFMSWV</td>
<td></td>
</tr>
<tr>
<td>KLFFAKCLV</td>
<td>VLWEGGHDL</td>
<td>YLLNYAGRI</td>
<td>RLEELLPAV</td>
<td></td>
</tr>
<tr>
<td>VLOAGFFLL</td>
<td>AIDDFCLFA</td>
<td>KVVSLVILA</td>
<td>LLVFACSAV</td>
<td></td>
</tr>
<tr>
<td>TLKADAMQL</td>
<td>GLFQEAYPL</td>
<td>YQLGDYFFV</td>
<td>GMVIACL</td>
<td></td>
</tr>
<tr>
<td>MSDIFHALV</td>
<td>MVVKVNAAL</td>
<td>FMTALVSL</td>
<td>WLISTYAVRI</td>
<td></td>
</tr>
<tr>
<td>GMRDVSFEI</td>
<td>FGFLATAG</td>
<td>ILAKFL</td>
<td>IVFVNL</td>
<td></td>
</tr>
<tr>
<td>QLPLESDAV</td>
<td>SLYPPLCFK</td>
<td>MTPSPFYTV</td>
<td>LLVAPMPTA</td>
<td></td>
</tr>
<tr>
<td>KVGCNETV</td>
<td>IIPPATHYV</td>
<td>IIDQVPFSV</td>
<td>YLNI</td>
<td></td>
</tr>
<tr>
<td>ILYQVPFSV</td>
<td>YLMKDKNLI</td>
<td>AIMEKNI</td>
<td>LLLNSLGSV</td>
<td></td>
</tr>
<tr>
<td>GLKISLCGI</td>
<td>ALGLGVSL</td>
<td>MCMCPFLF</td>
<td>FMFNELLAL</td>
<td></td>
</tr>
<tr>
<td>WELETILVFV</td>
<td>ALYWALMES</td>
<td>GLDPTGVAV</td>
<td>GMSPVCLPI</td>
<td></td>
</tr>
<tr>
<td>WQDGGWQS</td>
<td>LLIEGIFFI</td>
<td>SILNLRLFL</td>
<td>GLPSLSCUL</td>
<td></td>
</tr>
<tr>
<td>VMILIGIES</td>
<td>RLLNKIVSSEL</td>
<td>KVEKYL</td>
<td>YLVAYQATV</td>
<td></td>
</tr>
<tr>
<td>SVMDPLIYA</td>
<td>IMSSFQAT</td>
<td>FTLVATVSI</td>
<td>ILLVAVSFV</td>
<td></td>
</tr>
<tr>
<td>GMFGCCFAA</td>
<td>RLLDDTPEV</td>
<td>SLDSLVLHLL</td>
<td>LLVQAGFFL</td>
<td></td>
</tr>
<tr>
<td>VLAGYGAGI</td>
<td>VILWFSFG</td>
<td>VLNLTLMFMV</td>
<td>CFOQAKWY</td>
<td></td>
</tr>
</tbody>
</table>
When is overfitting a problem?

![Graph showing peptide scores and numbers of peptides]
When is overfitting a problem?
Gibbs clustering (multiple specificities)

Multiple motifs!

Cluster 1

Cluster 2
When is overfitting a problem?

Always
How to training a method. A simple statistical method: Linear regression

Observations (training data): a set of x values (input) and y values (output).

Model: $y = ax + b$ (2 parameters, which are estimated from the training data)

Prediction: Use the model to calculate a y value for a new x value

Note: the model does not fit the observations exactly. Can we do better than this?
Overfitting

\[ y = ax + b \]

2 parameter model
Good description, poor fit

\[ y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g \]

7 parameter model
Poor description, good fit

**Note:** It is not interesting that a model can fit its observations (training data) exactly.

To function as a prediction method, a model must be able to generalize, i.e. produce sensible output on new data.
How to estimate parameters for prediction?

A Regression Problem

$y = f(x) + \text{noise}$

Can we learn $f$ from this data?

Let’s consider three methods…
Model selection

Which is best?

- Linear Regression
- Quadratic Regression
- Join-the-dots
The test set method

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
The test set method

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set

(Linear regression example)
The test set method

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Linear regression example)
Mean Squared Error = 2.4
The test set method

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Quadratic regression example)
Mean Squared Error = 0.9
The test set method

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Join the dots example)
Mean Squared Error = 2.2

So quadratic function is best
How to deal with overfitting? Cross validation

Cross validation

Train on 4/5 of data
Test/evaluate on 1/5

⇒

Produce 5 different methods each with a different prediction focus
Model over-fitting

2000 MHC:peptide binding data
PCC=0.99

Evaluate on 600 MHC:peptide binding data
PCC=0.70
Model over-fitting (early stopping)

Evaluate on 600 MHC:peptide binding data
PCC=0.89
What is going on?

Temperature vs. years graph showing a trend of increasing temperature over time, with fluctuations and a notable rise at the end.
5 fold training

Which method to choose?
5 fold training

Pearson correlation

Train
Test
Eval
The Wisdom of the Crowds

- The Wisdom of Crowds. Why the Many are Smarter than the Few. James Surowiecki

One day in the fall of 1906, the British scientist Fracis Galton left his home and headed for a country fair... He believed that only a very few people had the characteristics necessary to keep societies healthy. He had devoted much of his career to measuring those characteristics, in fact, in order to prove that the vast majority of people did not have them. ... Galton came across a weight-judging competition...Eight hundred people tried their luck. They were a diverse lot, butchers, farmers, clerks and many other no-experts...The crowd had guessed ... 1.197 pounds, the ox weighted 1.198
The wisdom of the crowd!

- The highest scoring hit will often be wrong
  • Not one single prediction method is consistently best
- Many prediction methods will have the correct fold among the top 10-20 hits
- If many different prediction methods all have a common fold among the top hits, this fold is probably correct
Method evaluation

- Use cross validation
- Evaluate on concatenated data and **not** as an average over each cross-validated performance
Method evaluation
Method evaluation

![Method evaluation graph](image)
Method evaluation
How many folds?

• Cross validation is always good!, but how many folds?
  - Few folds -> small training data sets
  - Many folds -> small test data sets

• 560 peptides for training
  - 50 fold (10 peptides per test set, few data to stop training)
  - 2 fold (280 peptides per test set, few data to train)
  - 5 fold (110 peptide per test set, 450 per training set)
Problems with 5fold cross validation

- Use test set to stop training, and test set performance to evaluate training
  - Over-fitting?
- If test set is small, Yes
- If test set is large, No
- Confirm using “true” 5 fold cross validation
  - 1/5 for evaluation
  - 4/5 for 4 fold cross-validation
Conventional 5 fold cross validation
“Nested (or true)” 5 fold cross validation
When to be careful

• When data is scarce, the performance obtained used “conventional” versus “nested” cross validation can be very large.

• When data is abundant the difference is in general small.
Training/evaluation procedure

- Define method
- Select data
- Deal with data redundancy
  - In method (sequence weighting)
  - In data (Hobohm)
- Deal with over-fitting either
  - in method (SMM regulation term) or
  - in training (stop fitting on test set performance)
- Evaluate method using cross-validation