

Answers

Q1: How many training data is in the training set?

Number of training data: 1908

Q2: How many of the examples have actually been used in the training?

80% of 1908 = 1527 el. 1528

Q3: What part of the training set has been used for testing?

Last 20 % (bottom part)

Q4: What époque is obtaining the best linear correlation and what is the value?

Maximal test set pearson correlation coefficient sum = 0.910800 in epoch 134

Q5: What influence has this on the training?

The training will be stopped after époque 134 (to avoid overtraining, this can diverge between training setups, sometimes we stop on best Matthews correlation and sometimes on lowest error, depending on the problem)

Q6: How many training data is in this training set?

Number of training data: 1960

Q7: How many data has been used for testing?

20% of 1960 = 392

Q8: What époque is obtaining the best linear correlation and what is the value?

Maximal test set pearson correlation coefficient sum = 0.804000 in epoch 52

Q9 : Is this different from the previous training?

Yes, significantly lower

Both training sets contain only unique peptides.

Q10: What might the reason be of the difference between the two training sessions?

Homology between the data used to stop the training and the rest of the training set in session 1 will lead to an overestimation of the performance.

We now try to evaluate the performance on a set where the data have not been used in training or test.

Q11: How many data points are in this evaluation set?

Number of evaluation data: 212

Q12: What is the Pearson's correlation coefficient for this evaluation set?

Pearson coefficient for N= 212 data: 0.81467

Here we have used the same training set as in the first training session.

Q13: What is the difference between the Pearson's correlation coefficient between the evaluation set and the test set?

It is much lower for the evaluation set

Q14: Would you expect the sequences in this evaluation set to be more or less similar to sequences used for training than the sequences in the test set?

Less similar

Q15: How many data points are in this evaluation set?

212 data (same set as before)

Q16: What is the Pearson's correlation coefficient for this evaluation set?

Pearson coefficient for N= 212 data: 0.77605

Q17: How many data points are in this evaluation set?

212 data (same number but not the same set as before)

Q18: What is the Pearson's correlation coefficient for this evaluation set?

Pearson coefficient for N= 212 data: 0.89324

Q19: What could be the reason for the difference in correlation coefficients between the two evaluation sets?

Q20: Which set will be more similar to the training set?

Eval2 is more similar to train2 than eval1.

Use the evaluation set you think is the more homologous to this training set to evaluate the training in session 1.

Compare the test set and evaluation set values.

Test:

Maximal test set pearson correlation coefficient sum = 0.910800

Eval:

Pearson coefficient for N= 212 data: 0.93576

Q21: Would you be surprised to know that this evaluation set and test set are nearly identical?

You shouldn't

Q22: Why (not)?

Because the Pearson correlation of both sets are comparable and very high