7 Model evaluation

7.1 Evaluation of binary classifiers

Binary classifiers are mathematical or computational models that classify an input data set and produce the output with two labels.

Evaluation of models

The performance of different models can be evaluated under the same test dataset.

- Algorithms
- Scoring schemes
- Statistical analysis

Test data

It should contain both homologous and non-homologous alignments.

- Positive: homologous
- Negative: non-homologous

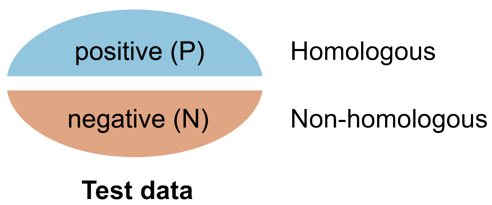


Figure 7.1: Test dataset for homologous and non-homologous

Model output

Different models often output different formats of scores.

- Raw scores, bit scores, z-scores
- P-values, e-values

Threshold values are used to separate the result into positives and negatives.

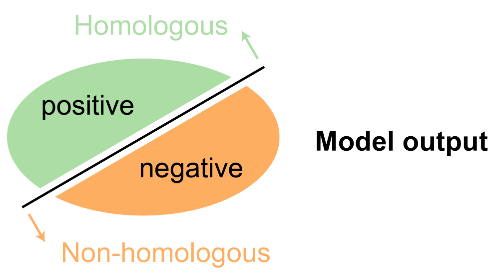


Figure 7.2: Model output for homologous and non-homologous

7.2 Confusion matrix

The output of a model produces two false and two correct classifications.

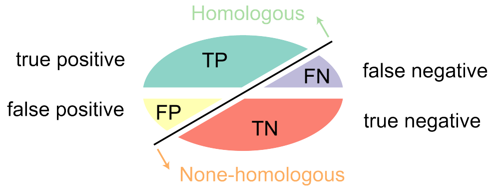


Figure 7.3: Four outcomes of model classification

Example of model output

A test dataset contains 10 positives and 10 negative.

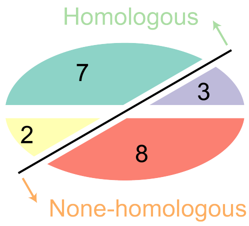


Figure 7.4: An example of the four outcomes

- 7 true positives
- 8 true negatives
- 2 false positives
- 3 false negatives

Confusion matrix

The classification result can be formed into a matrix format.

Table 7.1: Confusion matrix

		Test data	
		Homologous	Non-homologous
Model classification	Homologous	TP	FP
	Non-homologous	FN	TN

Example of confusion matrix

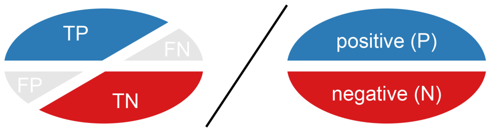
7 TPs	2 FPs
3 FNs	8 TNs

7.3 Basic evaluation measures

Various measures can be derived from the confusion matrix.

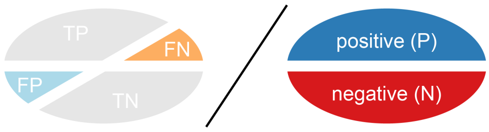
Accuracy

$$\frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N}$$



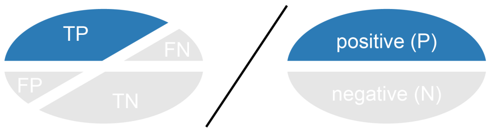
Error rate

$$\frac{FP + FN}{TP + FP + TN + FN} = \frac{FP + FN}{P + N}$$



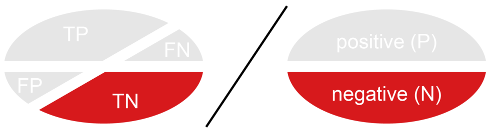
Sensitivity, True positive rate, Recall

$$\frac{TP}{TP + FN} = \frac{TP}{P}$$



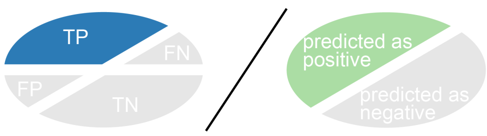
Specificity, True negative rate

$$\frac{TN}{FP + TN} = \frac{TN}{N}$$



Precision, Positive predictive value

$$\frac{TP}{TP + FP}$$



7.4 Measures with multiple thresholds

The test data set needs to be sorted by scores, and then confusion matrices can be calculated for multiple threshold values.

Example of making confusion matrices with multiple thresholds

Test data set

Label	N	P	P	N	N	N	P	P	P	N	P	N	P	P	N	N	P	P	N	N
Score	27	4	17	9	11	2	15	19	22	3	23	7	10	25	11	1	26	28	24	3

Sorted test data set

Label	P	N	P	P	N	P	P	P	P	P	N	N	P	N	N	P	N	N	N	N
Score	28	27	26	25	24	23	22	19	17	15	11	11	10	9	7	4	3	3	2	1
Threshold			↑ 1						↑ 2							↑ 3				

1st threshold (score = 25.5)

2 TPs	1 FPs
8 FNs	9 TNs

2nd threshold (score = 16)

7 TPs	2 FPs
3 FNs	8 TNs

3rd threshold (score = 3.5)

10 TPs	6 FPs
0 FNs	4 TNs

ROC and precision-recall

These measure are based on the confusion matrices of all possible threshold values.

- ROC (Receiver operating characteristic) plot
- Precision-recall plot

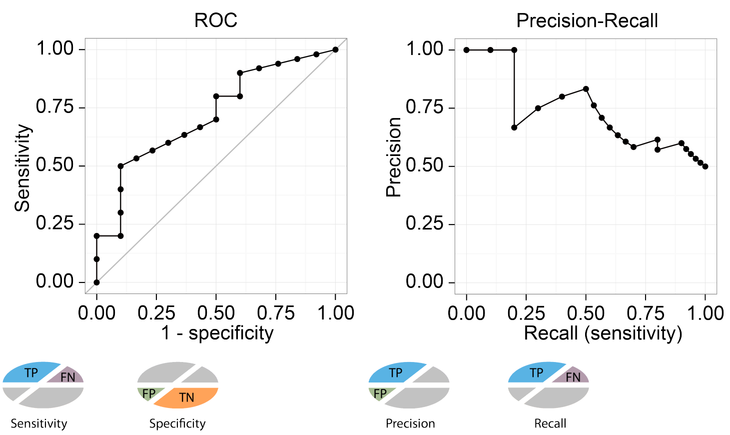


Figure 7.5: ROC and precision-recall plots

Exercise 7.1

Draw an ROC curve for the following specificity and sensitivity values.

Threshold	Specificity	1 - Specificity	Sensitivity
10	1	0	0
9	0.8	0.2	0.8
8	0.6	0.4	0.8
7	0.6	0.4	1
6	0.4	0.6	1
5	0.2	0.8	1
4	0	1	1