Evaluation of classifiers quality

Supervised classification finds and assess predictive genes

Goal: find a set of genes (or metagenes, principal components, etc.) which predicts some biological or clinical characteristics of the samples (e.g., is sample X a tumor?).

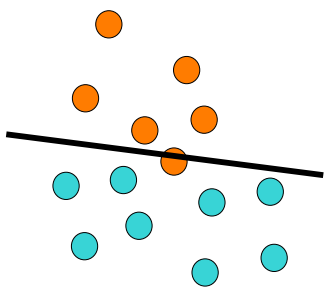
Two tasks are carried out:

1. Training. Using samples whose class is known, construct a classifier—i.e., a procedure that tells to which class a sample belongs.
2. Testing. Run the classifier on samples *not* used for training and see if it makes correct class predictions.

Linear classifiers

Linear methods find classification rules such as *'if genes A and B are up and genes C, D, and E are down, then sample is a tumor'*.

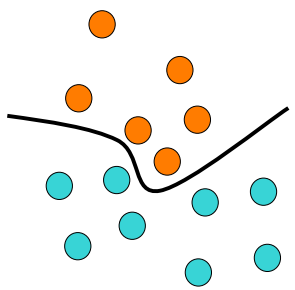
Popular algorithms include diagonal discriminants, decision trees, Fisher's discriminants, etc.



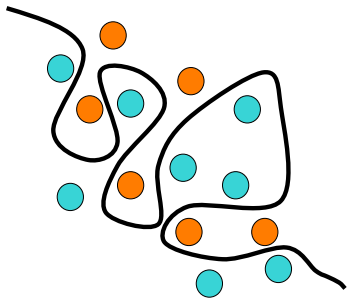
Nonlinear classifiers

Nonlinear methods find possibly complex mathematical functions separating the classes. They are typically not biologically interpretable.

Popular algorithms include neural networks and support vector machines.



Overfitting



- Overfitting is overlearning producing classifiers that do not generalize well
- It is *doomed* to happen when the number of variables (genes) exceed the number of points (samples), it is called the *curse of dimensionality*
- Overfitting is much more likely to occur when complex classification rules (e.g. nonlinear) are being learned

5

Evaluation criteria: accuracy

classification \ reality	true	false
true	TP	FP
false	FN	TN

Accuracy
= probability of errors
= (FN+FP)/N
- N=TP+TN+FP+FN
- misleading when classes are unbalanced

6

Evaluation criteria: sensitivity

classification \ reality	true	false
true	TP	FP
false	FN	TN

Sensitivity
= probability detecting a true positive
= TP/(FN+TP)
- E.g. probability that the classifier correctly detects cancers

7

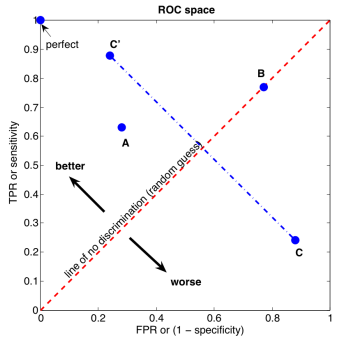
Evaluation criteria: specificity

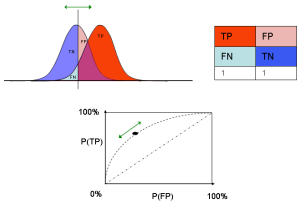
classification \ reality	true	false
true	TP	FP
false	FN	TN

Specificity
= probability of not detecting a true negative
= TN/(TN+FP)
- E.g. probability that the classifier correctly detects non cancers

8

Evaluation criteria: receiver operating characteristic

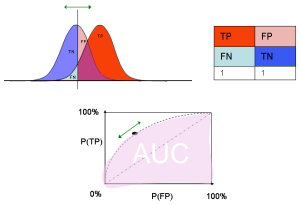




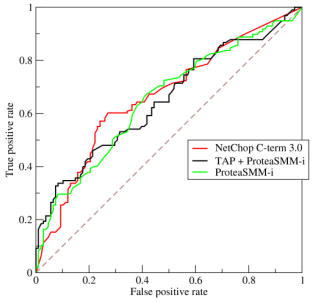
(from Wikipedia)

9

Evaluation criteria: area under the ROC



(from Wikipedia)



10

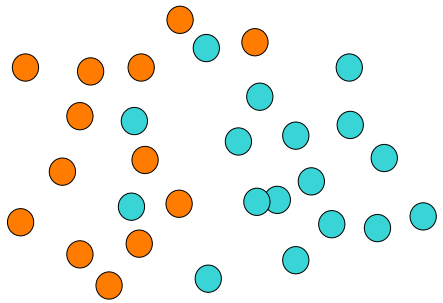
The fundamental rule of classifier evaluation

Training and testing must be *absolutely* separated

seems simple in theory, but errors are easily made in practice when dealing with complex classication algorithms (there are countless flawed studies)

11

K-fold cross-validation provides an objective assessment of classifiers

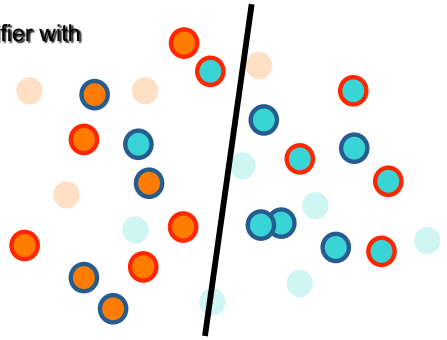


K-fold cross-validation provides an objective assessment of classifiers

Split samples in *K* groups

K-fold cross-validation provides an objective assessment of classifiers

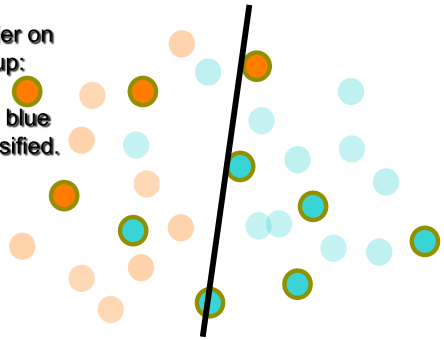
Train classifier with *K*-1 groups



K-fold cross-validation provides an objective assessment of classifiers

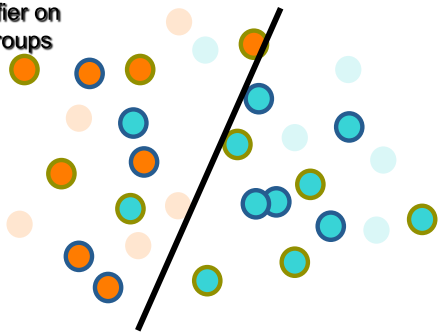
Test classifier on left-out group:

1 red and 1 blue are misclassified.



K-fold cross-validation provides an objective assessment of classifiers

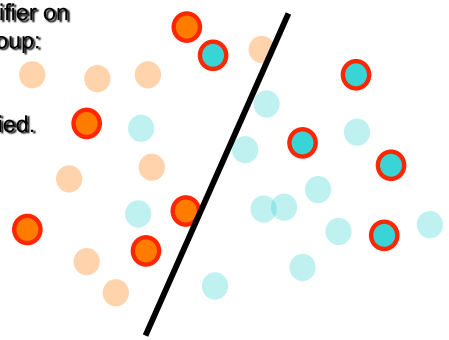
Train classifier on other *K*-1 groups



K-fold cross-validation provides an objective assessment of classifiers

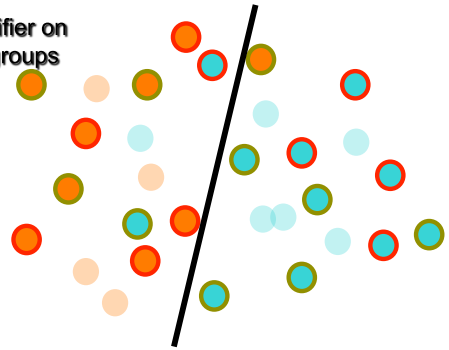
Test classifier on left-out group:

1 blue is misclassified.



K-fold cross-validation provides an objective assessment of classifiers

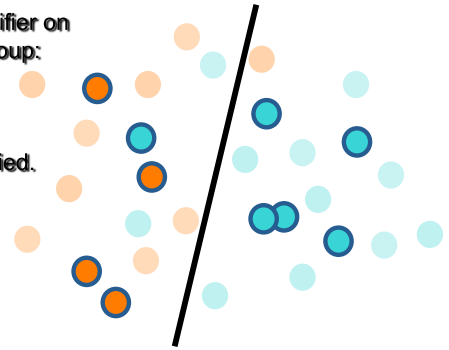
Train classifier on other *K*-1 groups



K-fold cross-validation provides an objective assessment of classifiers

Test classifier on left-out group:

1 blue is misclassified.



K-fold cross-validation provides an objective assessment of classifiers

- ✓ **Error measures are averaged over the *K* runs of the training/test cycle.**
- ✓ **The whole procedure may be averaged over many random *K*-partitions of the data.**
- ✓ **Run whole procedure on randomly permuted data to get *p*-values.**
- ✓ **Total separation between testing and training data must be maintained over the complete procedure.**

Feature selection bias results in grossly overestimated prediction performances

Typical ***incorrect*** feature selection (biased) set up

1. select the *N* genes most differentially expressed between condition A and B using the entire dataset

2. tune classifier with CV using the *N* genes

Problem: genes selection is in fact supervised learning! Training and testing are not separated. Overfitting is *not* ruled out

Ambroise & McLachlan, (2001), *PNAS* 99, 6562-6.

Feature selection bias results in grossly overestimated prediction performances

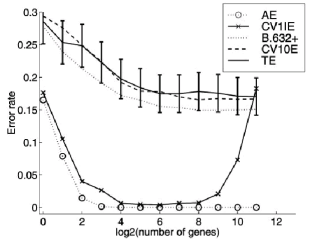


Fig. 1. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each. TE, test error.

Ambroise & McLachlan, (2001), *PNAS* 99, 6562-6.

Many authors have used cross-validation, while performing gene selection on the whole data set.

This resulted in small classification error, but it is wrong.

Feature selection bias results in grossly overestimated prediction performances

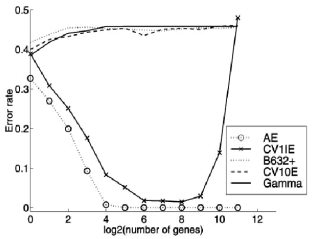


Fig. 5. Error rates of the SVM rule averaged over 20 noninformative samples generated by random permutations of the class labels of the colon tumor tissues.

Ambroise & McLachlan, (2001), *PNAS* 99, 6562-6.

Feature selection bias can lead to (erroneously) excellent performances although only noise is present.

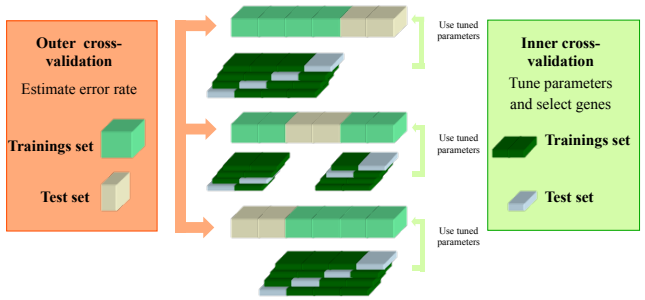
Data are overfitted.

There are other biases beyond feature selection

- Parameter selection bias: learning parameter are tuned through CV on the entire dataset
- Model selection bias: classification algorithm is selected from CV results resting on the entire dataset
- Ideally testing should be done on blinded data and result analysed by independent researchers

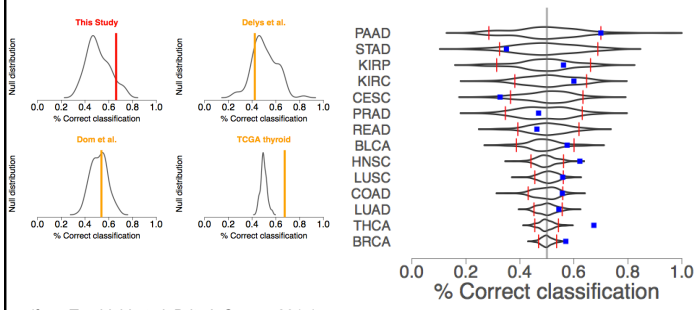
24

Nested CV avoids selection biases



Ruschhaupt *et al.* Stat. Application Genet. Mol. Biol. 2005

Error variance is large for small datasets



(from Tarabichi et al. Brit. J. Cancer 2015)

26

Rules of thumb

- Think negative control: run your code on sample-permuted data
- When tractable, estimate the p-value of your classification results
- Try to get enough data for completely independent testing (and avoid CV)

27