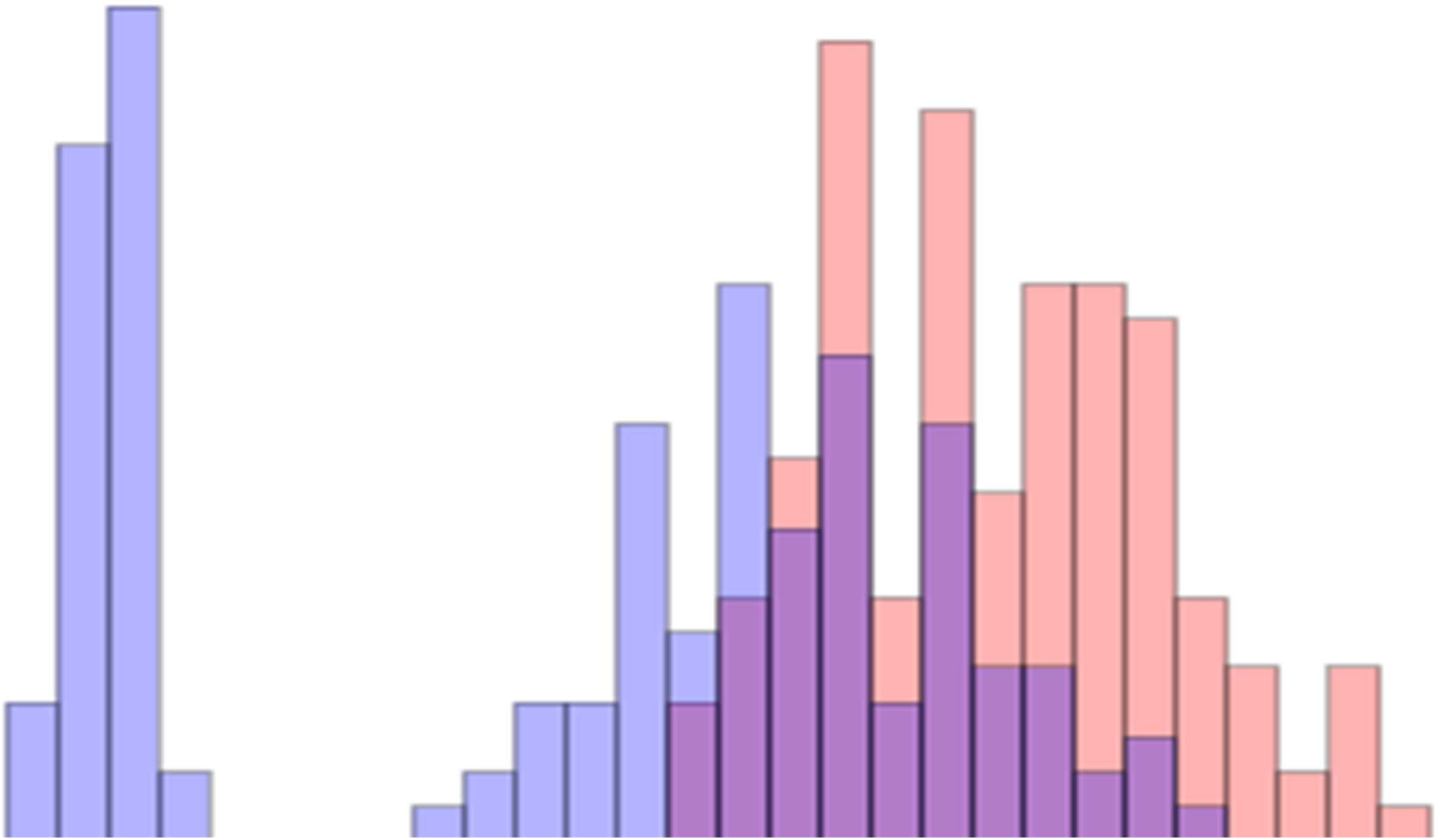


True or False?

- Classification error of the training set is a good measure of the true classification error

- A good estimate of the actual, true error is all we are interested in when building a classification system

Evaluation

› Marco Loog

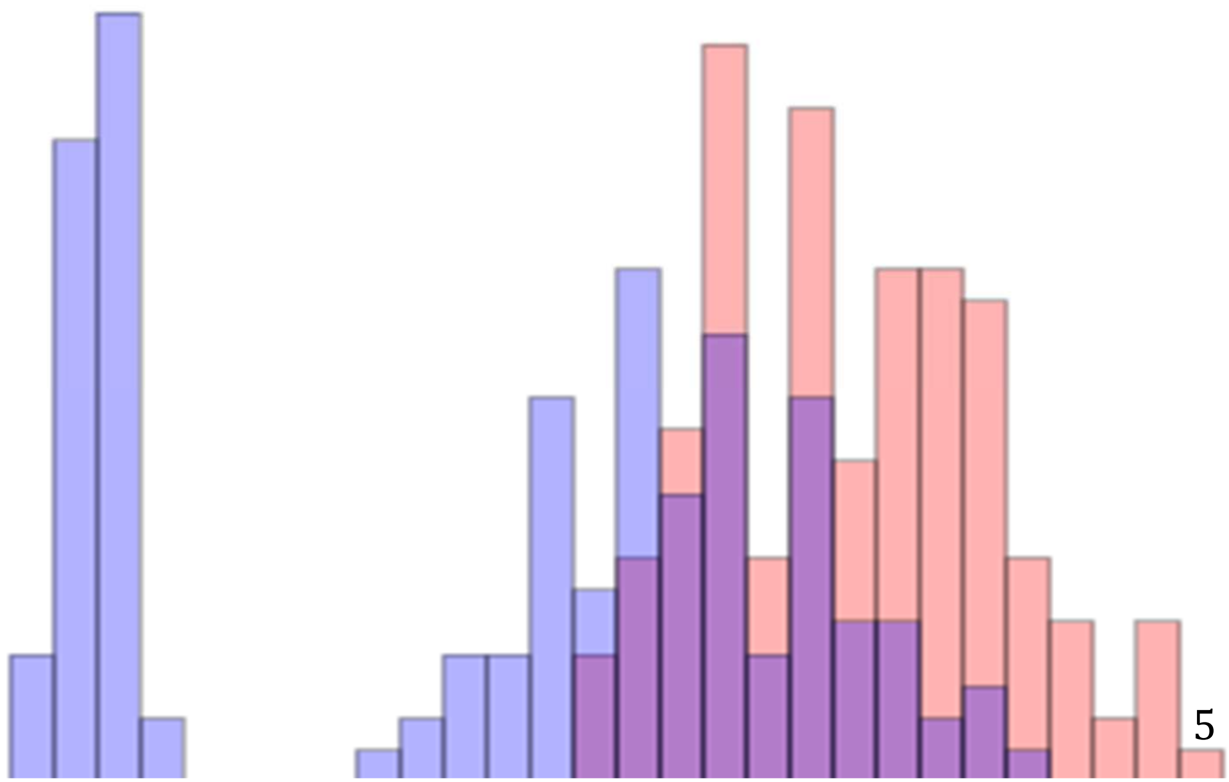
Past

- Regularization
- Bias-variance

- How classifiers and regressers can overtrain…

Let’s Recap a Bit

› Bias-variance for histogram …



Present

› Evaluating learners [focus is on classifiers]
  - Cross validation
  - Learning curves
  - Feature curves
  - Complexity curves
  - Confusion matrix

› Curse of dimensionality
  - Bias / variance

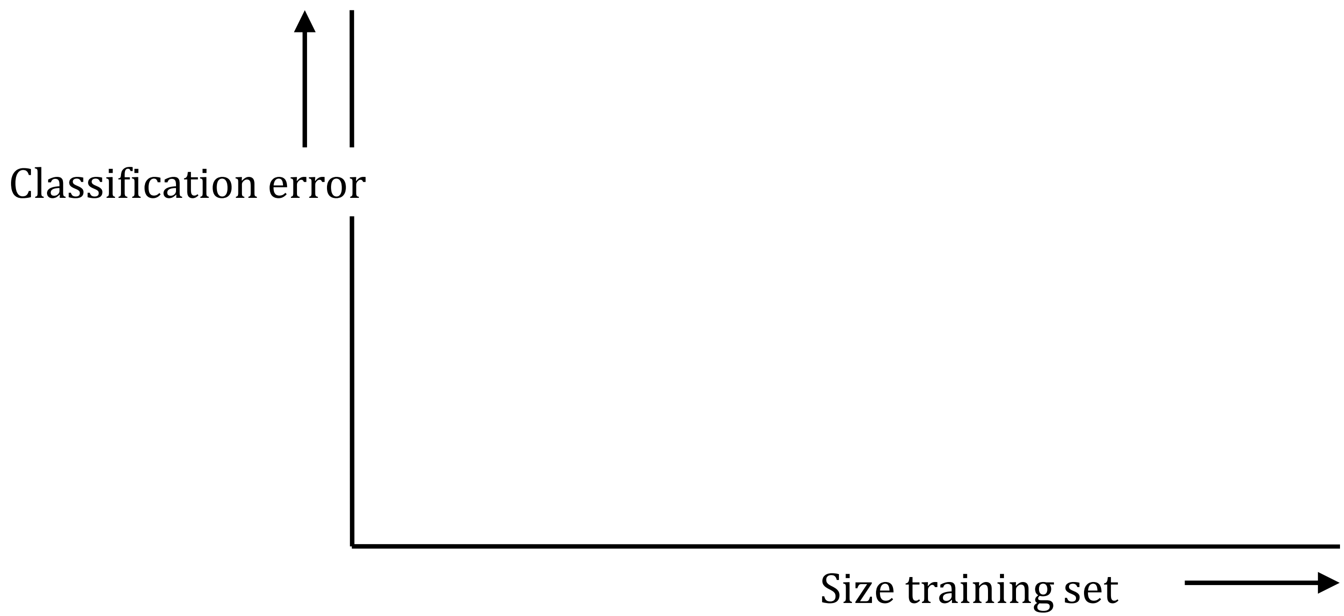
Determining Errors

Determining Errors

› Apparent or resubstitution error
= error classifier makes on its training data set

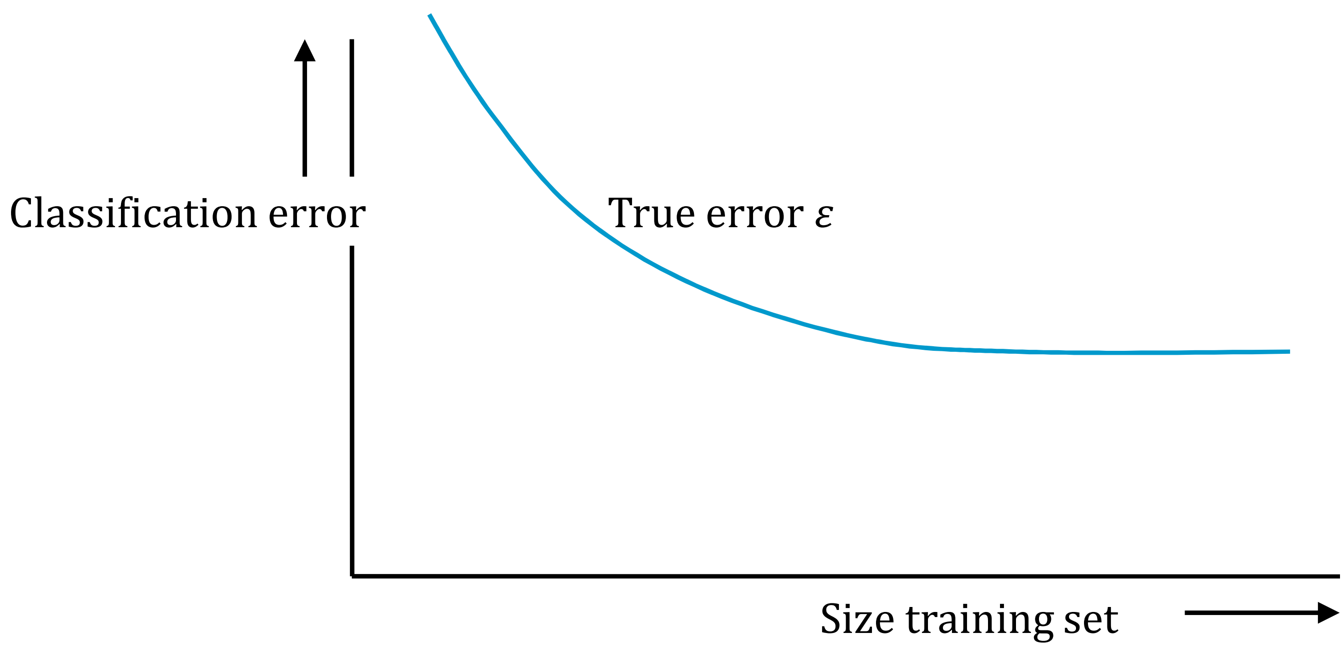
True Classification Error

› How does it behave?



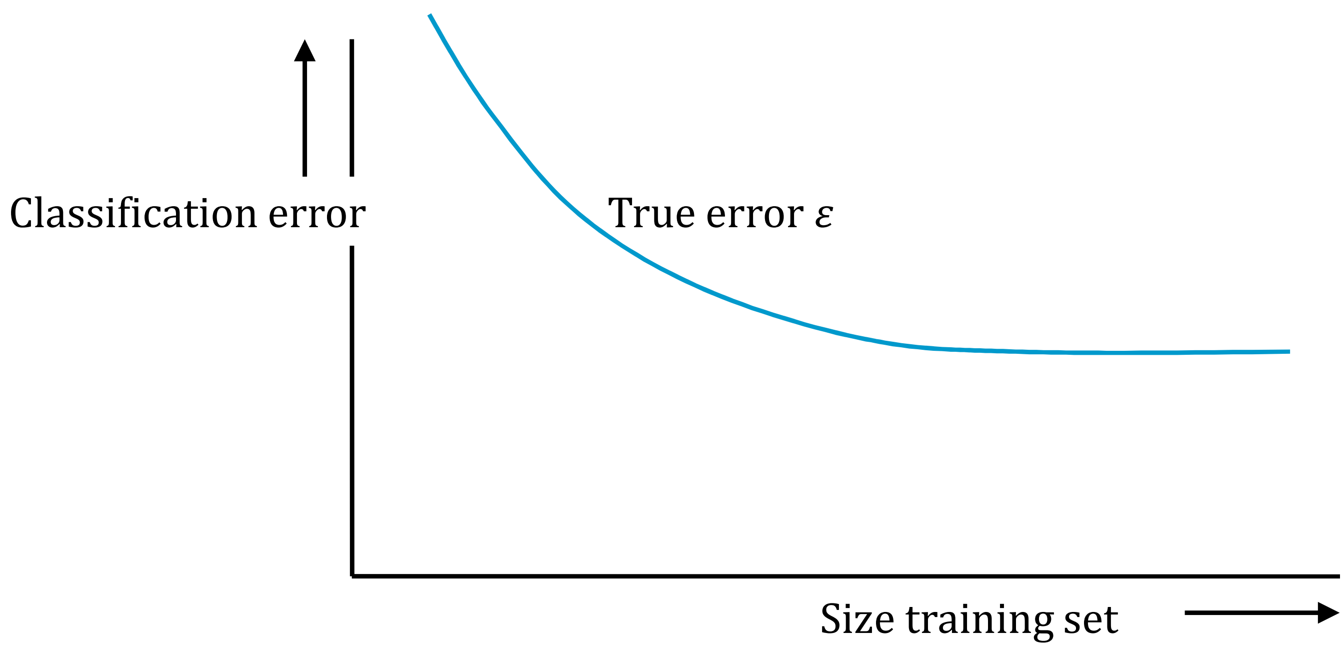
True Classification Error

› How does it behave?

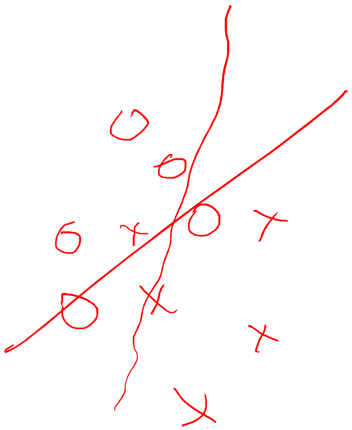


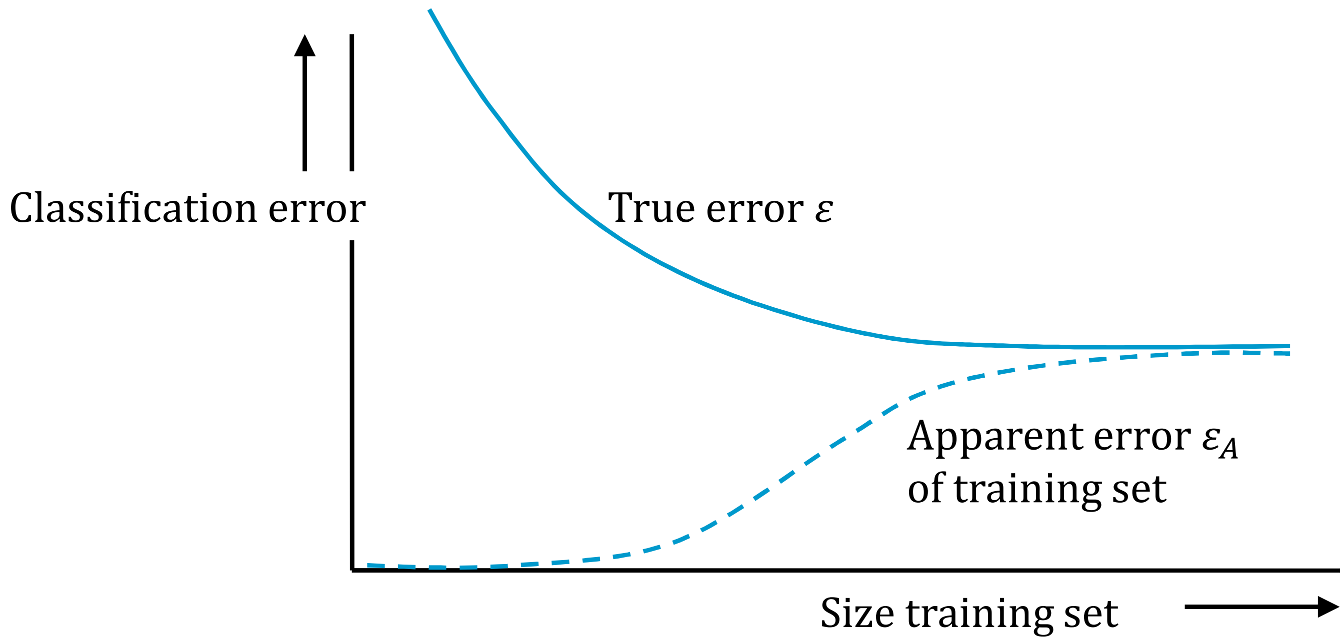
True Classification Error

› How does apparent error behave?



Apparent Classification Error



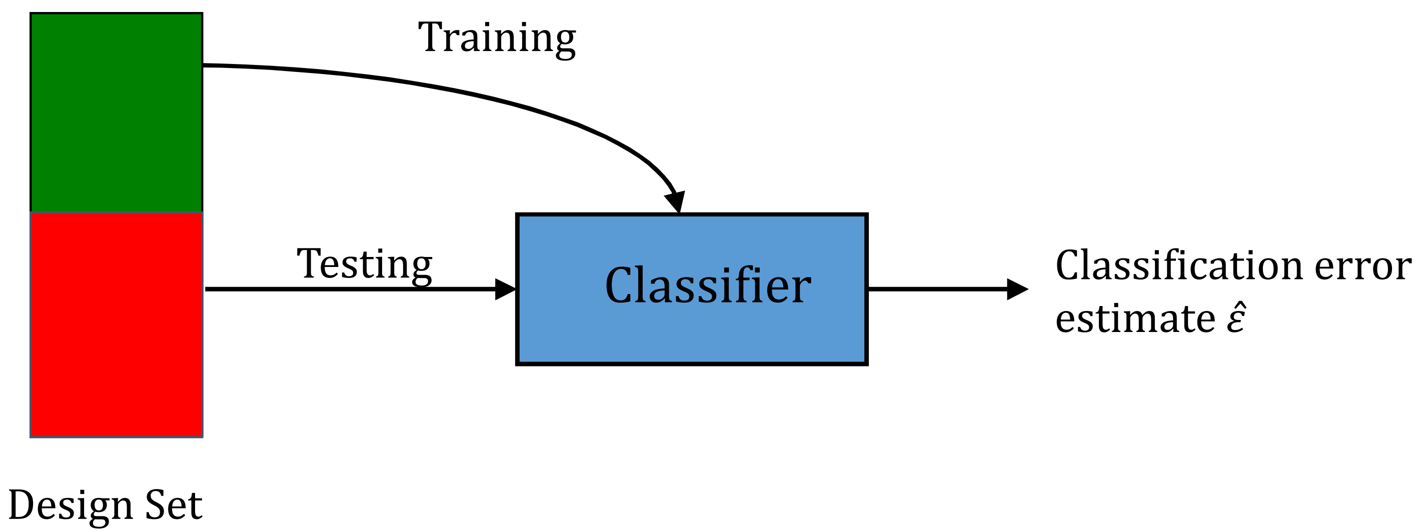


Determining Errors

- Apparent or resubstitution error
  = error classifier makes on its training data set

- How do we determine the true error in practice?

Error Estimation by Test Set



Other training set → other classifier
Other test set → other error estimate

How Variable?

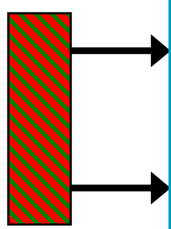
$$\sigma_{\hat{\varepsilon}}^2 = \mathrm{Var}(\hat{\varepsilon} \mid \text{test set size } N) = \frac{\varepsilon(1-\varepsilon)}{N}$$

$$\sigma_{\hat{\varepsilon}} = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}}$$

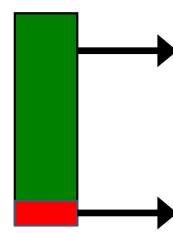
N \ ε	0.01	0.03	0.1
10	0.031	0.054	0.095
100	0.010	0.017	0.030
1000	0.003	0.005	0.009

Training Set Size vs. Test Set Size

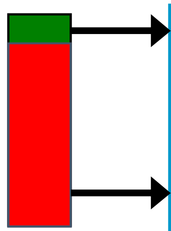
Large training set → good classifiers
Large test set → reliable, unbiased error estimate
In practice often just a single design set is given



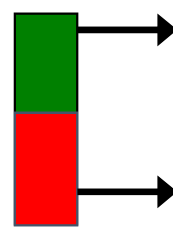
Same set for training and testing may give a good classifier, but will definitely yield an optimistically biased error estimate



A small, independent test set yields an unbiased, but unreliable [large variance] error estimate for a well trained classifier



A large, independent test set yields an unbiased and reliable [small variance] error estimate for a badly trained classifier

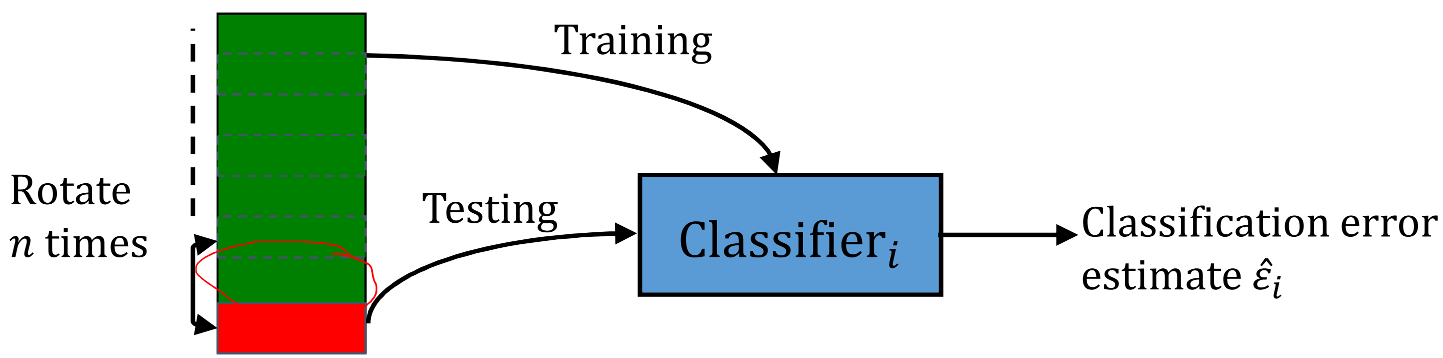


50-50 is an often used standard solution for the sizes for training and testing

But is it really good?

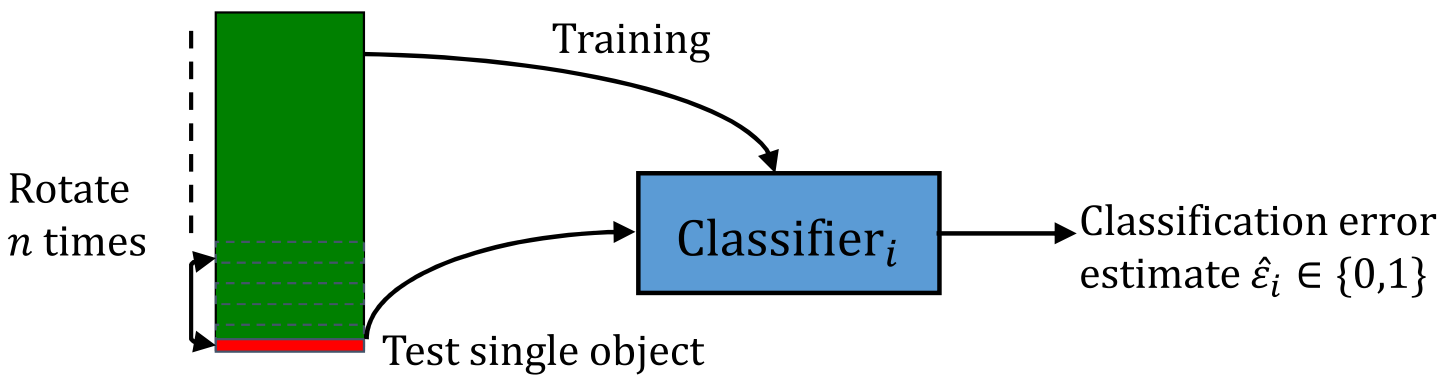
Cross Validations

Cross Validation

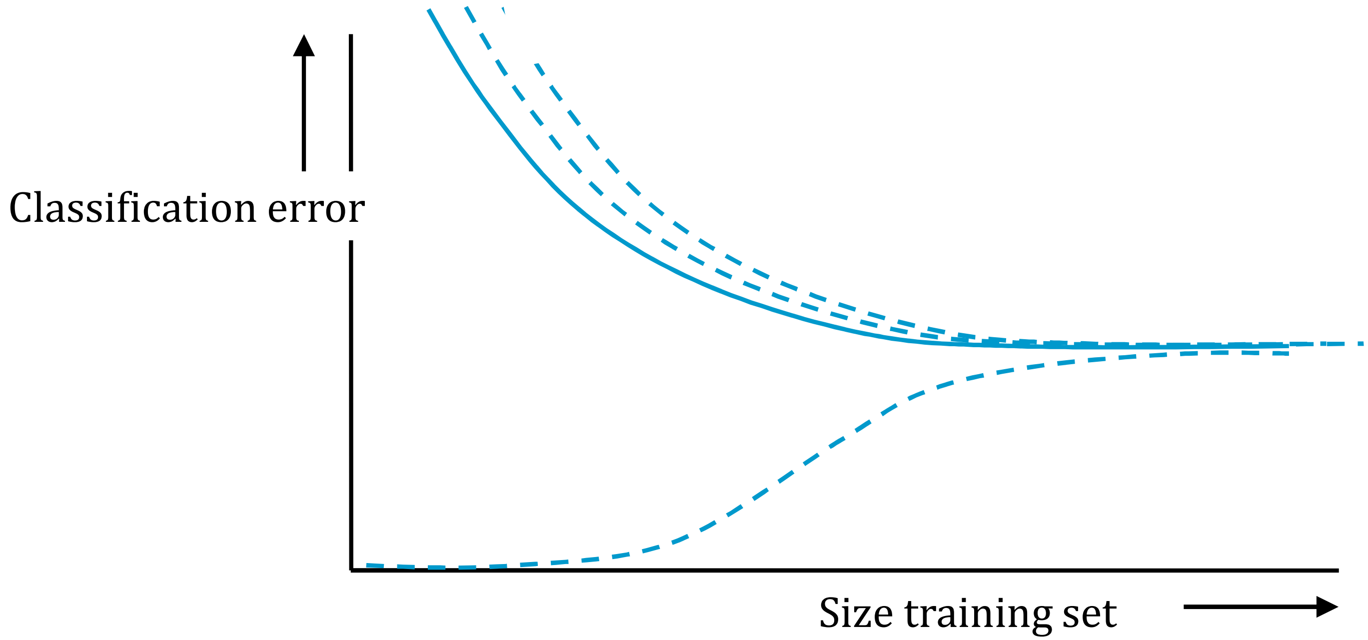


Leave-one-out Procedure

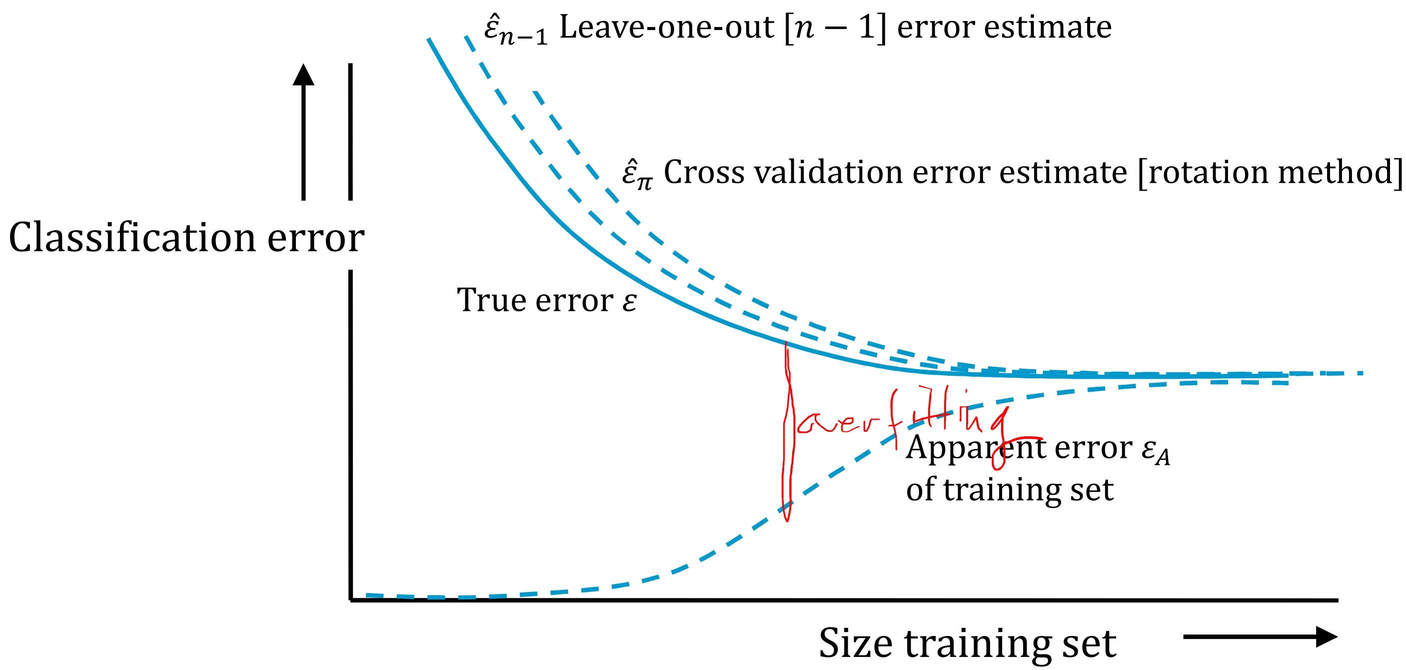
› n equals training set size



Cross Validation Curves and Related



Cross Validation Curves and Related

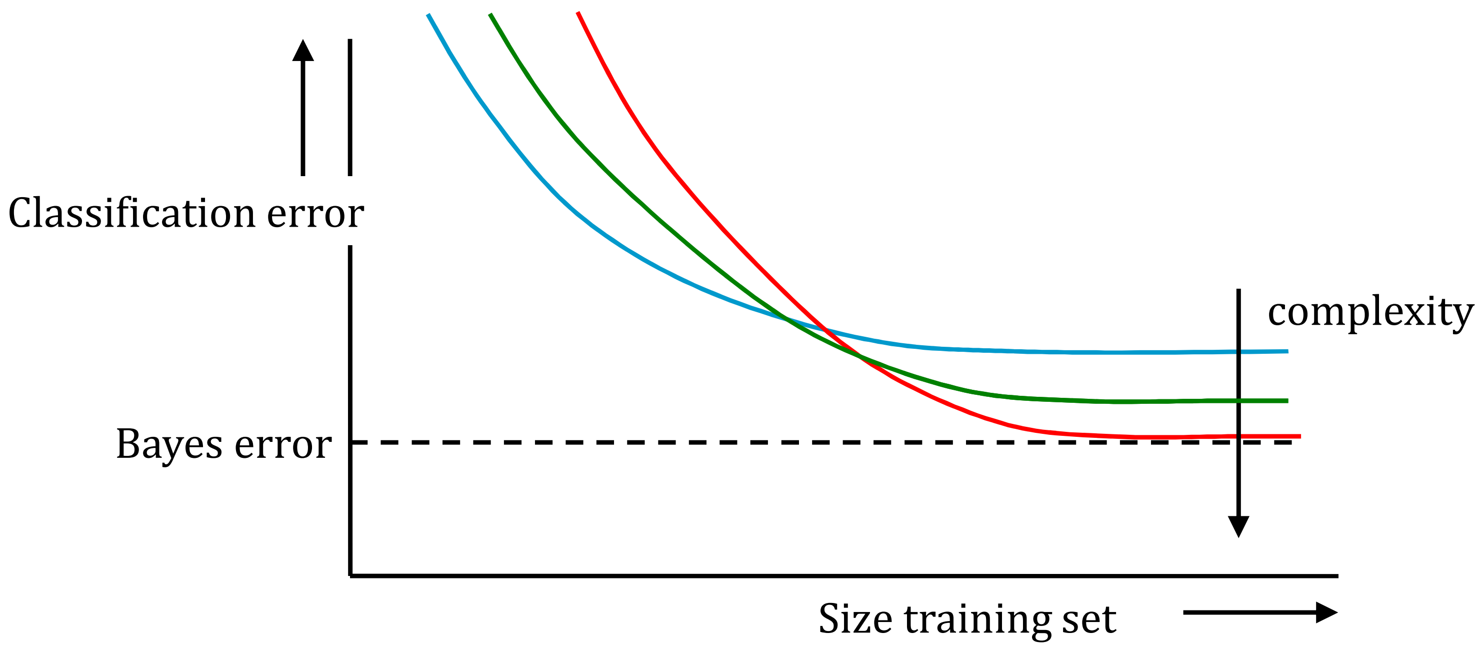


Learning Curves

Learning Curves

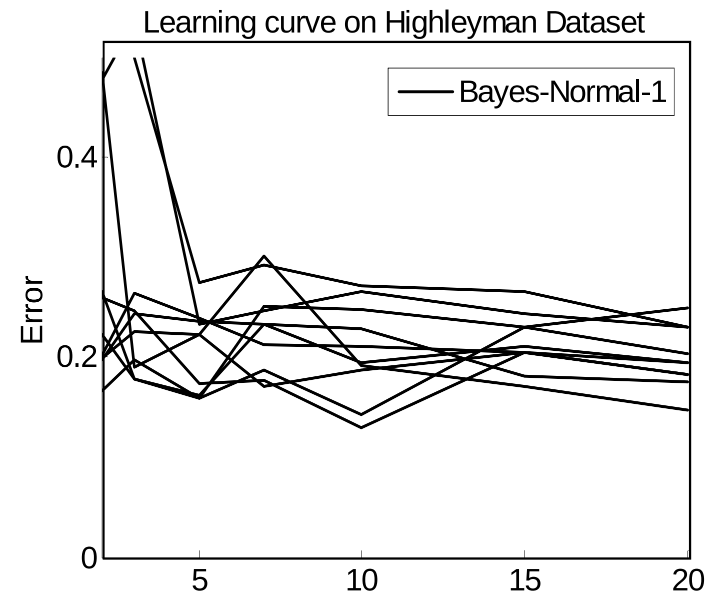
- Curves that plot [estimated] classification errors against the number of samples in training set
  - Usually plot error both on training and on test set
  - Gives insight, e.g. into
    - Amount of overtraining
    - Usefulness of additional data
    - How different classifiers compare
    - ...

Different Classifier Complexities



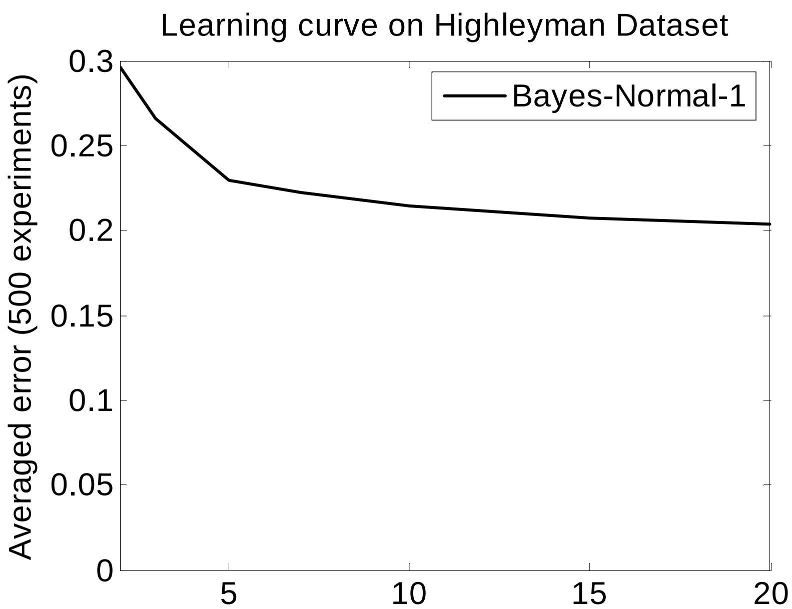
Real-world Learning Curves

› Small sample sizes have a large variability




```
a = gendath([200 200]);
for j=1:10
  e = cleval(a,ldc,[2,3,5,7,10,15,20],1);
  hold on; plote(e);
end
```

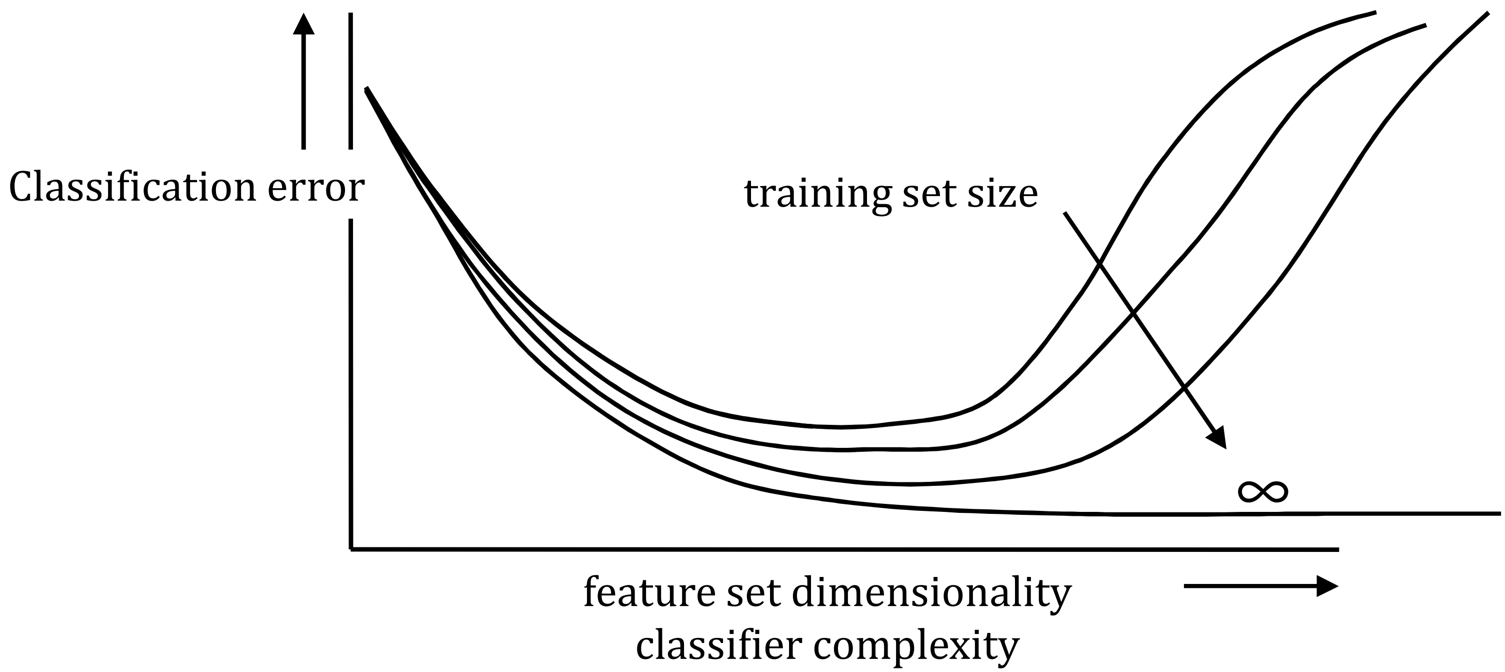
Averaged Learning Curve



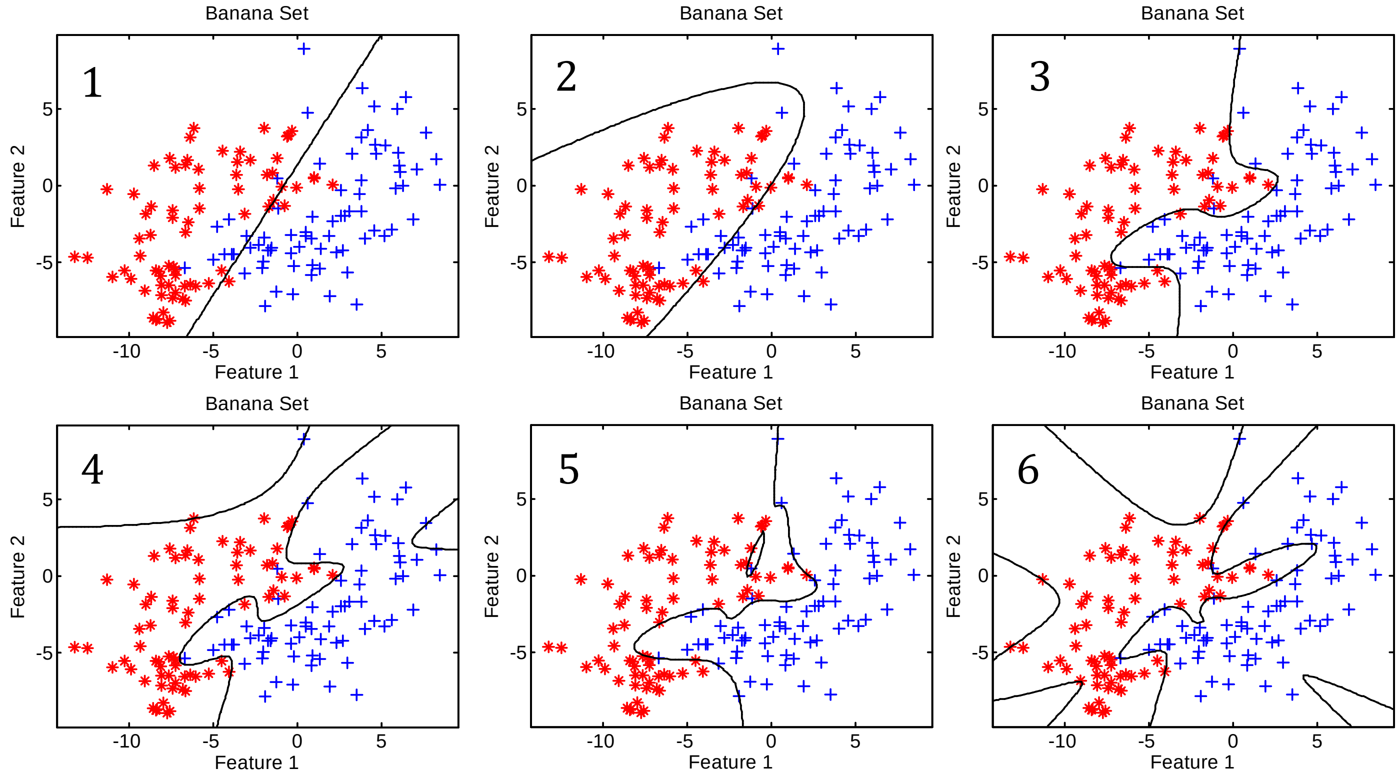

```
a = gendath([200 200]);
e = cleval(a,ldc,[2,3,5,7,10,15,20],500);
plote(e);
```

Feature Curves

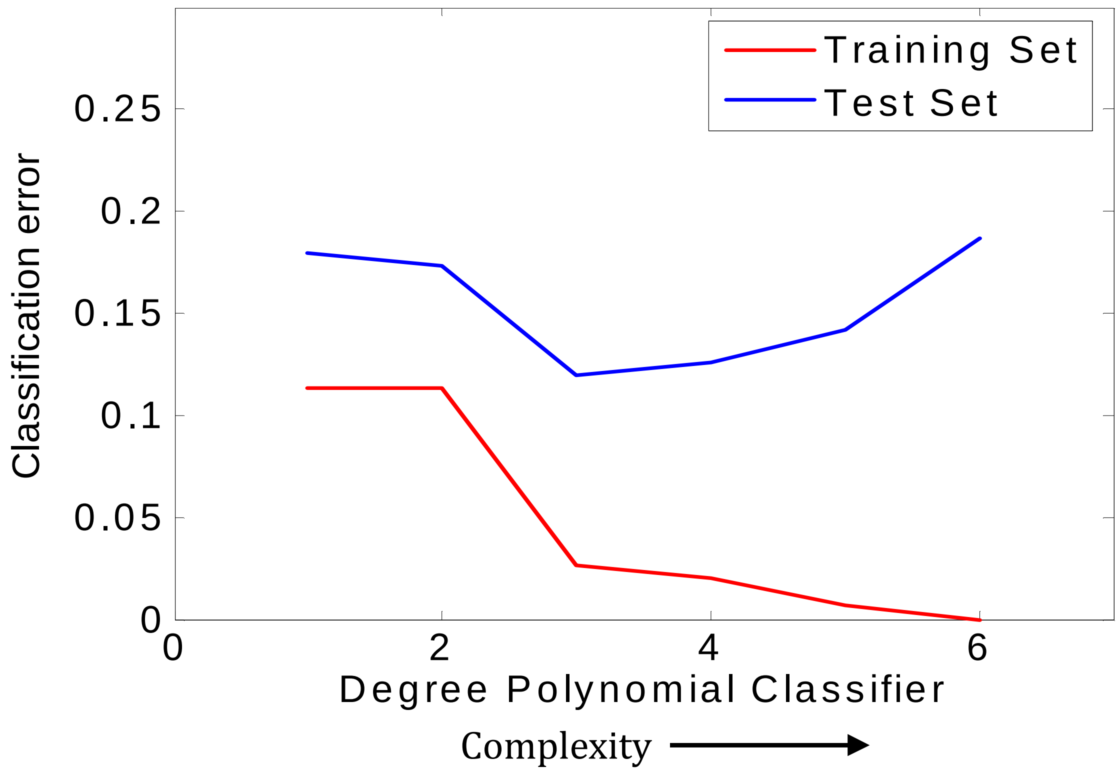
Feature Curves



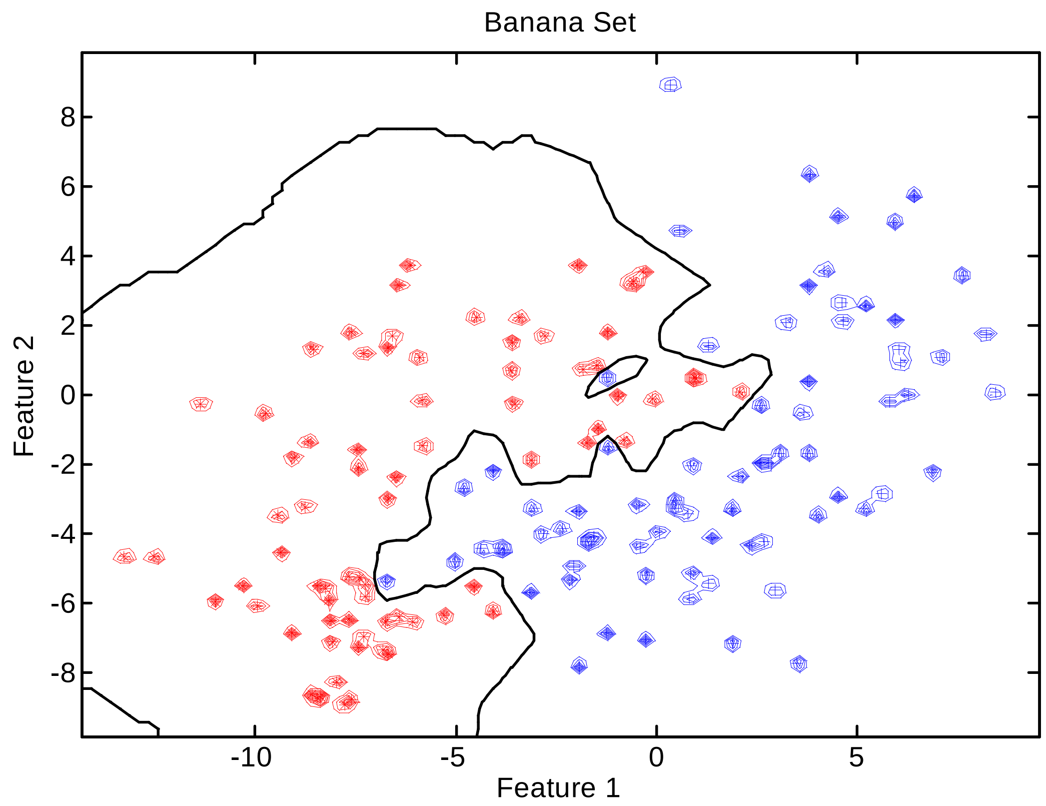
Polynomial Complexity Example

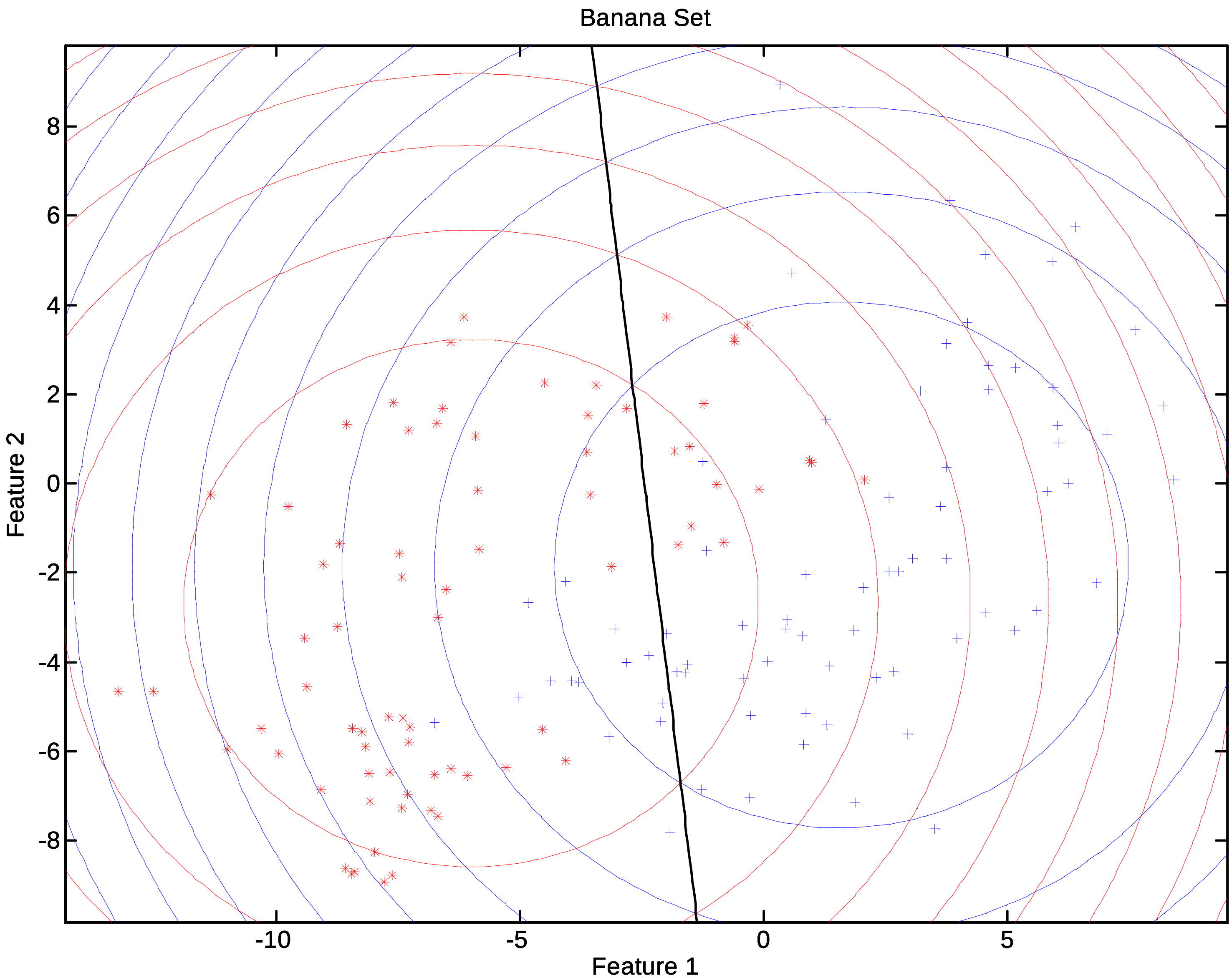


Polynomial Complexity Example



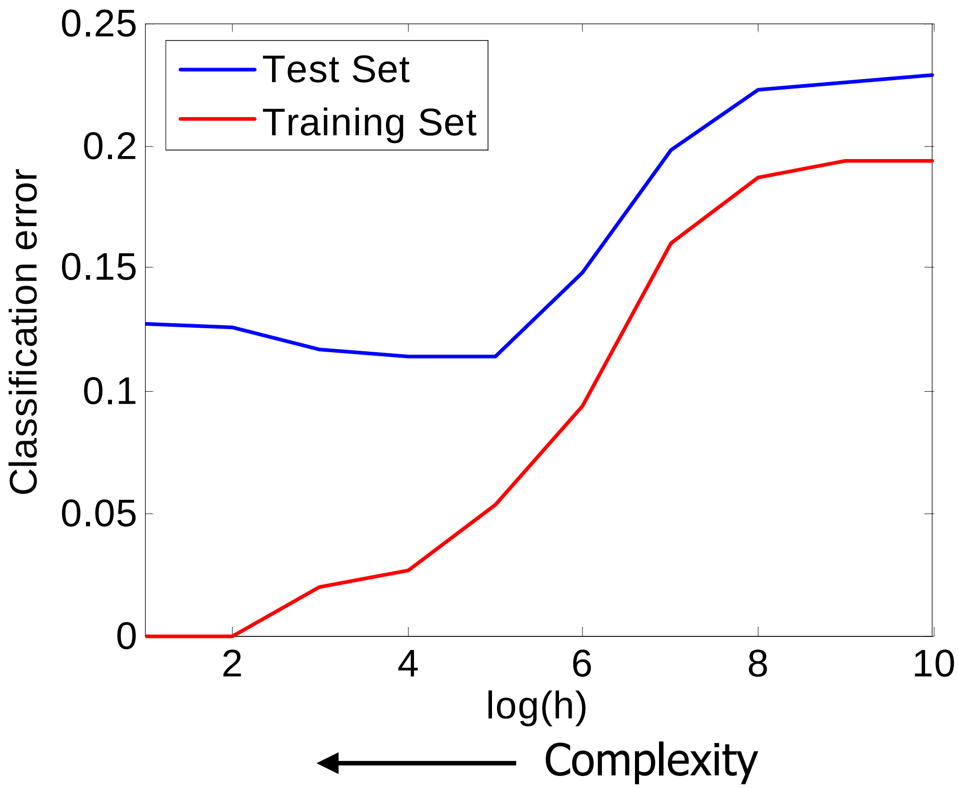
Parzenc Complexity Example





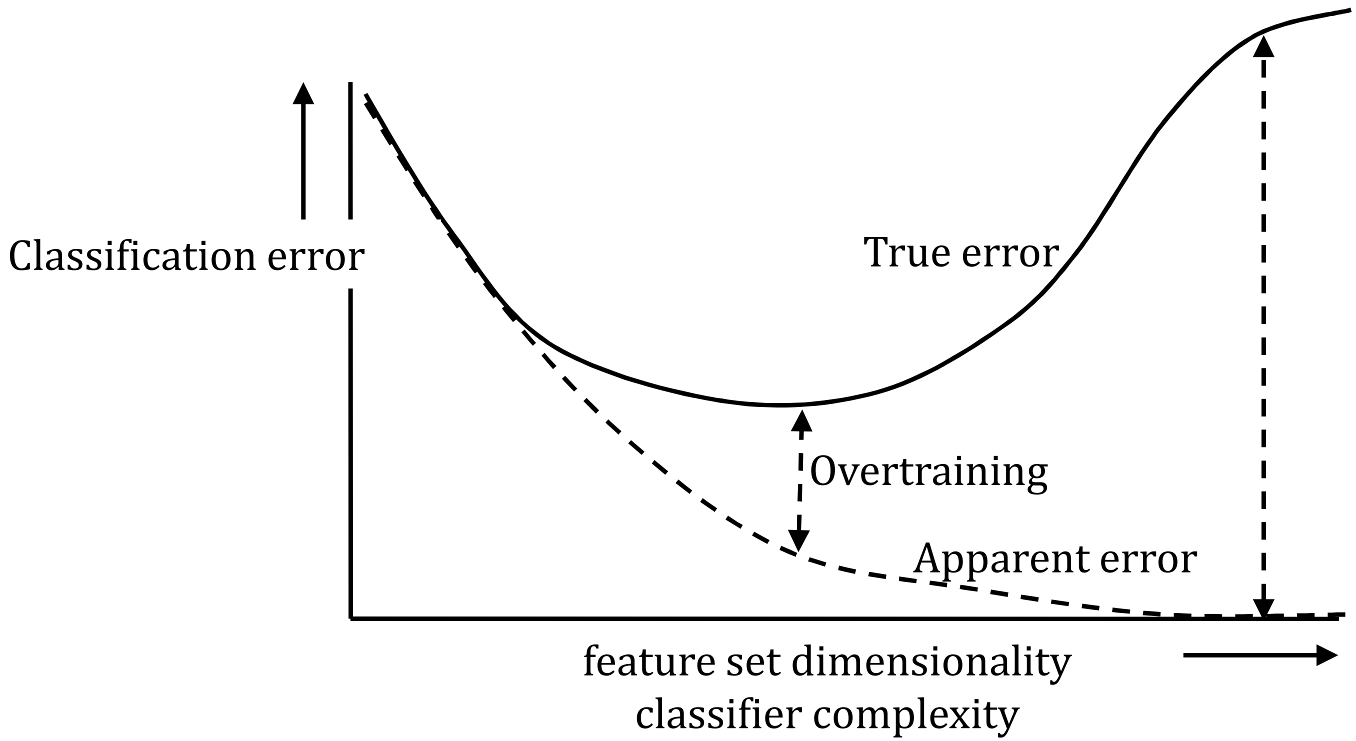
Banana Set
Feature 2
Feature 1
8
6
4
2
0
-2
-4
-6
-8
-10
-5
0
5

Parzenc Complexity Example



Curse of Dimensionality

Curse of Dimensionality



Some Concluding Claims…

- Larger training sets yield better classifiers
- Independent test sets needed for unbiased error estimates
- Larger test sets yield more accurate error estimates
- LOO cross validation “optimal”, but might be infeasible
- More complex classifiers need larger training sets
  - Same holds for larger feature set sizes
- Small training sets need simpler classifiers or smaller feature sets
- There is no single best classifier!

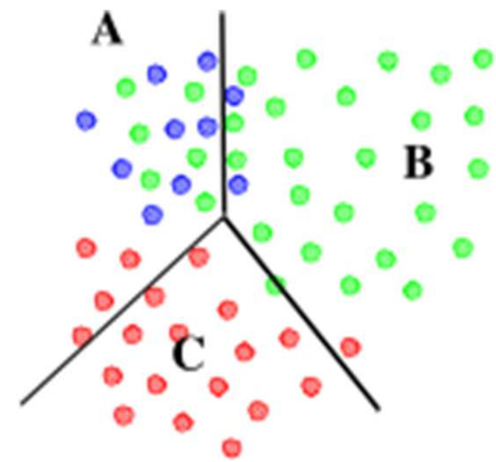
Confusion Matrices

› Provides counts of class-dependent errors : How many object have been classified as A that should have been classified as B?

Give a more detailed view than overall error rate

Can be used to estimate overall cost for particular classifier

Confusion Matrices



$$N_A = 10,\ N_B = 30,\ N_C = 20$$

$$E = \frac{c_{12} + c_{13} + c_{21} + c_{23} + c_{31} + c_{32}}{N_A + N_B + N_C}$$

$$E = 14/60 = 0.2333$$

$C = \text{confmat}(\Lambda, L)$

Λ real labels

L obtained labels

classified to

objects from	A	B	C
class A	8	2	0
class B	6	23	1
class C	4	1	15

0.20 error in class A
0.23 error in class B
0.25 error in class C

0.228 averaged error

› Q?