Methods for Intelligent Systems

Lecture Notes on Classifiers Evaluation and Comparison

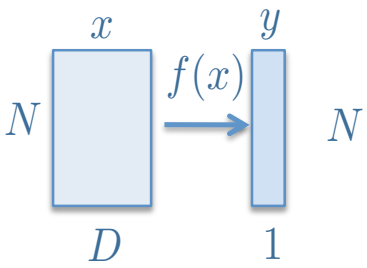
2009 - 2010

Simone Tognetti
tognetti@elet.polimi.it

Department of Electronics and Information
Politecnico di Milano

Classification Problems

Feature Extraction → Dimensionality Reduction → Classification



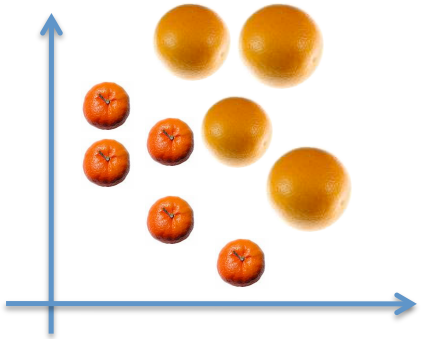
Feature Projection
- LDA
- PCA

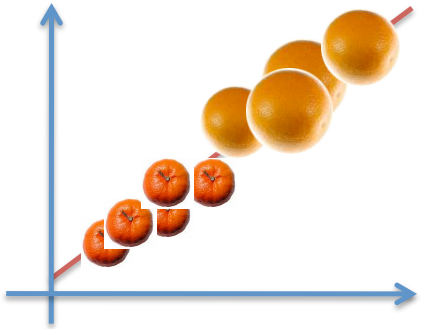
Feature Selection
- SFS,SBS
- Genetic Algorithm

- Linear functions
- Quadratic function
- Tree-based
- Logistic
- SVM
- K-NN

Evaluation

- Cross Validation



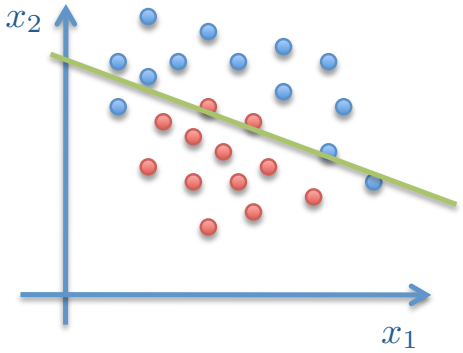


Type of classifiers(1)

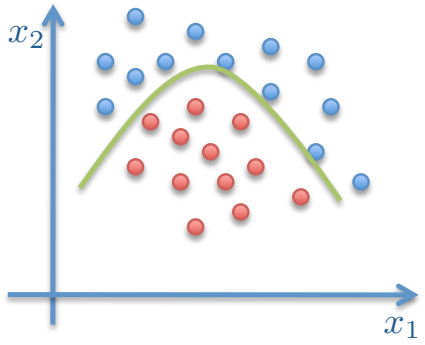
Define the function type that will be estimated from data $y = f(x, \bar{\theta})$

- Linear, Quadratic

$\theta_0 x + \theta_1 > 0$

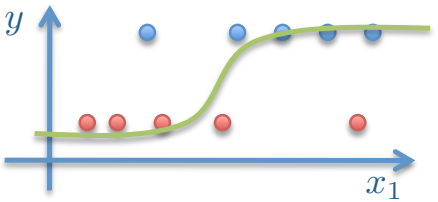


$\theta_0 x^2 + \theta_1 x + \theta_2 > 0$



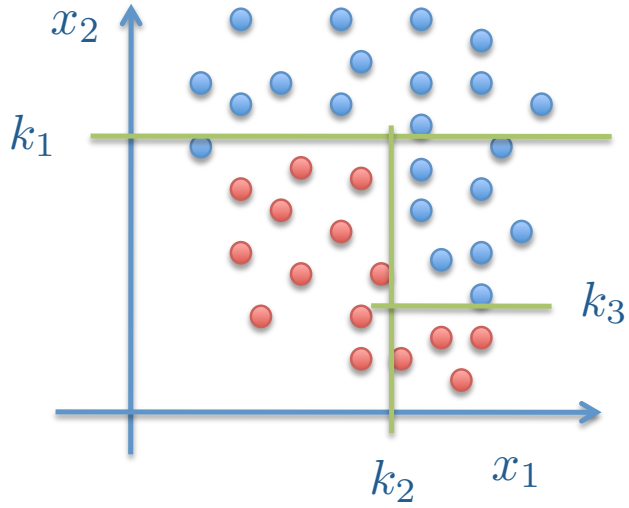
- Logistic

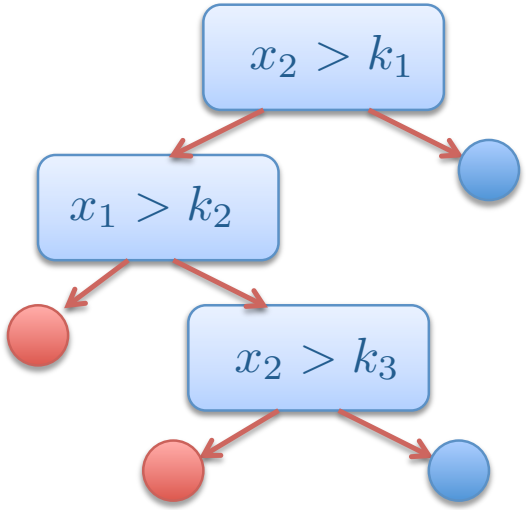
$$\frac{1}{1 + e^{-\sum_i \beta_i x_i}}$$



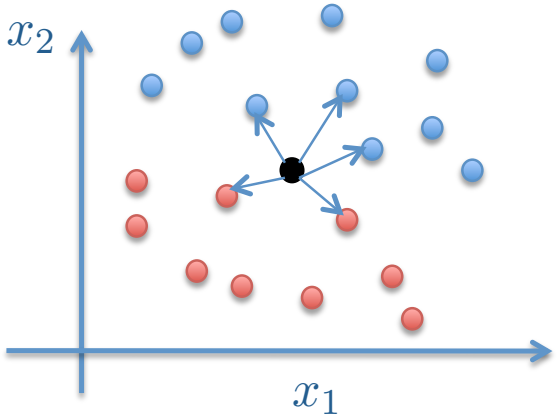
Type of classifiers(2)

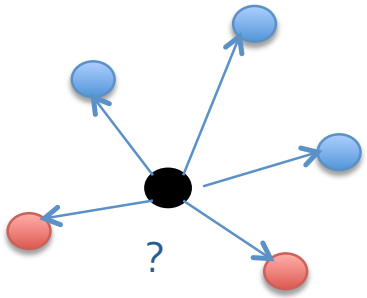
- Tree based (i.e. KD Tree)





- Instance based classifier: Classify samples according to its neighbors



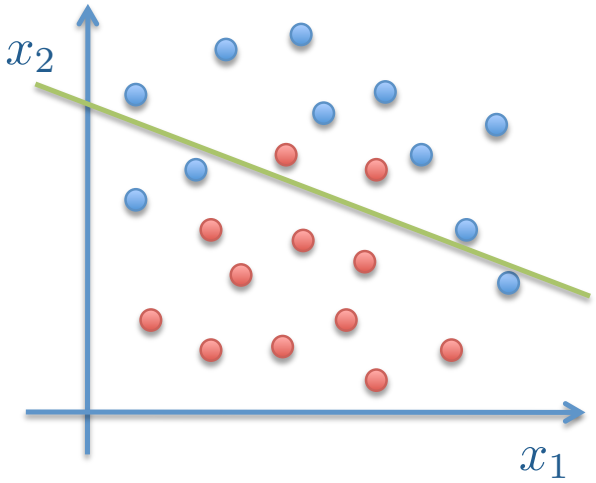


Consider K(=6) closest neighbors
Predicted class according to the majority of neighbors →

Classifier performance

- Confusion matrix

  Describes the classifier performance. Two class example



Original \ Predicted			
	8	3	11
	2	10	12
	10	13	23

$$C_f = \begin{bmatrix} 8 & 3 \\ 2 & 10 \end{bmatrix}$$

$$N = \sum_{i=1}^{D}\sum_{i=1}^{D} C_f(i,j) = 23 \qquad CC = \sum_{i=1}^{D} C_f(i,i) = 18 \qquad WC = N - CC = 5$$

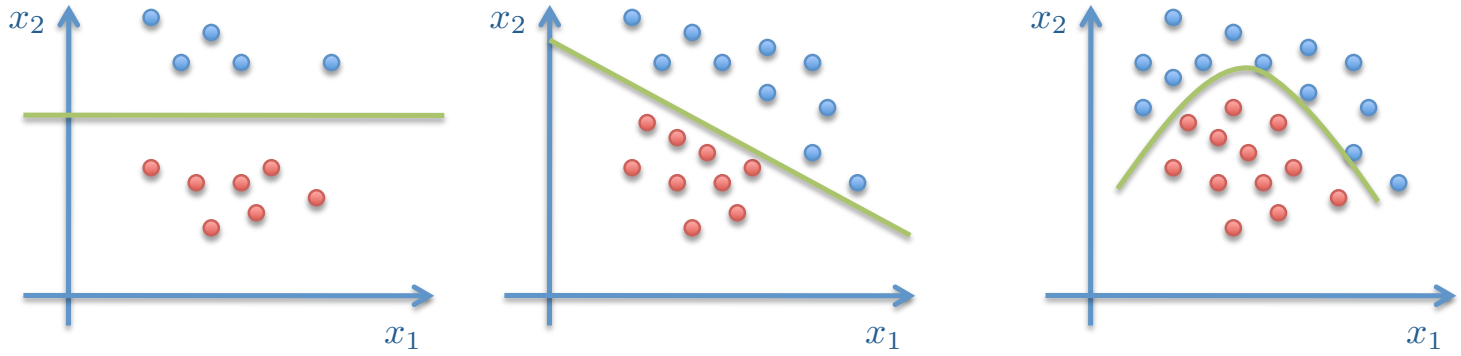
$$CCR = \frac{CC}{N} = 0.78 \qquad ER = \frac{WC}{N} = 1 - CCR = 0.22$$

$$TPR_i = \frac{C_f(i,i)}{\sum_{j=1}^{D} c_f(i,j)} \qquad FPR_i = \frac{\sum_{j=1}^{D} c_f(j,i) - C_f(i,i)}{N - \sum_{j=1}^{D} c_f(i,j)}$$

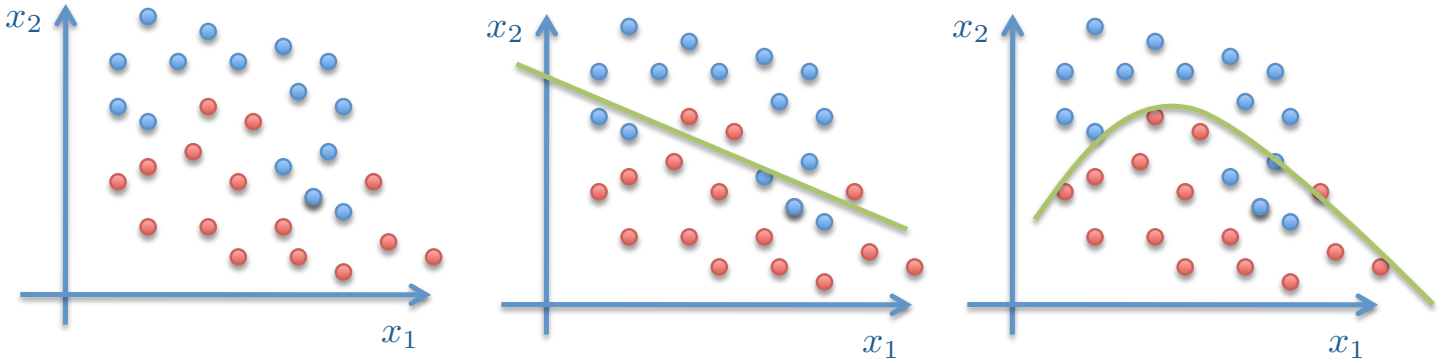
$$TPR_1 = 8/11 = 0.73$$
$$FPR_1 = 2/12 = 0.17$$

Classifier evaluation on training data

- Which data we should use to evaluate the classifier performance?
  - Data may be available at different times

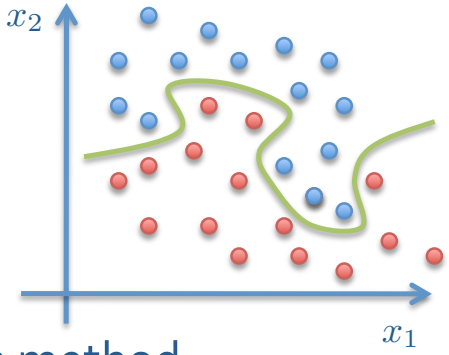


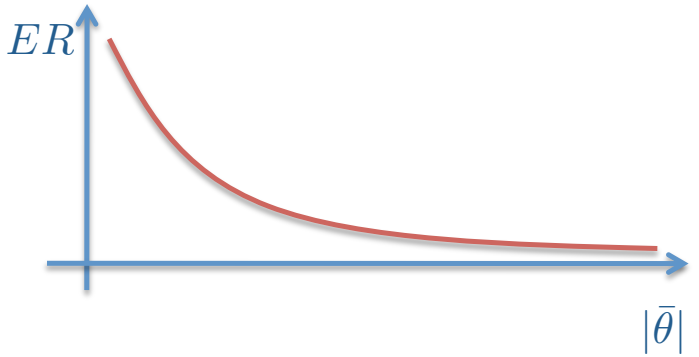
  - Over-fitting problem



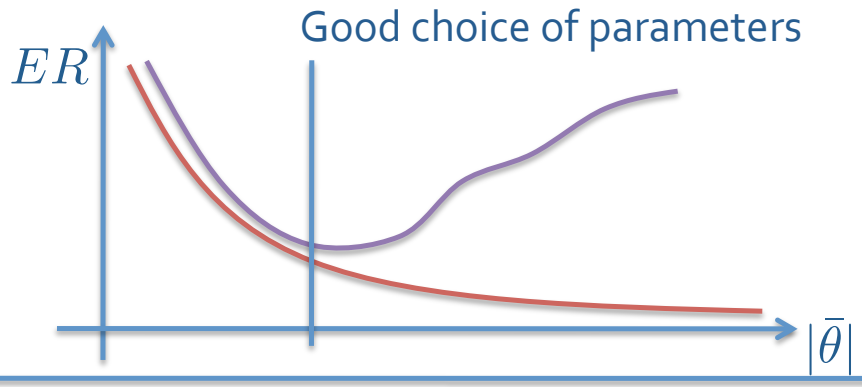
Over fitting and Generalization

- Why not this one?
  - **Training data**: data used to estimate the classifier performance
  - Training error (i.e. Error Rate on training data) decreases with the number of parameters



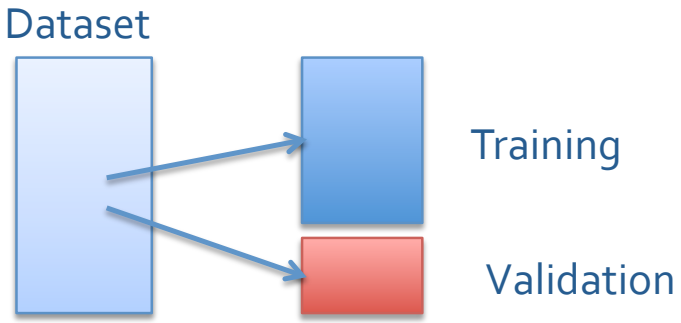


- Use of a validation method
  - **Validation data:** data that are NOT used to estimate the classifier but are used to estimate the performance

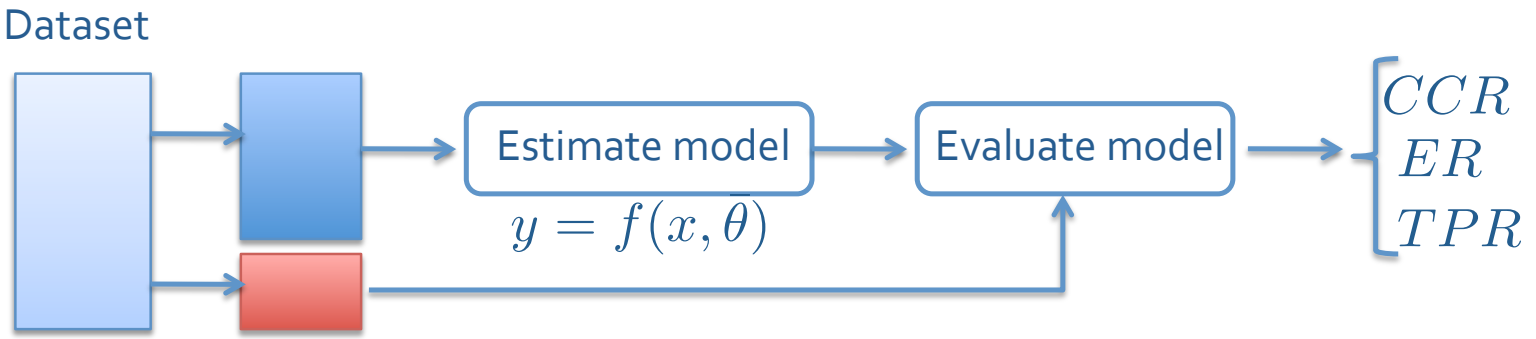


Simple Cross validation

- Randomly split the initial dataset into 2 subset:
  - Training set (usually 2/3 of the total samples)
  - Test set (usually 1/3 of the total samples)



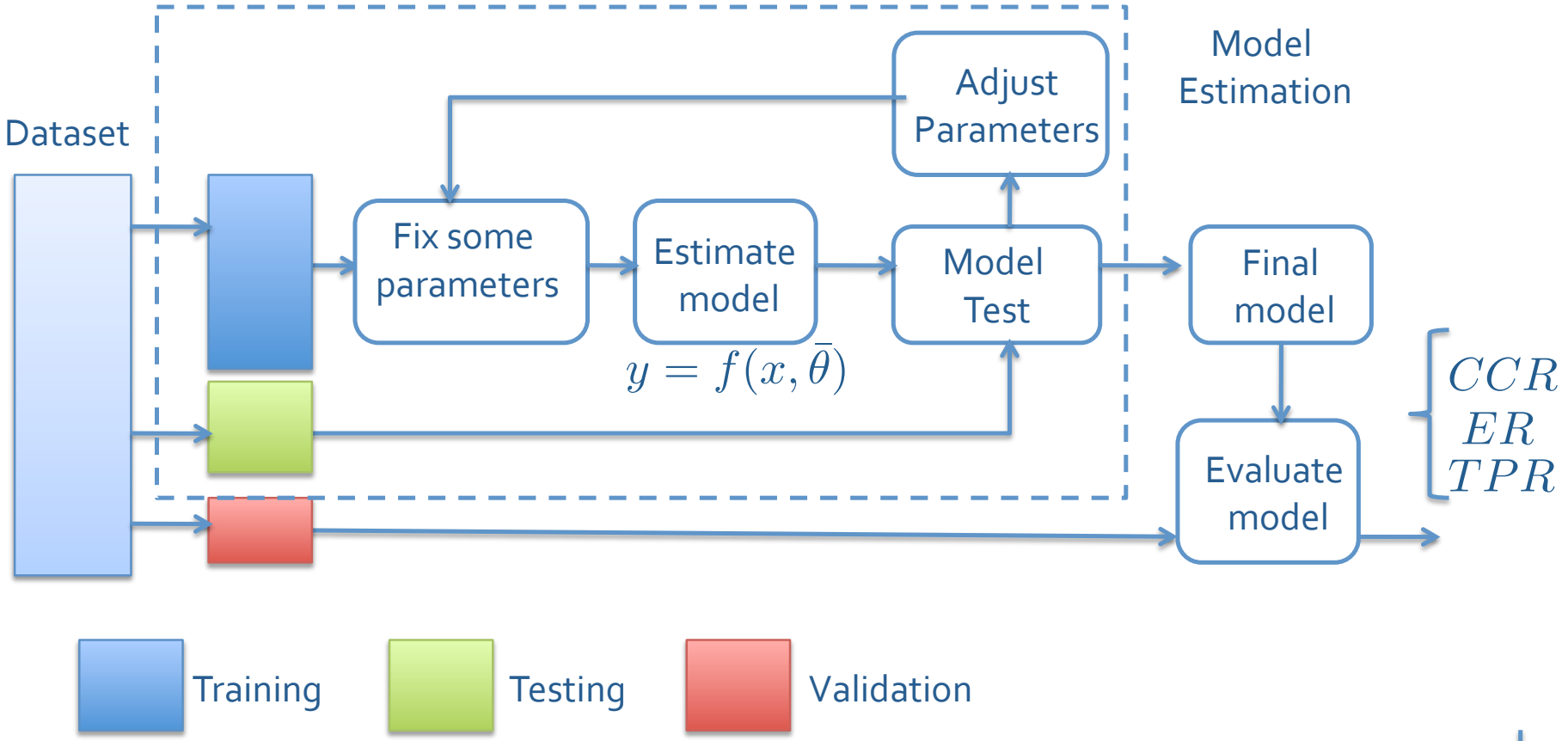
- Build the classifier considering only the training set
- Evaluate the performance of the classifier on validation set



- This approach performs well on large dataset

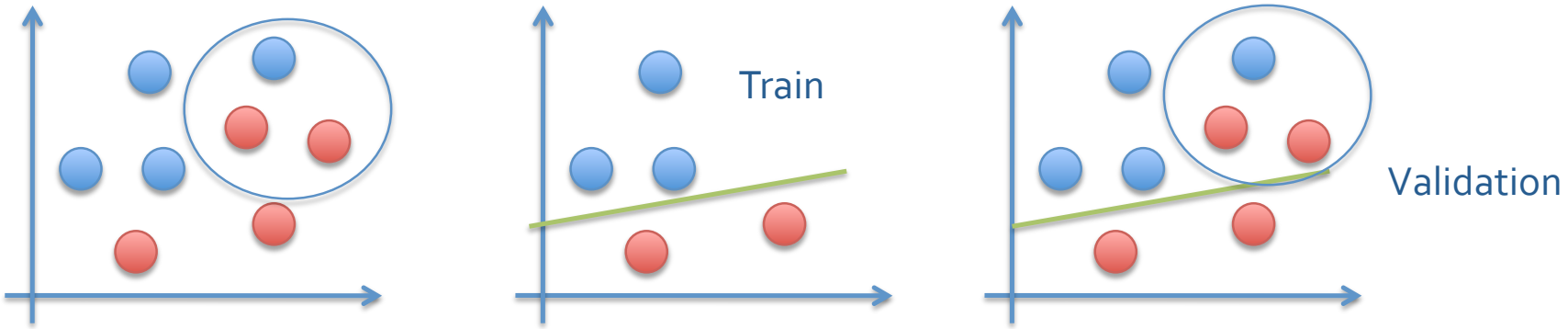
Training Testing and Validation

- Some classifier needs a parameter optimization phase
- Use a extra part of dataset
  - **Testing Data**: Data used during training to adjust parameters of classifier



Repeated Holdout

- Usually large dataset are not available
- Classification is always **biased** because of the choice of the samples

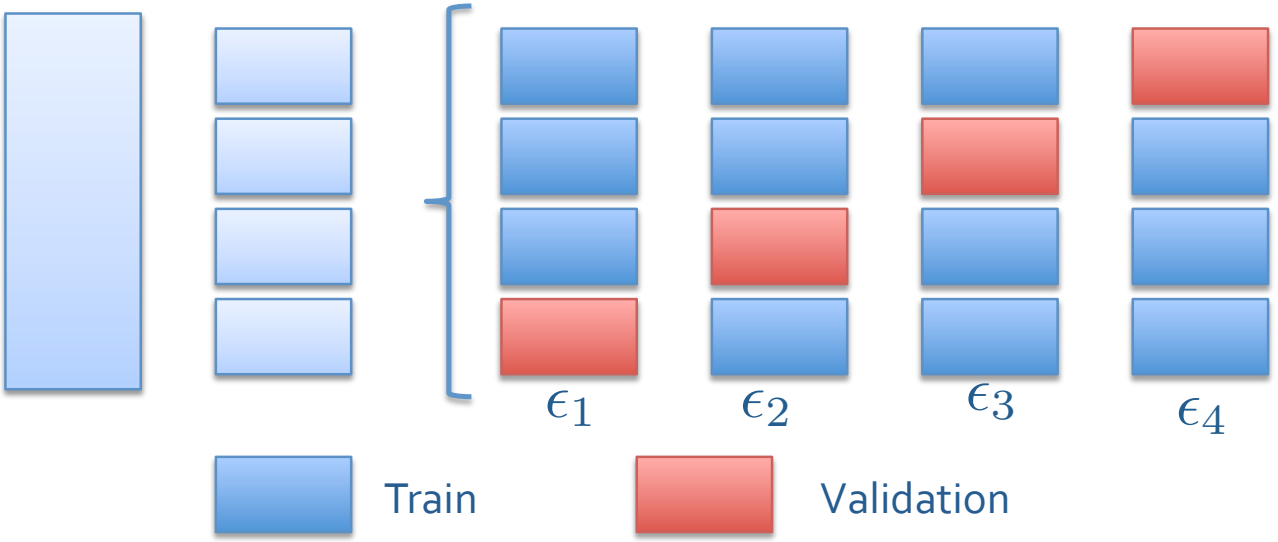


- A solution is given by **repeated holdout method** that repeats the validation process with different sub-samples
  - For i=1..K
    - Extract randomly training (2/3) & validation(1/3) data
    - Estimate model on training
    - Evaluate model on validation ϵ_i
  - Estimate performance

$$E[\epsilon_i] = \mu_i \quad E[(\epsilon_i - \mu_i)^2] = \sigma_i^2$$

K-Fold Cross validation

- Used to avoid overlapping of dataset
- Split the dataset into K folds
  - Use K-1 folds for training and 1 ford for validation

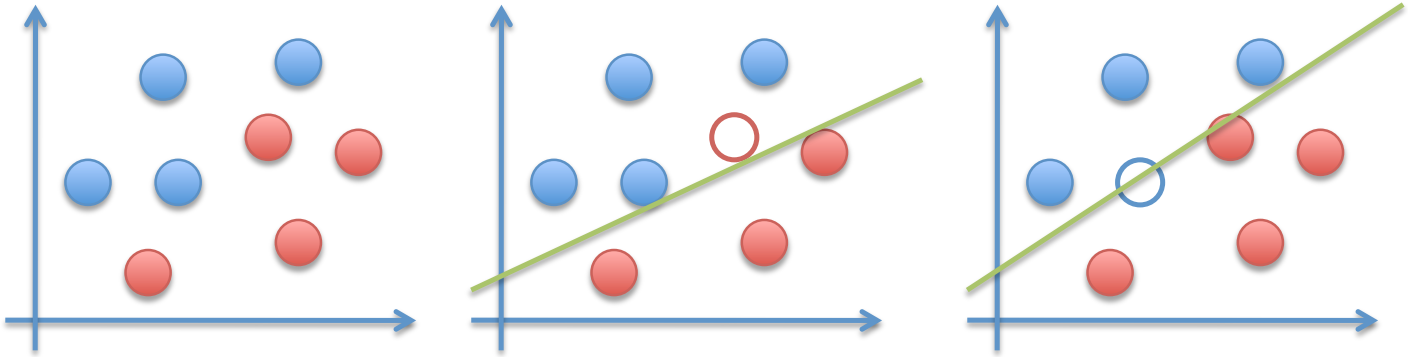


  - Estimate performance

$$E[\epsilon_i] = \mu_i \qquad E[(\epsilon_i - \mu_i)^2] = \sigma_i^2$$

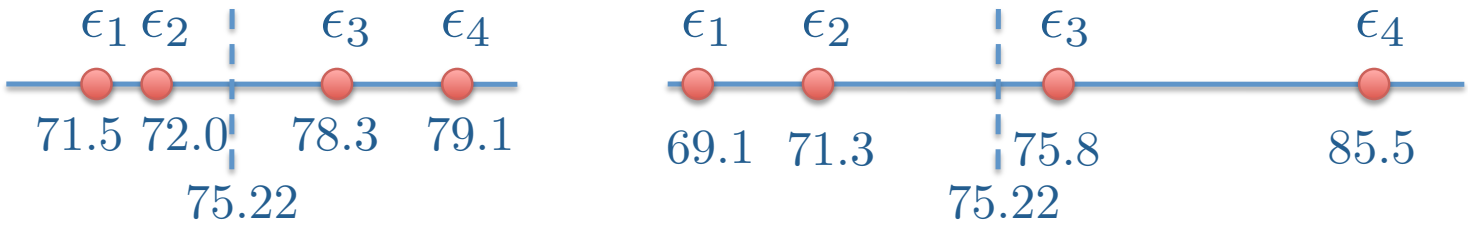
Leave one out cross validation

- K-fold cross validation with K=N
  - Use N-1 sample for training
  - Remove one sample for validation
- More slow, but used **only** with few data
- Reduce the variance of performance (Not good) σ_i^2
  - Removing a sample, the classifier does not change too much
  - Cross validation may not be well performed



Confidence interval

- Average is nothing without variance
- We obtained an estimated average error rate (or accuracy rate)



  - mean depends on the number of folds
  - Having less folds means we are less confident

- How close will the real error go to the estimated error?
- With what probability?
- We can compute a range (confidence interval) of accuracy inside which we can be sure to fall with a certain probability

Confidence interval

- Suppose we want to estimate the true error of our classifier
  - We consider different realization of our evaluation (i.e error on different fold)
  - Hp: errors follow a unknown Normal distribution

$$\epsilon \sim N(\mu, \sigma^2)$$

  - Number of folds n
  - Estimated mean $\bar{X}_\epsilon = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$ estimated variance $\bar{S}_\epsilon^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\epsilon_i - \bar{X}_\epsilon)^2$

- The difference between the true error and the estimated error follows a t-student distribution

$$T = \frac{\bar{X}_\epsilon - \mu}{\sqrt{\frac{\bar{S}_\epsilon^2}{n}}}$$

  - μ is the true error we are looking for

Confidence interval

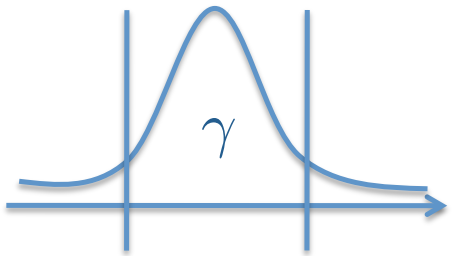
- High probability that the difference between true error and estimated error fall into a give range

$$Pr(-A \leq T \leq A) = \gamma \qquad (-A, A)$$

Confidence interval for T with significance γ

$$Pr(-A \leq \frac{\bar{X}_\epsilon - \mu}{\sqrt{\frac{\bar{S}_\epsilon^2}{n}}} \leq A) = \gamma = 0.95$$

0.95 Probability that T falls into that range



Bilateral case:

$$\gamma = 1 - 2Pr(T \geq A)$$

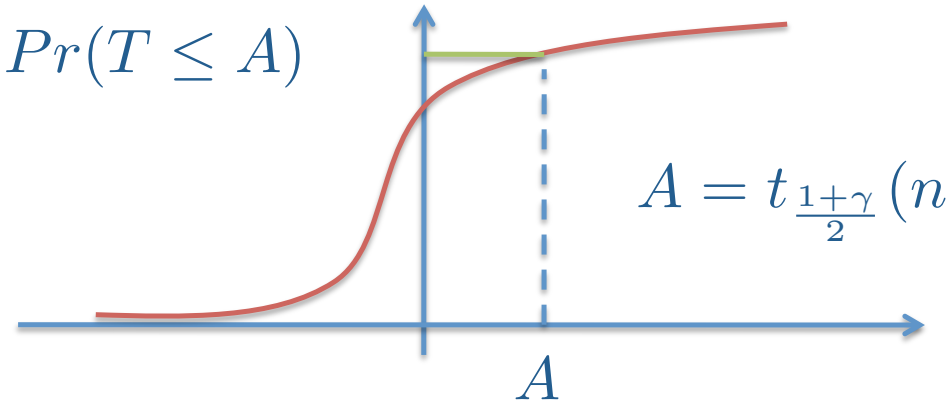
$$Pr(T \geq A) = 1 - Pr(T \leq A)$$

$$\gamma = 1 - 2[1 - Pr(T \leq A)]$$

$$Pr(T \leq A) = \frac{1 + \gamma}{2}$$

Confidence interval (bilateral)

$$Pr(\frac{\bar{X}_\epsilon - \mu}{\sqrt{\frac{\bar{S}_\epsilon^2}{n}}} \leq A) = \frac{1+\gamma}{2} = 0.975$$



Quantile of t-student distribution with n-1 degree of freedom

$$\bar{X}_\epsilon - t_{\frac{1+\gamma}{2}}(n-1)\sqrt{\frac{\bar{S}_\epsilon^2}{n}} \leq \mu \leq \bar{X}_\epsilon + t_{\frac{1+\gamma}{2}}(n-1)\sqrt{\frac{\bar{S}_\epsilon^2}{n}}$$

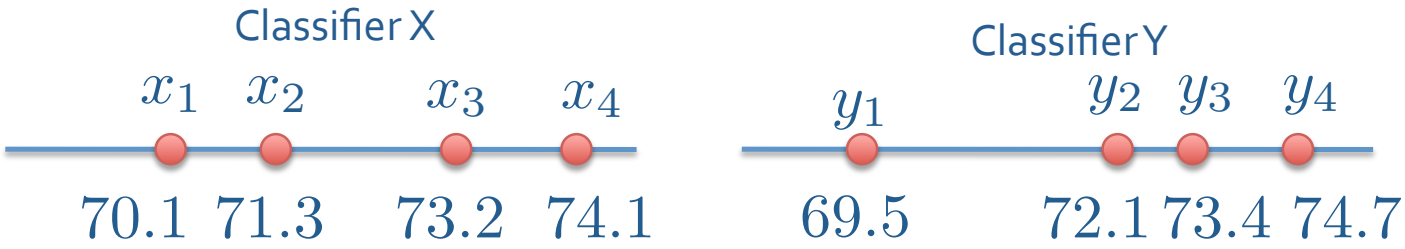
Confidence interval example

- Matlab example…

```
clear all
close all
mu = 75; % Population mean
sigma = 2; % Population standard deviation
n = 20; % Sample size
e = normrnd(mu,sigma,n,1); % Random sample from population


xbar = mean(e) % Sample mean
s = std(e) % Sample standard deviation


gamma = 0.95 % P(-A< T < A) =0.95
t = tinv((1+gamma)/2,n-1) %correspond to  P(T <= A) =(1+gamma)/2

lb = xbar-t*s/sqrt(n)
ub = xbar+t*s/sqrt(n)


figure;
subplot(131),plot(e,'o')
hold on
subplot(131),plot(zeros(1,n)+ub,'r');
subplot(131),plot(zeros(1,n)+xbar,'k');
subplot(131),plot(zeros(1,n)+lb,'g');
```

Classifier comparison

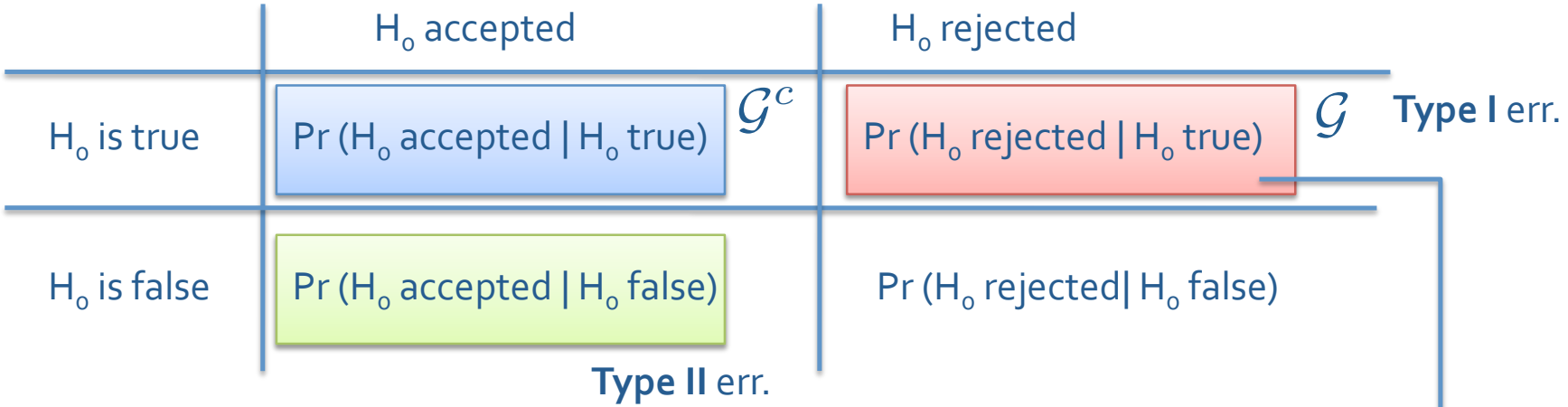
- Suppose we want to establish how much a set of classifier are different in terms of performance



- Solution: Hypothesis test on their mean
  - H_0 non significant difference on mean
  - H_1 mean are different
- A significance test measures how much evidence there is in favor of rejecting the null hypothesis (accepting the alternative hypothesis)
- Student's paired t-test tells us whether the means of two samples are significantly different

Hypothesis Test (from statistics)

- Set of stochastic variables $x_1, \ldots, x_n \quad iid \sim N(\mu, \sigma^2)$
  - Unknown parameters
- Hypothesis on one parameters $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
- Critical region $(x_1, \ldots, x_n) \in \mathcal{G}$
  - If samples belong to critical region H_0 is rejected



- Significance level $\alpha = Pr_{\mu_0}(H_0 \textbf{ rejected} | H_0 \textbf{ true})$
- Confidence interval for μ_0

$$\gamma = 1 - \alpha = Pr_{\mu_0}(H_0 \textbf{ accepted} | H_0 \textbf{ true})$$

Hypothesis Test (from statistics)

- Set of stochastic variables $x_1, \ldots, x_n \quad iid \sim N(\mu, \sigma^2)$
- Hypothesis on mean $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
- **Variance is unknown**
- Critical region

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{(n)}} \geq t_{1-\frac{\alpha}{2}}(n-1)$$

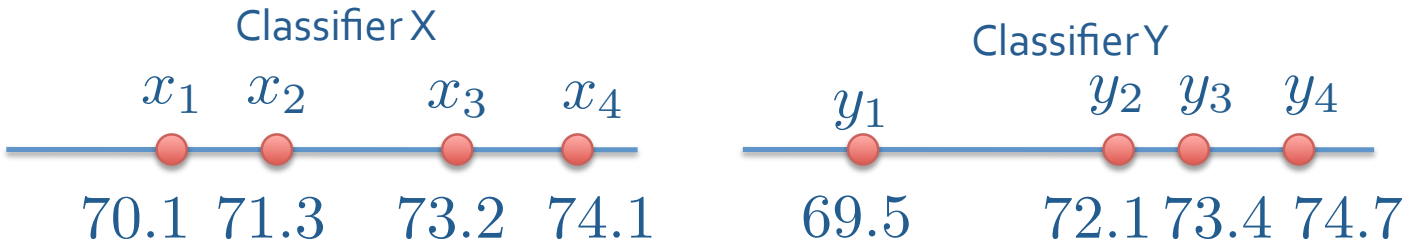
- Confidence interval (two tailed)

$$\left[\bar{x} - t_{\frac{1+\gamma}{2}}(n-1)\frac{s}{\sqrt{(n)}}, \bar{x} + t_{\frac{1+\gamma}{2}}(n-1)\frac{s}{\sqrt{(n)}}\right]$$

$$\gamma = 1 - \alpha$$

Student's Paired t-test

- Errors follows a gaussian distribution with same variance



$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

$$\bar{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 \qquad \bar{S}_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{Y})^2$$

- Paired test couples samples x_i, y_i
- Statistic with t-student distribution

$$n = m \qquad T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\bar{S}_x^2 + \bar{S}_y^2}{n}}}$$

Student's Paired t-test

- Hypothesis test with $H_0 : \mu = 0, H_1 : \mu \neq 0$
- From previous results..
- Critical region

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{S_x^2 + S_y^2}{n}}} \geq t_{1-\frac{\alpha}{2}}(n-1)$$

  - If the abs difference of mean fall outside the interval, the null hypothesis can be rejected
- Confidence interval

$$\left[(\bar{X} - \bar{Y}) - t_{\frac{1+\gamma}{2}}(n-1)\sqrt{\frac{S_x^2 + S_y^2}{n}}, (\bar{X} - \bar{Y}) + t_{\frac{1+\gamma}{2}}(n-1)\sqrt{\frac{S_x^2 + S_y^2}{n}}\right]$$

Student's Paired t-test

- Matlab example

```
close all
clear all
mu = 77; % Population mean
sigma = 2; % Population standard deviation
n = 5; % Sample size
x = normrnd(mu,sigma,n,1); % Random sample from population

mu = 75.6; % Population mean
sigma = 2; % Population standard deviation
y = normrnd(mu,sigma,n,1);

xbar = mean(x);
ybar = mean(y);
sx = var(x);
sy = var(y);

%significance of test
alpha = 0.05;

t = tinv(1-alpha/2,n-1);
s = sqrt((sx+sy)/n);

fprintf('t=%.3f\n', t);
fprintf('s=%.3f\n', s);
fprintf('xbar-ybar = %f\n', xbar-ybar);
lb = -t*sqrt(s/n);
ub = t*sqrt(s/n);

fprintf('IC = (%.3f,%.3f)\n', lb,ub);


if (abs(xbar-ybar)/sqrt(s/n)>= t)
    fprintf('H_0 rejected\n')
else
    fprintf('H_0 accepted\n')
end
```