Introduction to Validation and Theoretical Issues [IIA – Lect.______]

Alessio Micheli

micheli@di.unipi.it

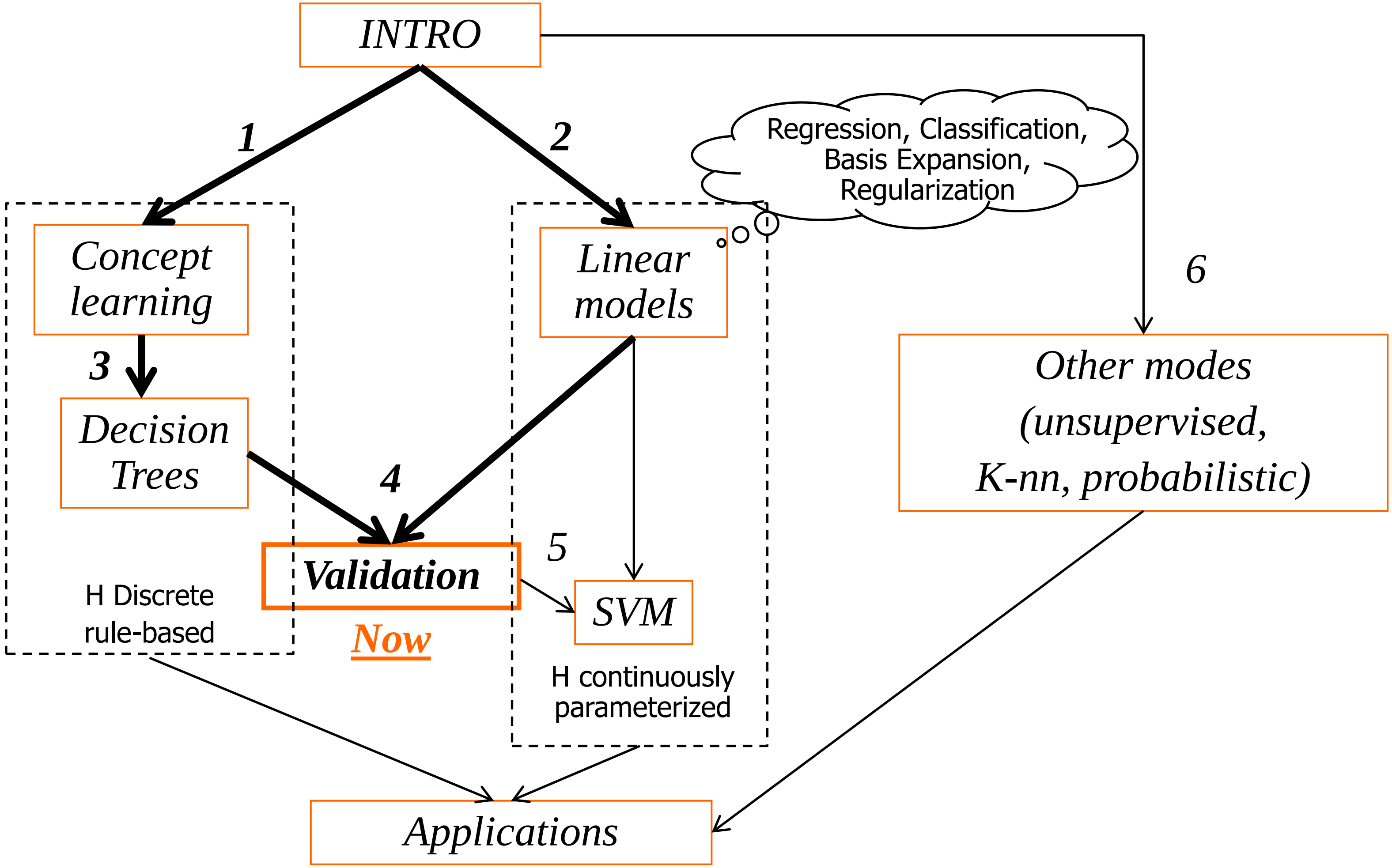
Dipartimento di Informatica
Università di Pisa - Italy

Computational Intelligence & Machine Learning Group

DRAFT, please do not circulate! *2023*

In the course flow

Dip. Informatica
University of Pisa



Obiettivi (overview)

Dip. Informatica
University of Pisa

Questioni fondamentali del ML:

evaluate generalization capabilities (of your hp)

- ruolo <u>essenziale</u> della validazione
- cenni (dell'esistenza) di fondamenti teorici (supporto al significato del ML)

- Aspetto sia teorico che pratico per un uso *consapevole* del ML
- Raccogliamo gli spunti raccolti finora solo indirettamente dedicandoci una (questa) lezione

ML issues

Quando un modello di ML
è un buon modello?
Usare il ML versus usare bene il ML

Machine Learning: generalization (I)

Dip. Informatica
University of Pisa

- *Learning*: search for a *good function* in a function space from known data

Def

- Good w.r.t. generalization error: it measures how accurately the model predicts over novel samples of data (low error, high accuracy and vice versa)

[Repetita from lect. 1]

Generalization (II)

Dip. Informatica
University of Pisa

- Inductive learning hypothesis
  - Any *h* that approximates *f* well on training examples will also approximate *f* well on new (unseen) instances x (**?**)
  - I.e. is it really valid? And at which extent?

- Punto centrale, ma come obiettivo del ML:
  - Teoria che supporta in che condizioni ciò si verifca
  - Guida la scelta del "best model" (tra modelli diversi o configurazioni diverse: iperparametri, livello di training, ...)
  - Va verificato nelle applicazioni

Generalization (III)

Dip. Informatica
University of Pisa

[Repetita from lect. 1]

- Generalization: crucial point of ML!!!
- **Learning** phase **(training, fitting)**: build the model from know data – *training data* (and bias)
- **Predictive** phase **(test)**: apply to new examples (we take the inputs ***x'*** and we compute the response by the model; we compare with its target *d'* that the model has never seen): evaluation of the predictive hypothesis, i.e. of the generalization capability

Note: *performance* in ML = *predictive accuracy*

estimated by the error computed on the (Hold out) Test Set

Def

- [*repetita*] **Overfitting**: A learner overfits the data if it outputs a hypothesis h(·)∈H having true error ε and empirical (TR) error E, but there is another h'(·)∈H having E'>E and $\varepsilon' < \varepsilon$

Premise: which measure?

Dip. Informatica
University of Pisa

Recap:

To evaluate typically we measure (see def. in previous lectures)

- For *classification*: MSE for the loss, accuracy or mean error rate for the outcome
  - but also precision, recall or specificity, sensitivity (accounting for False Positive, False Negative), …
- For *regression*: MSE, Root MSE (S), Mean Absolute Error, Max Absolute Error, ….
  - but also statistics measures such R (correlation coefficient/index), etc.

- Of course high error ←→ low accuracy (both for training, test, etc.)
  - E.g. poor fitting with high training error,
  - E.g. poor generalization with high test error, …

Validation: Two aims

Very important!

Dip. Informatica
University of Pisa

Def

- ***Model selection:*** estimating the performance (*generalization error*) of different learning models in order to choose the best one (to generalize).
  - this includes search the best *hyper-parameters* of your model (e.g. polynomial order, lambda of ridge regression, …).

  It returns a model

Def

- ***Model assessment:*** having chosen a final model, estimating/evaluating its prediction error/ risk (*generalization error*) on new *test* data (measure of the quality/performance of the ultimately chosen model).

  It return an estimation

Gold rule: Keep separation between goals and use separate data sets

Defs

Hold out

Dip. Informatica
University of Pisa

- If data set size is sufficient: e.g. 50% TR, 25% VL, 25% TS **disjoint sets**

Training	Validation	Test
D_{t+v} *Development/design set*		$D \backslash D_{t+v}$

- TR: *Training set* is used to fit [**training**]
- VL*: Validation set* (or *selection set*) is used to select the best models (among different models and/or hyper-parameters configurations) [**model selection**]
- TR+VL sometimes are joitnly called development/design set , i.e. used to build the final model
- TS*:Test set* is used for estimation of generalization error (of the final model) [**model assessment**]

Notes:

1) the estimation made for model selection (on vl set) [is for model selection purpose], it is not a good estimation for the assessment phase/risk test.

2) **Test set results** cannot be used for model selection (or call it validation set)

Dip. Informatica
University of Pisa

Test or model selection?

- What if test set is used in a (repeated) design cycle?
  - We are making *model selection* and not reliable *assessment* (estimation of expected generalization error)
    - and *we wouldn't be able to do that on future examples*
    - Blind test set concept (e.g. for ML competitions)
    - Image an exam exercise: if you see the solutions it is not a test!
  - In that case, used test set error provides an overoptimistic evaluation of the true test error (→ we will see how easy is to obtain very high classification accuracy over a random task even using the test set only implicitly)

Gold rule:

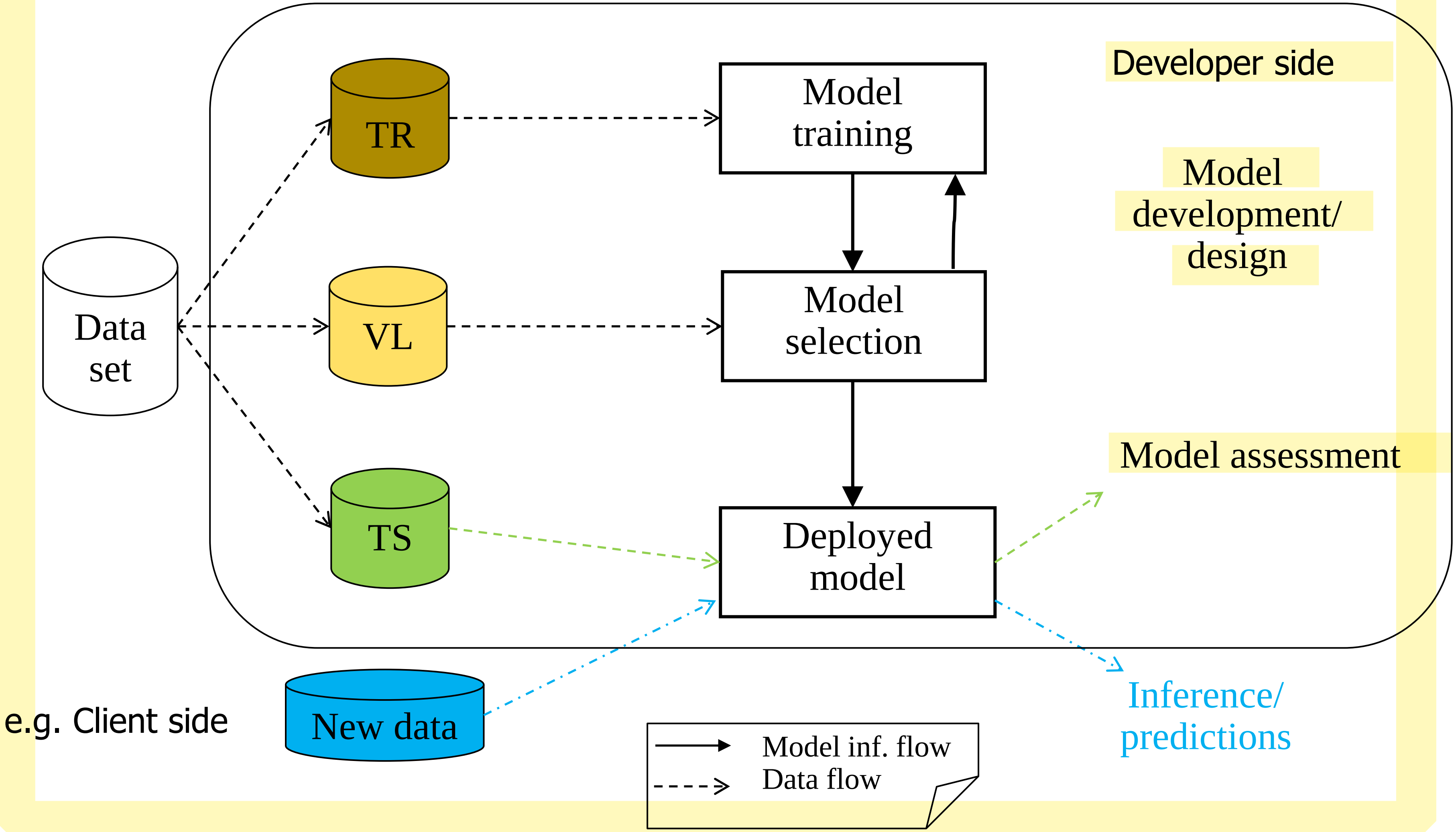
Keep separation between goals and use separate sets

(**TR** for training, **VL** for model selection, **TS** for risk estimation)

TR=training set, VL= validation set, TS=test set

TR/VL/TS by a simple schema

Dip. Informatica
University of Pisa



A simple meta-algorithm

Dip. Informatica
University of Pisa

- Separate TR (training), VL (validation) and TS (test) sets
- Search best $h_{w,\lambda}()$ changing the model hyper-parameters λ [e.g. the polynomial order, the lambda for ridge regression]:
  - For each different values of λ (grid search)
    - Search best $h_{w,\lambda}()$ that minimize error/empirical loss (fitting the TR set) finding the best w parameters,

      where best = minimum error on TR set [e.g. $argmin_{\mathbf{w}}\ Loss(\mathbf{w})\ in\ L_2$]
- Select the best best $h_{w,\lambda}()$: where best = minimum error on the VL set
- (Optional: Now it is also possible to fit $h_{w,\lambda}(x)$ on TR+VL with best λ model)
- Evaluate the final $h_{w,\lambda}(x)$ on the TS

This is a double cycle: Search best can be a *for* on a grid of values in the external cycle: for each λ value you train a model $h_{w,\lambda}$ (in the internal cycle, e.g. the gradient descent cycle) and then compute the results (accuracy) on the VL set.
Then take the best value of λ i.e the model with min VL err or max VL accuracy etc..

Search on a grid
(e.g. with 2 hyper-parameters)

Dip. Informatica
University of Pisa

Def

- Find *hyper-parameters* value (i.e. parameters that are not directly learnt, which are not modified by training)
- Search best hyper-parameter can be a <<FOR>> over a grid of candidate values. For each trained model $h_{\mathbf{w},\lambda}$ compute the results (accuracy) on the VL set. Then take the one with the minimum error or the max accuracy.

Hyper-param.	Lambda 0.1	Lambda 0.01	Lambda 0.001
Degree 1	Res1	Res4	Res7
Degree 2	Res2	Res5	Res8
Degree 4	**Res3**	Res6	Res9

E.g. "Res1" is computed on the VL set, by the model with and Polynomial-Degree=1 and Lambda=0.1 trained on the TR set

- Example: The best one is *Res3* → (Degree 4, lambda=0.1) is the winner
- We can automatize it!!! Parallelization is easy (independence of the trials)
- Alternatives exist to reduce the cost or to automatize the search *

Exercise

λ	TR	VL	TS
0.5	75	70	70
0.1	80	75	70
0.01	90	70	72

Accuracy (% di classificazione corretta)

- In che ordine si usano le porzioni di dati per calcolare i valori in tabella?
- Quale modello (ossia lambda) si sceglie?
- Che fenomeni si osservano?

Controesempio (separare TR, VL e TS)

Dip. Informatica
University of Pisa

- 20-30 esempi, 1000 variabili di input random,
- *random target* 0/1
- Scelgo 1 modello con una sola variabile/feature che indovina *'per caso'* al 99% sul dataset e poi su qualsiasi split successivo in training, validation e test set.

Perfect result (a model with accuracy 99%)? What is wrong?

99% non è una buona stima dell' errore di test (quella corretta e' 50%)

1. Errore stimato su training o validation per model selection NON è utile per stima del rischio! Dati di TR o VL non vanno usati per scopi di test

2. Usare *tutto* il data set per feature/model selection lede la correttezza dell stima (risultati biased – «Feature Selection bias»).

- Test set è stato usato implicitamente all'inizio*.
- Test deve essere separato prima, prima di qualsiasi model selection o design del modello (incluso selezione di features)

Un test set esterno fornisce invece la stima corretta del 50% (*random coin result* !).

E' la correttezza della stima che è in giudizio, non la possibilità di risolvere il task!

Delicato confrontando metodi diversi e usando tecniche di K-fold cross-validation che in se non garantiscono correttezza della procedura di validazione

The table for the Counterexample

Dip. Informatica
University of Pisa

Input variable value

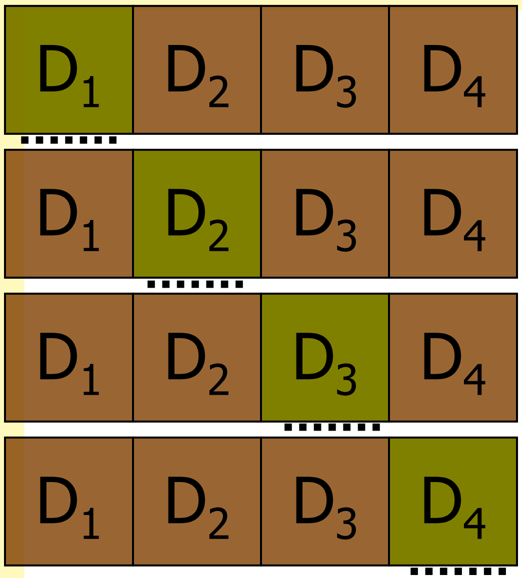
Pattern

	1	2	…	26	27	28	…	1000	**Target**
1	…	…	…	1	1	1	…	…	1
2				0	0	0			0
3				1	1	1			1
4				0	0	0			0
…				0	0	0			0
…				1	1	1			1
…				0	0	0			0
20	…	…	…	1	1	1	…	…	1
TS1				1	0	0			1
TS2				0	1	0			0
TS2				1	1	1			1
Accuracy				100%	33%	66%			

Hold out and **K-fold** cross validation

Dip. Informatica
University of Pisa

Hold out CV *can make insufficient use of data*



Fold 4

Def K-fold Cross-Validation

- Split the data set D into k mutually exclusive subsets $D_1, D_2, \ldots, D_K$
- Train the learning algorithm on $D \backslash D_i$ and test it on D_i
- Summarize averaging all the D_i results (*diagonal*)
- NOTE: This technique can be used both for the validation set or for the test set
- *It uses all the data for training and validation or testing*

Issues:

- How many folds? 3-fold, 5-fold , 10-fold, …., 1-leave-out
- Often computationally very expensive
- Combinable with validation set, double-K-fold CV, ….

An example of model selection and assessment (with K-fold CV)

Dip. Informatica
University of Pisa

- Split data in TR and Test set (here simple hold-out or a K-fold CV)

- [Model selection] Use K-fold CV (internal) over TR set, obtaining new TR e VL set in each split, to find best hyper-parameters of your model (e.g. polynomial order, lambda of ridge regression, ...): How? Apply a **grid-search** with many possible values of the hyper-par.
  – i.e. for example a k-fold-CV for λ =0.1, a k-fold-CV CV for λ =0.01 , ... and then take the best λ (comparing the mean errors computed over the validation sets obtained by the all the folds of each k-fold CV, ... the results on the diagonal in the previous slide)

- Train on the whole TR set the final model

- [Model assessment] Evaluate it on the external Test set

Validation: summarizing

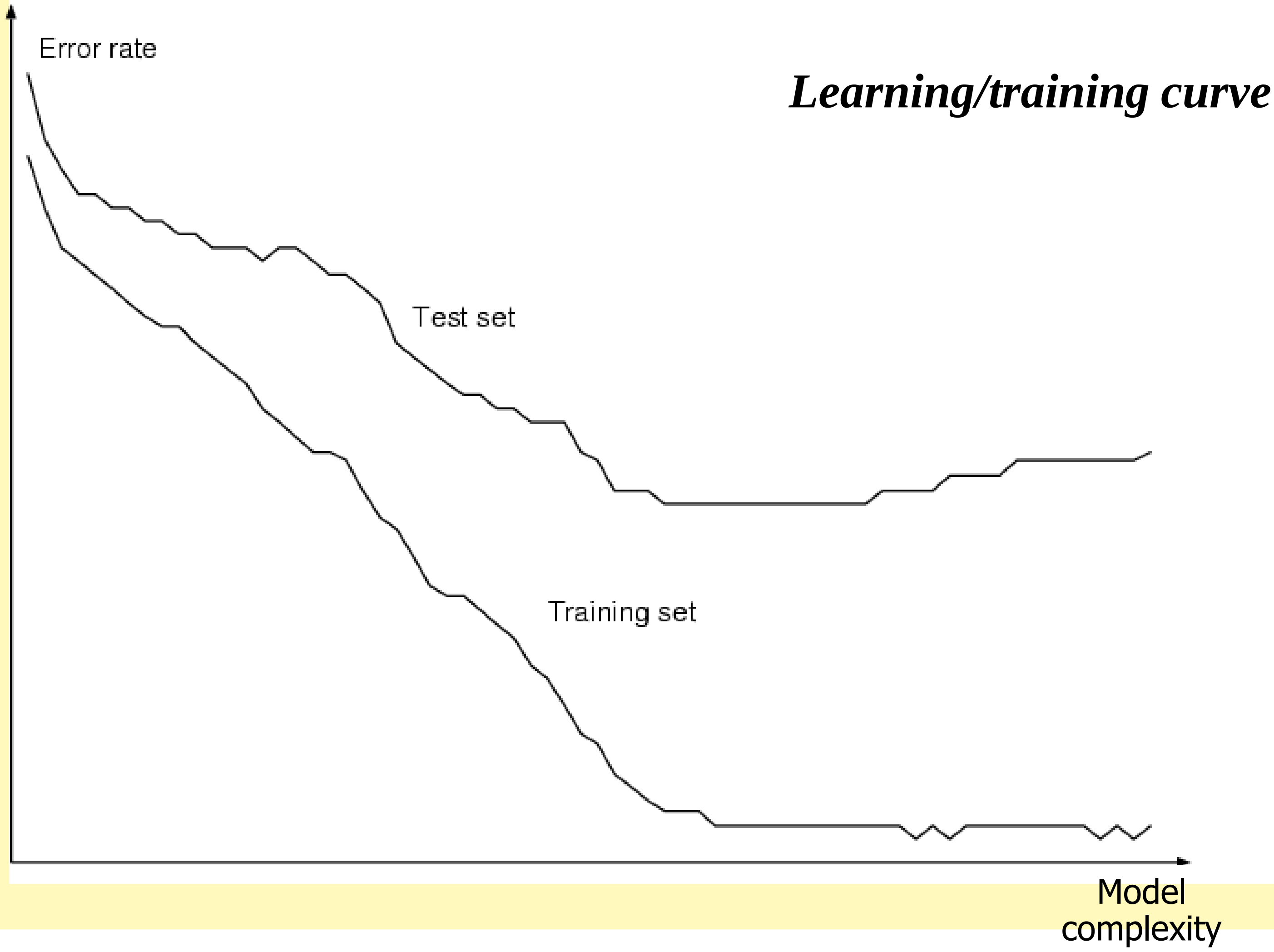
Dip. Informatica
University of Pisa

- **Stima Empirica**: errore calcolato/stimato su (Hold out)
  - E.g. Hold out: training, validation set and test sets
  - K-fold cross validation (resampling)

- **Teoria**: E.g. Statistical Learning Theory [Vapnik] :
  - *sotto quali condizioni (matematiche) un modello è capace di generalizzare (buona generalizzazione)?* → cenni

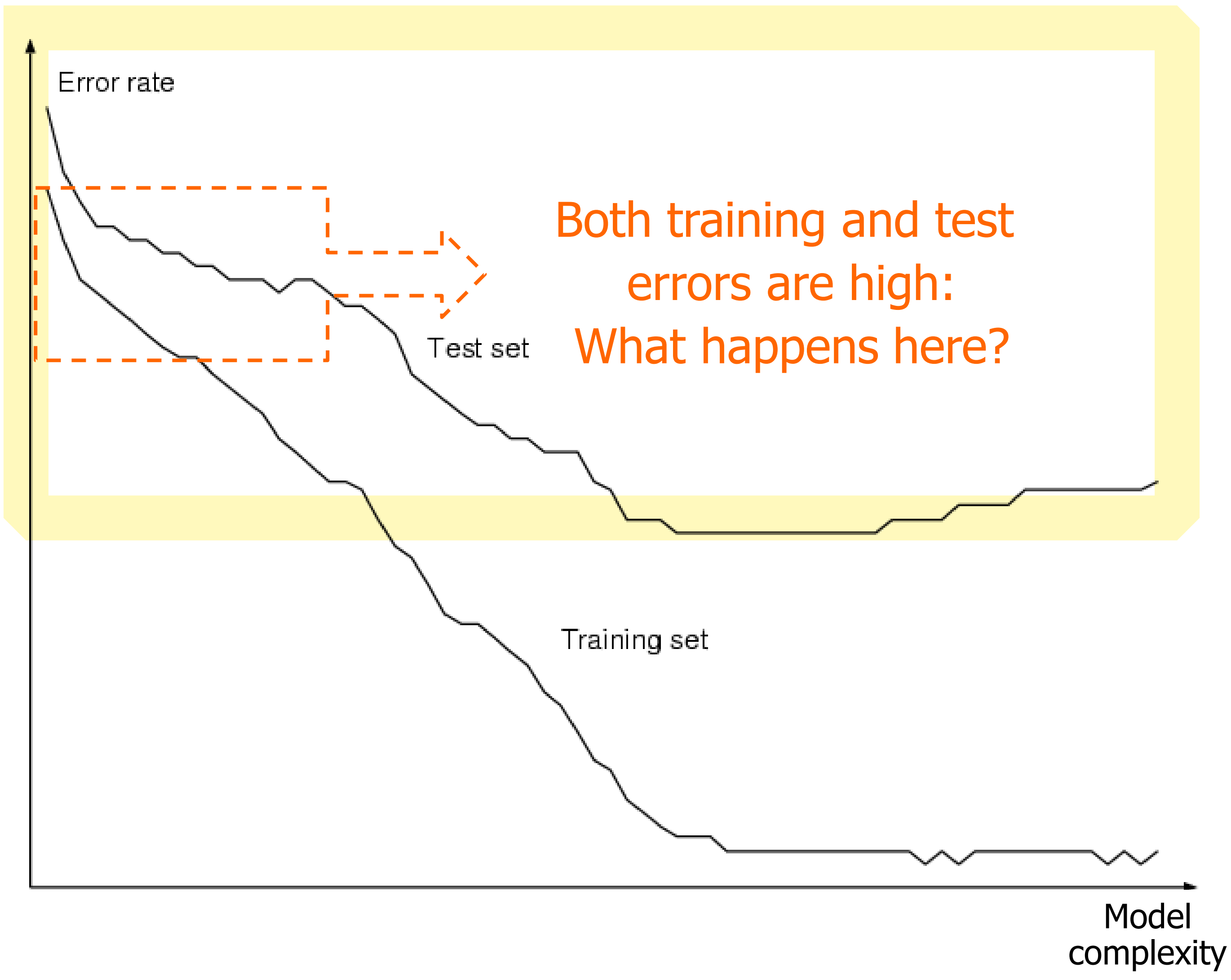
Typical behavior of learning

Dip. Informatica
University of Pisa



Typical behavior of learning

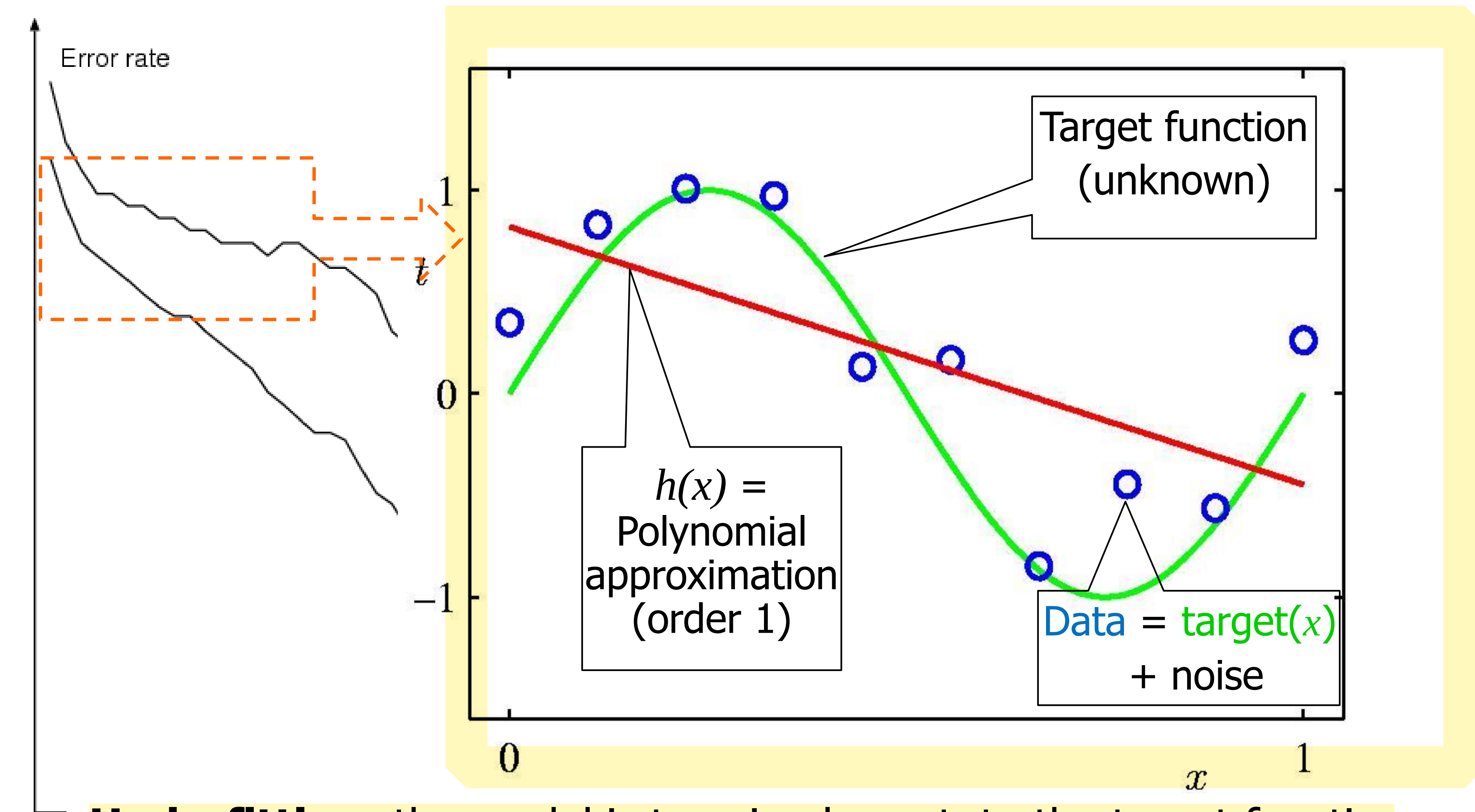
Dip. Informatica
University of Pisa



Both training and test errors are high:
What happens here?

Typical behavior of learning & Polynomial Curve Fitting (I)

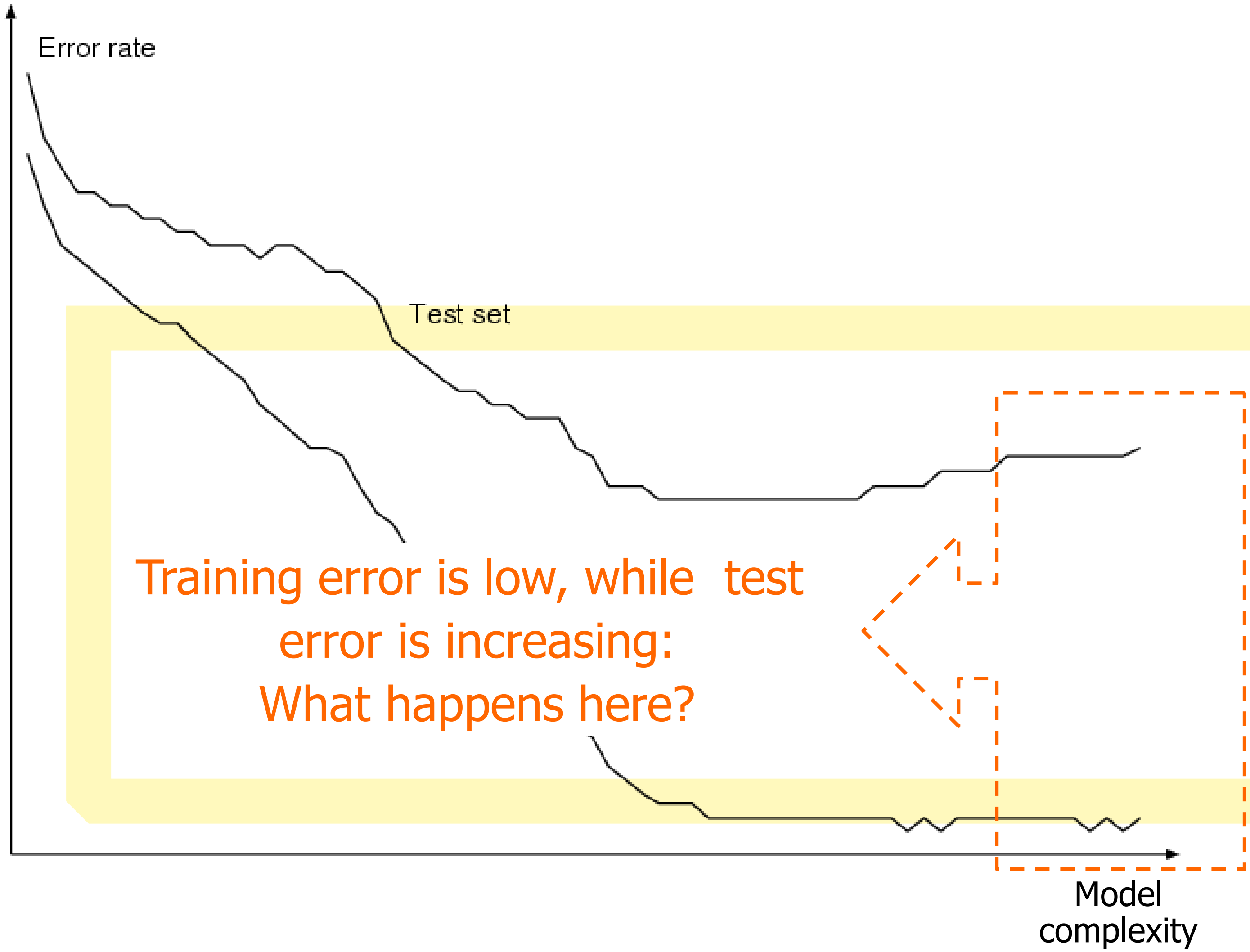
Dip. Informatica
University of Pisa



Underfitting: the model is too simple w.r.t. to the target function both for known data and new data

Typical behavior of learning

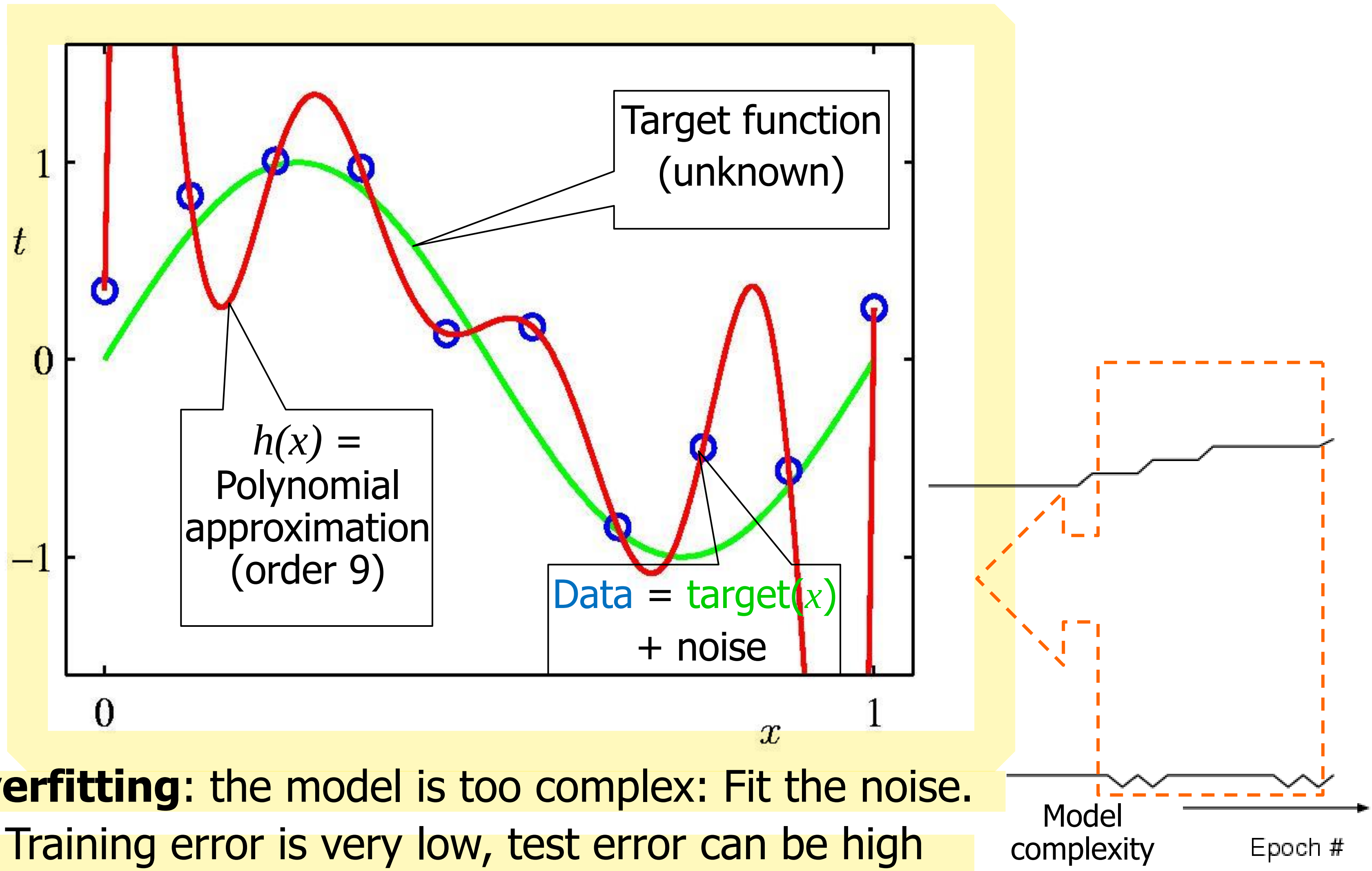
Dip. Informatica
University of Pisa



Training error is low, while test error is increasing:
What happens here?

Typical behavior of learning & Polynomial Curve Fitting (II)

Dip. Informatica
University of Pisa



Overfitting: the model is too complex: Fit the noise.

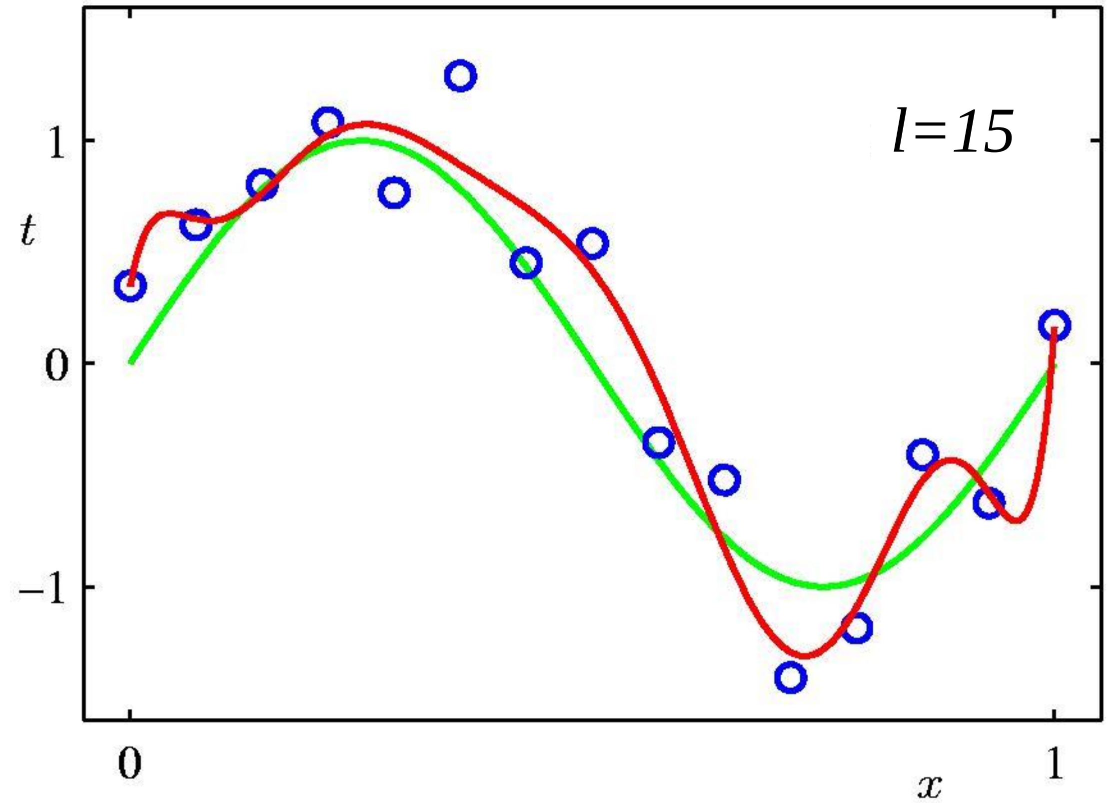
Training error is very low, test error can be high

Data Set Size: $l=15$

previous was 10

Dip. Informatica
University of Pisa

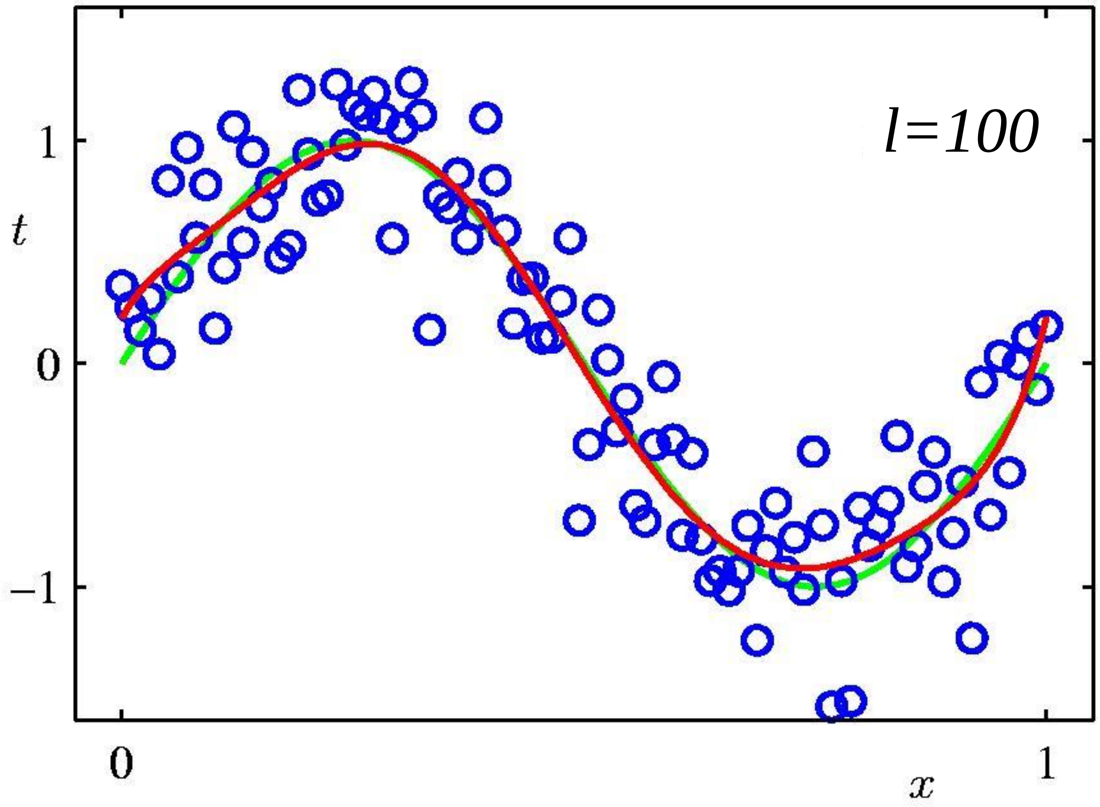
9th Order Polynomial (chancing the of number of data)



Data Set Size: *l=100*

Dip. Informatica
University of Pisa

9th Order Polynomial (even more data)



Toward SLT

Dip. Informatica
University of Pisa

Putting all together:

- The *generalization* capability (measured as a risk or test error) of a model
  - with respect to the training error
  - overfitting and underfitting zones

1. The role of model complexity
2. The role of the number of data

- Statistical Learning Theory (SLT): a general theory relating such topics

(Simplified) Formal Setting Statistical Learning Theory (SLT)

Dip. Informatica
University of Pisa

Defs

- Approximate unknown $f(\boldsymbol{x})$, d *(or y or t) is the target (d=true f +noise)*
- Minimize *risk function* $R = \int L(d, h(\boldsymbol{x}))dP(\boldsymbol{x}, d)$ True Error Over *all* the data
- Given
  - value from teacher (d) and the probability distribution $P(\boldsymbol{x}, d)$
  - a loss (or cost) function, e.g. $L(h(\boldsymbol{x}), d) = (d - h(\boldsymbol{x}))^2$
- Search h in H : Min R
- But we have only the finite data set $TR = (\boldsymbol{x}_{p,} d_p), \quad p = 1 \ldots l$
- To search h: minimize *empirical risk* (training error E), finding the best values for the model free parameters

$$R_{emp} = \frac{1}{l}\sum_{p=1}^{l}(d_p - h(\boldsymbol{x}_p))^2$$

- Emprical Risk Minimixzation (ERM) Inductive Principle
- *Can we use R_{emp} to approximate R?*

Vapnik-Chervonenkis-dim and SLT: a general theory (I)

Dip. Informatica
University of Pisa

Def

- Given the *VC-dim* (VC), a measure *complexity* of *H (flexibility to fit data)* (e.g. Num. of parameters for linear models/polynomials)

  Repetita: Can we use R_{emp} to approximate R?

Very important!

Def

- *VC-bounds in the form:* it holds with probability 1-δ that

  guaranteed risk

$$\boxed{R \leq R_{emp} + \varepsilon(1/l, VC, 1/\delta)}$$

VC-confidence

- First (basic) explanation:
  - ε is a function that grows with VC *(VC-dim)*, that decreases with (higher) l and $delta$.
  - We know that R_{emp} decrease using complex models (with high *VC-dim*) (e.g. the polynomial degree in the example)
  - $delta$ is the confidence, it rules the probability that the bound holds (e.g. low delta 0.01, it holds with probability 0.99)
- Now we can see how it can "explain" the *underfitting* and *overfitting* and the aspects that control them.

Vapnik-Chervonenkis-dim and SLT: a general theory (II)

Dip. Informatica
University of Pisa

Comments:

Very important!

- *VC-bounds in the form:* it holds with probability 1-δ that

Def

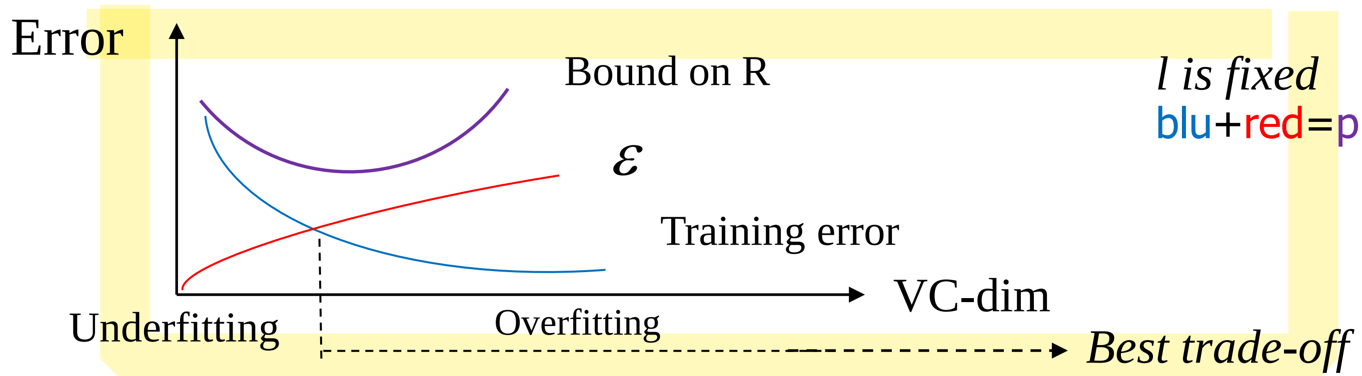
guaranteed risk $$R \boxed{\leq} \underline{R_{emp}} + \underline{\varepsilon}(1/l, VC, 1/\delta)$$ VC-confidence

Intuition:

- Higher l (data) → lower VC confidence and bound close to R
- Too simple model (low VC-dim) can be not suff. due to high R_{emp} (underfitting)
- Higher VC-dim (fix l) → lower R_{emp} but $VC\text{-}conf$. and hence R may increase (overfitting)

Def

- *Structural risk minimization*: minimize the bound !



l is fixed
blu+red=purple

- Concept of control of the model complexity (flexibility):
  trade-off between TR accuracy (fitting) and model complexity (VC-dim)

Discussion
Complexity control

Dip. Informatica
University of Pisa

- **Statistical Learning Theory (SLT):**
  - Permette inquadramento formale del problema della generalizzazione e (underfitting/)overfitting, fornendono limitazioni superiori analitiche e quantitative al rischio R di predizione su tutti i dati, indipendemente dal tipo di learning algorithm o dettagli del modello.
  - Il ML è ben fondato:
    - Il rischio del learning (e l'errore di generalizzazione) può essere analiticamente limitato, e solo pochi concetti sono fondamentali !
    - Si può trovare un buona approssimazione dell f target da esempi, pur di avere un buon numero di dati e una adeguata complessità del modello (misurabile formalmente con la VC-dim)
  - Porta a nuovi modelli (SVM) (e altri metodi che direttamente considerino il controllo della complessità nella costruzione del modello)
  - Fonda uno dei principi induttivi sul *controllo della complessità*

Domande aperte:

- Quali (altri) principi vi sono per fondare il controllo della complessità e <u>come operare in pratica</u>?
  - Come misurare la complessità (flessibilità per il fitting)?
  - Come trovare il bilanciamento ottimo tra fitting e complessità ?

Some Examples for Complexity Control

Dip. Informatica
University of Pisa

- Linear models (LM):
  - Complexity seems* related to number of free parameters *w*: input dimension / dim. of the basis expansion (e.g. polynomial degree)
  - Lambda parameter for the regularized version (using the model selection/validation techniques to find the proper value of lambda)

- Decision trees (DT): number of nodes (e.g. control by early stop, pruning)

- We will also see: direct approach to the complexity optimization through the SVM model

- <u>**Exercise**</u>: relate the complexity control to the approaches used in the different models, explaining the *underftting* and the *overfitting* from the point of view of the SLT upper bound on *R*: e.g. how to explain the role of the hyper-parameters lambda in Linear models or # of nodes in DT etc. in terms of SLT?

Conclusioni

Dip. Informatica
University of Pisa

- ML models flexiblity →
  - Use the power of ML without control is a way to produce *illusory results*
  - Control the tradeoff between model fitting and complexity
  - Fundamental role of validation approaches (for model selection and estimations)

- Il ML è ben fondato teoricamente
  - Domande fondamentali tramite Statistical Learning Theory (#ML)
  - ed altre (e.g. PAC- probably approximately corerct learning con cenni in AIMA cap. 18.5)

Bibliography

- AIMA , ed .3: **chap 18.4**
  (thought quite simplified !!!)

- Further readings (not mandatory!):
- Every good ML book:
  - see the bibliography of the SVM lecture

For fun

Dip. Informatica
University of Pisa

- Can I have just a look to the test set?

  See https://youtu.be/XvOsh15hLIs

By Sepe-Dukic past ML students

* It can hold also for the validation set used as test set ;-)

DRAFT, please do not circulate!

For information

http://www.di.unipi.it/~micheli/DID

Alessio Micheli

micheli@di.unipi.it

www.di.unipi.it/groups/ciml

Dipartimento di Informatica
Università di Pisa - Italy

Computational Intelligence & Machine Learning Group