JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Ciro Donalek (Caltech)

Supervised Learning

Jet Propulsion Laboratory

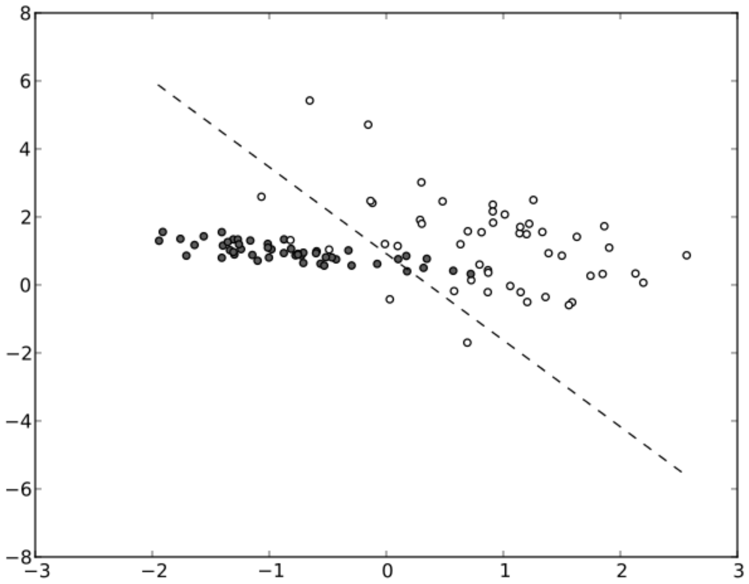
Keck Institute for Space Studies

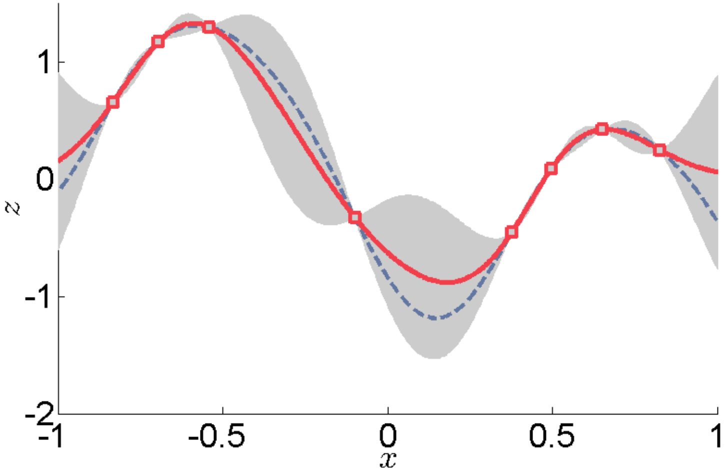
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Caltech

Outline

- Supervised Learning
- How to create a training set
- Steps
- Cross-Validation
- Overfitting

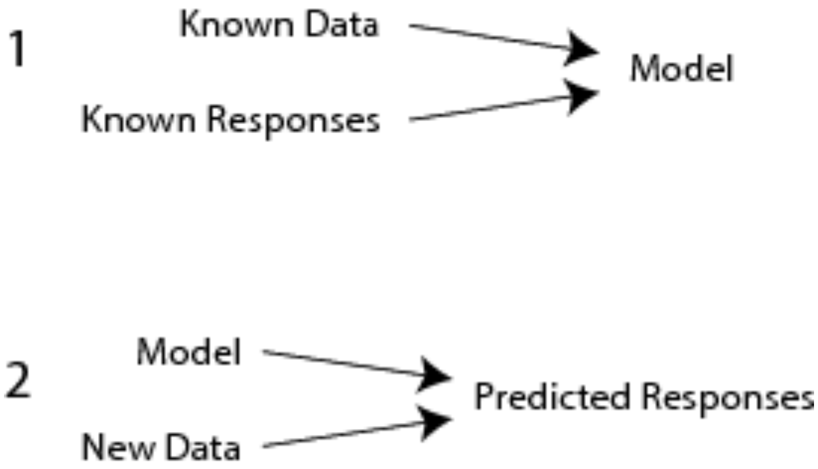




Pictures credit: Wikipedia.org

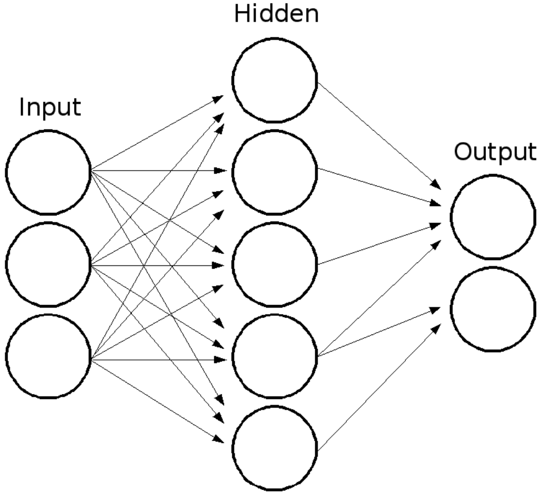
Supervised Learning

- For some examples the correct results (targets) are known and are given in input to the model during the learning process.
- Generalization: ability of a learning machine to perform accurately on new, unseen examples.



1 Known Data → Model
Known Responses → Model

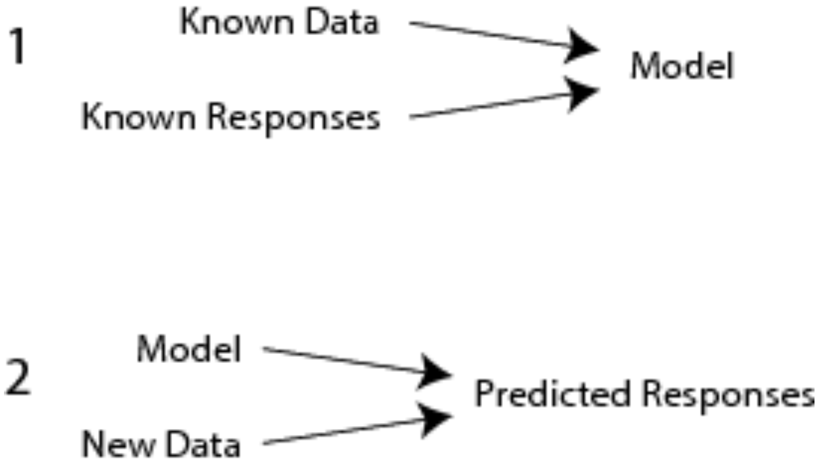
2 Model → Predicted Responses
New Data → Predicted Responses



Images credits: Mathworks

Supervised Learning

- Common tasks:
  - classification: categorical output, data can be separated in specific classes;
  - regression: for continuous outputs.

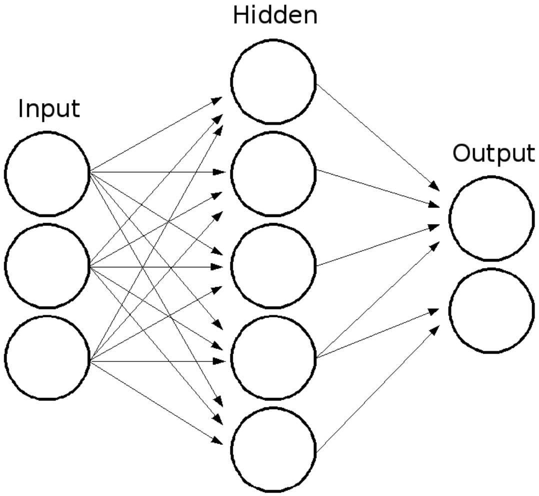


1 Known Data → Model

Known Responses → Model

2 Model → Predicted Responses

New Data → Predicted Responses



Images credits: Mathworks

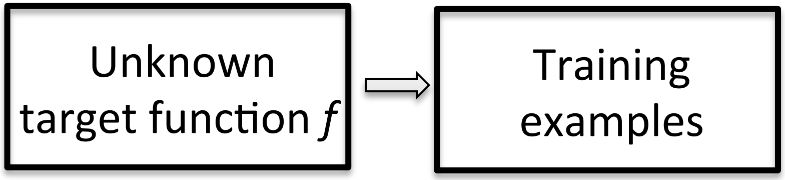
Supervised Learning

- Training data consists of a set of training examples $D = (\mathbf{x_1}, y_1),(\mathbf{x_2}, y_2),\ldots,(\mathbf{x_n}, y_n)$ where:
  - $\mathbf{x}_1, \ldots, \mathbf{x}_n$: input parameters (feature vector)
  - y : output value (target vector)
  - Example: credit approval
    - $\mathbf{x}$ = [age, gender, annual salary, current debt]
    - y = {accept, deny} = {+1, -1}

Training examples

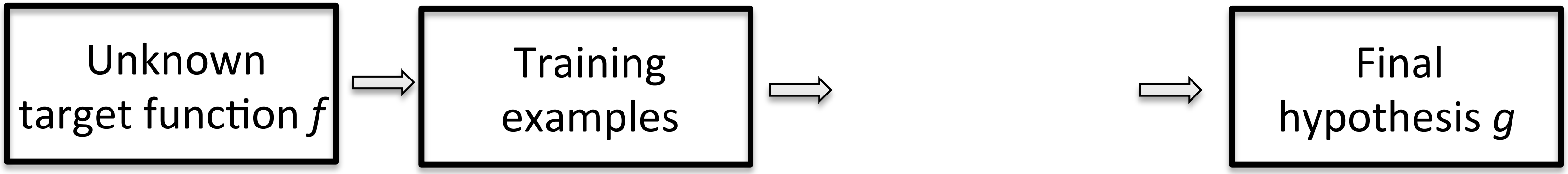
Supervised Learning

- Training data consists of a set of training examples.
- Target function: $f(\mathbf{x}_\mathrm{i}) = y_i$ (function to learn, unknown)
  - – example: ideal credit approval formula



Supervised Learning

- Training data consists of a set of training examples
- Ideal target function *f*
- Hypothesis: $g(\mathbf{x}_i) \approx y_i$ (approximate *f*, known)
  - – example: reliable credit score



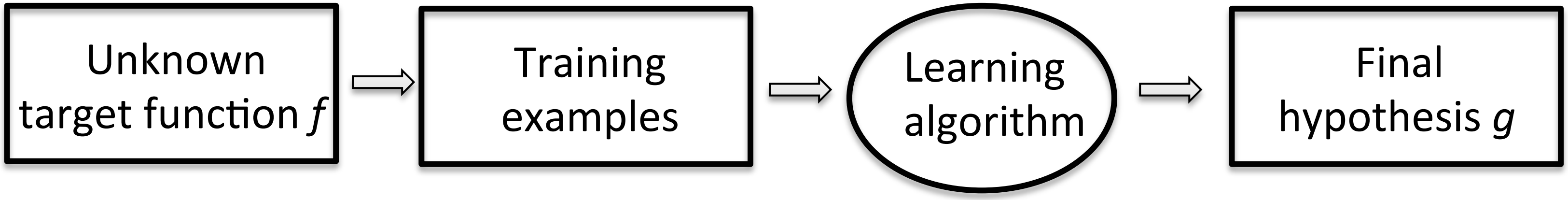
Unknown target function *f*

Training examples

Final hypothesis *g*

Supervised Learning

- Training data consists of a set of training examples
- Ideal target function *f*
- Hypothesis *g* that best approximate *f*
- Learning algorithm
  - – connect target function and hypothesis


Unknown target function *f*

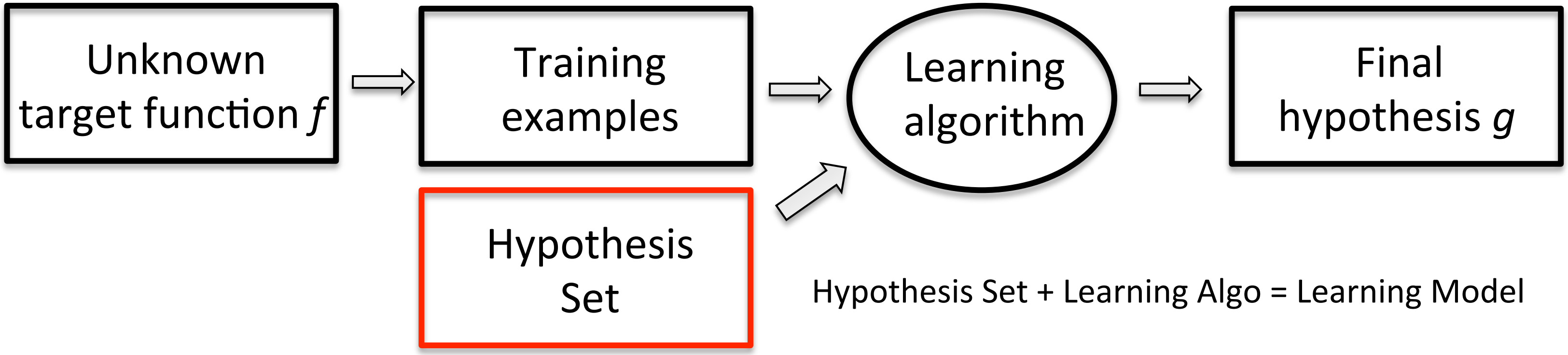
Training examples

Learning algorithm

Final hypothesis *g*

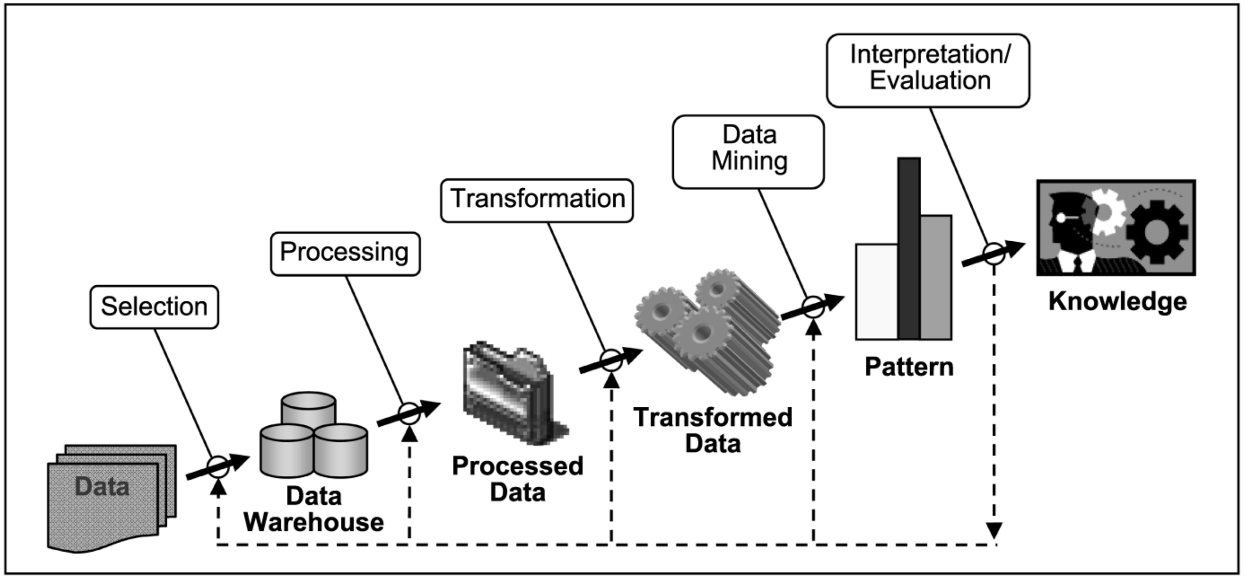
Supervised Learning

- Training data consists of a set of training examples
- Ideal target function *f*
- Hypothesis *g* that best approximate *f*
- Learning algorithm
  - – connect target function and hypothesis
- Hypothesis Set
- Predict new inputs: $y_{new}=g(\mathbf{x}_{new})$



Hypothesis Set + Learning Algo = Learning Model

Supervised Learning in a nutshell



Supervised Learning in a nutshell

① Define the data to be used as a learning set

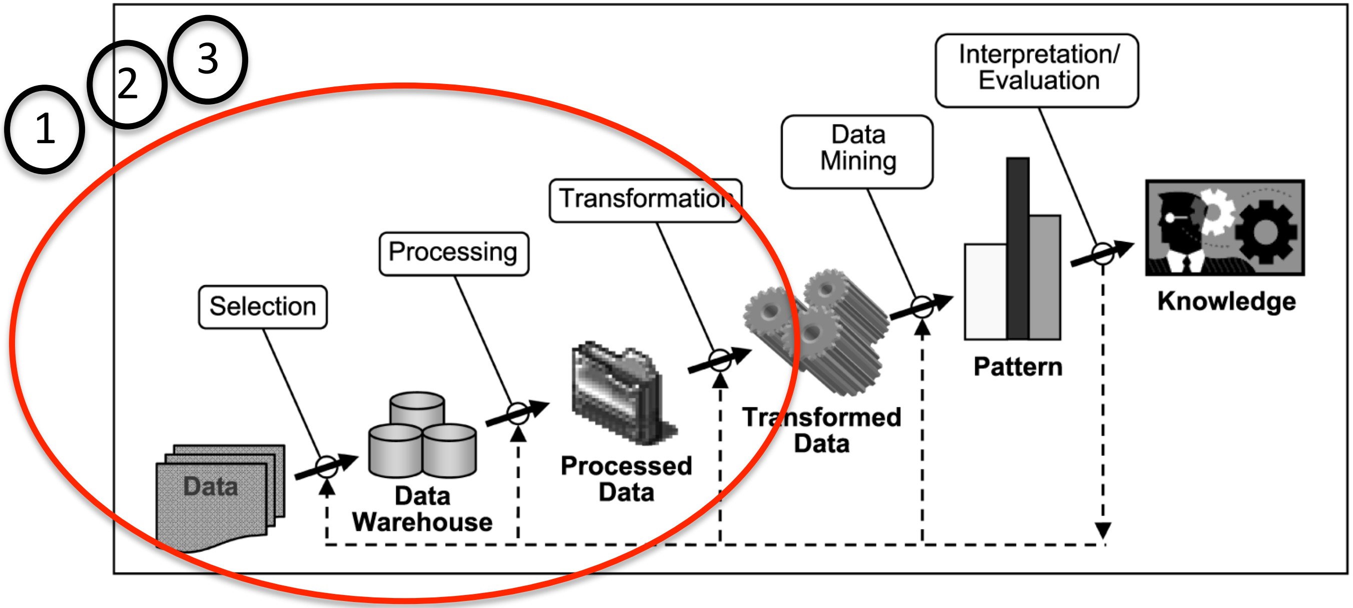
- eg, handwriting analysis: single character or entire words?

② Prepare the training set

- eg, create training, validation and test sets

③ Transform the input data in feature vectors (X,Y)

- eg, extract/select features to avoid the curse of dimensionality



Supervised Learning in a nutshell

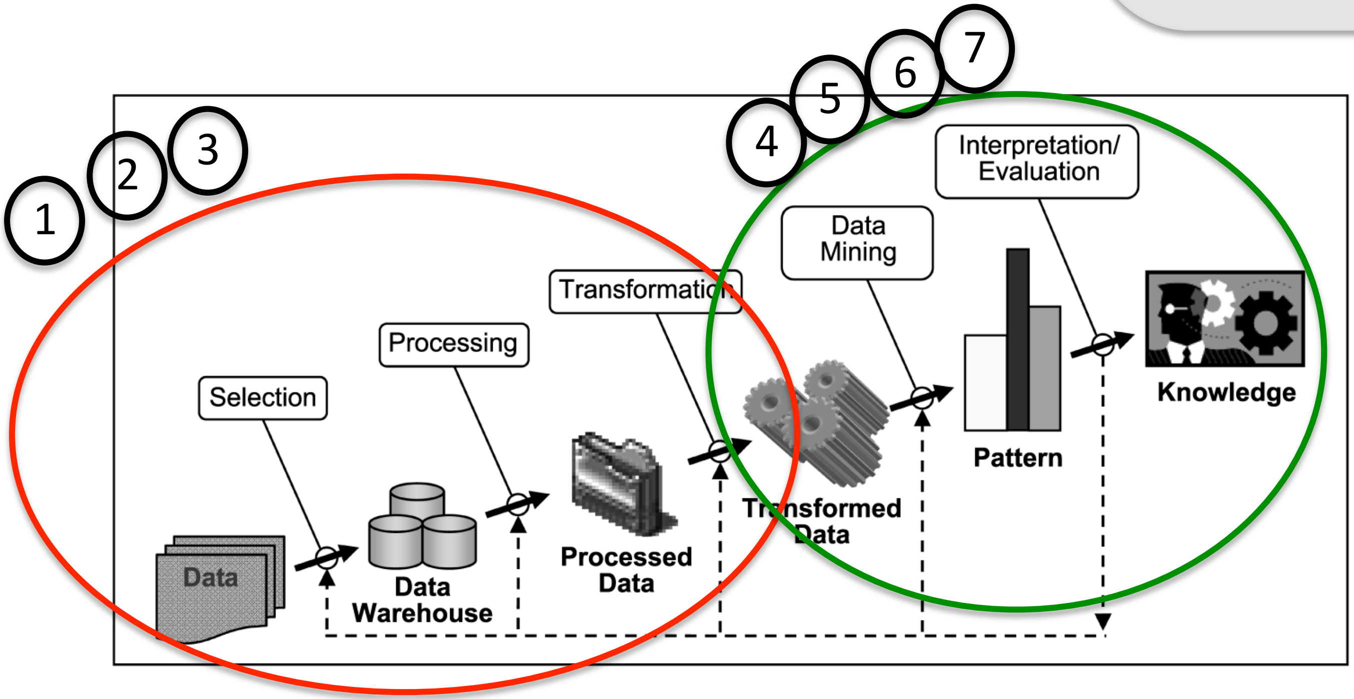
④ Choose the learning model
- eg, Neural Network and Back propagation

⑤ Choose a validation model
- eg, cross validation, random splits, etc

⑥ Run the algorithm, compute the accuracy and update until satisfied
- eg, minimize the loss, minimize the MSE, etc

⑦ Use final model to make predictions

Supervised Algorithms
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
GPR
...

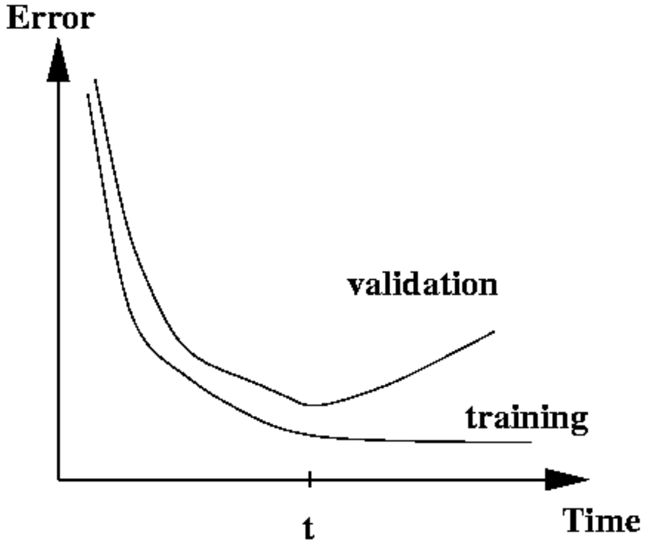


Learning set

- All eventualities must be covered: the learning dataset must be representative of the underlying model.
- Split the data in three independent data sets:
  - training set;
  - validation set;
  - test set.

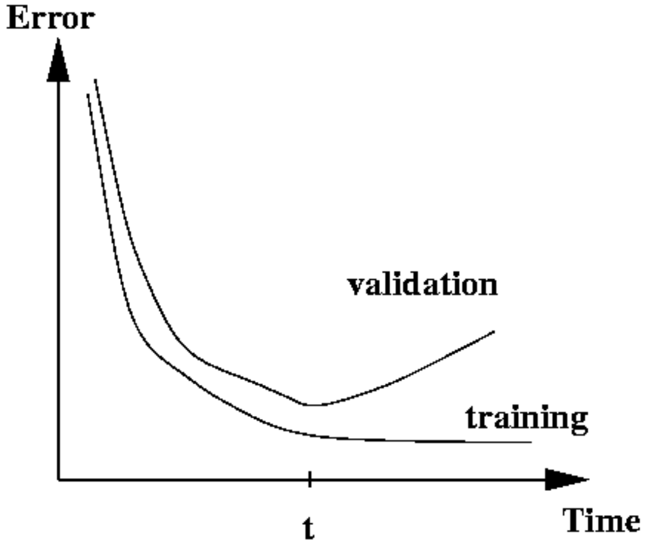
Training Set

- **Training set**: a set of examples used for learning, where the target value is known.
- The goal of the learning algorithm is to build a model which makes accurate predictions on the training set.
- Training set accuracy does not give a good indication about the generalization power of the model.
- Add a Validation set.



Validation Set

- Set of examples used to tune the architecture of a classifier and estimate the error.
- Used for model selection.
- The validation data has to be representative of the range of inputs the classifier is likely to encounter.
- How to create it?
  - – gather new data;
  - – random split:
    - 80-20
    - cross-validation

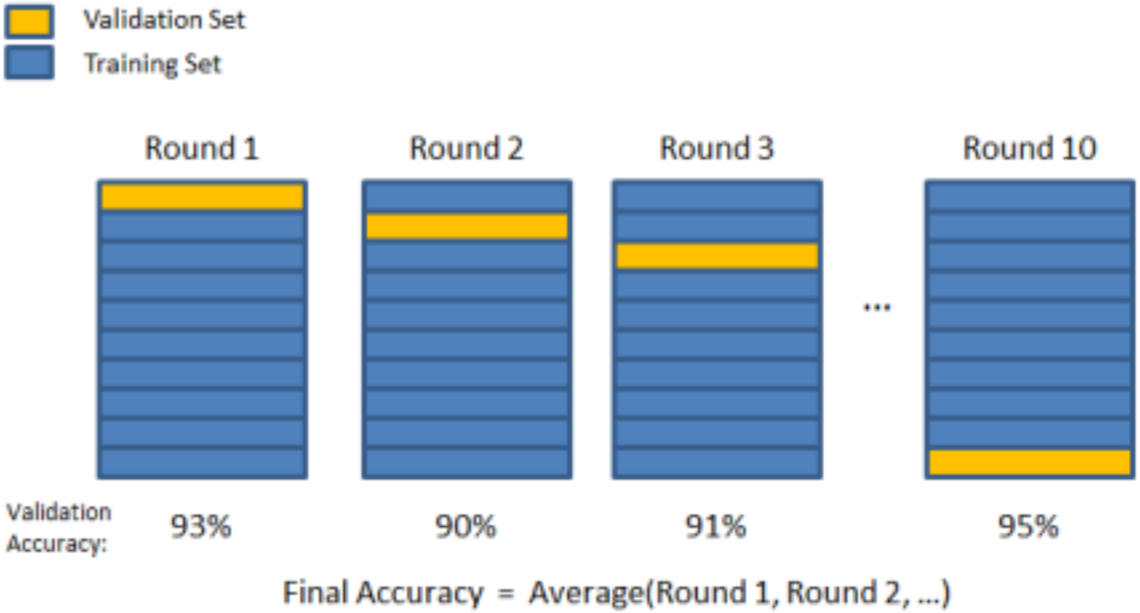


Cross-Validation

- C-V techniques are used for assessing how the results of a statistical analysis will generalize to an independent data set.
- Exhaustive Cross-Validation
  - leave one out cross validation (LOOCV)
  - leave p-out cross validation
- Non-exhaustive Cross-Validation
  - *k*-fold cross validation
  - repeated random sub-sampling validation
- Choose also according to your model/task.

K-fold cross-validation

- How it works:
  - randomly partition the original into *k* subsamples;
  - of the *k* subsamples, one is retained as the validation data for testing the model, and the remaining *k* − 1 are used as training data;
  - the process is then repeated *k times*, with each of the *k* subsamples used exactly once as the validation data;
  - the *k* results can be averaged (or otherwise combined) to produce a single estimation.



Picture credit: chrisjmccormick

Repeated random sub-sampling

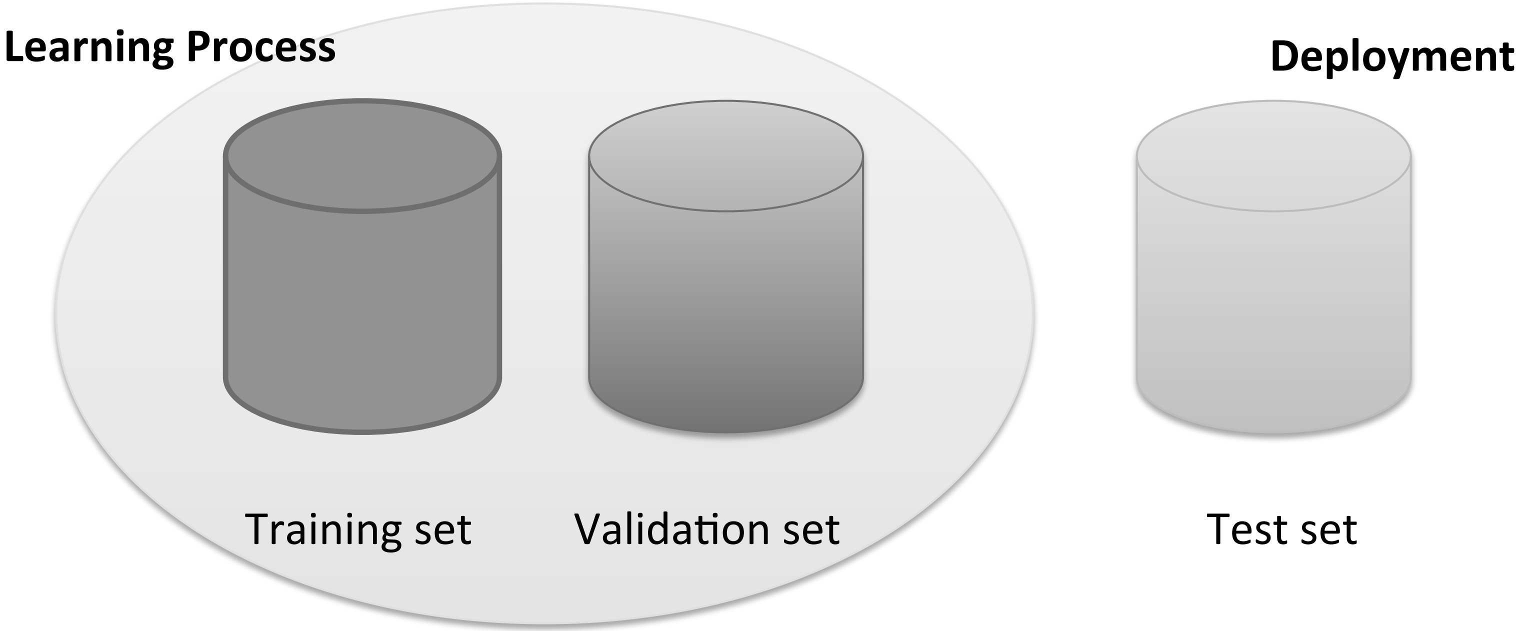
- Repeated Random Sub-Sampling
  - at each step: randomly split the dataset into two subsets: training and validation;
  - compute the validation errors.
  - average the results over the splits.
- Advantage: proportion of the sets not dependent on the number of folds.
- Disadvantage: some samples may never be selected for validation, some may be selected more than once.

LOOCV

- Leave One Out Cross Validation
  - use a single **observation** from the original sample as the validation data, and the remaining observations as the training data;
  - repeat n-times such that each observation in the sample is used once as the validation data;
  - computationally expensive.

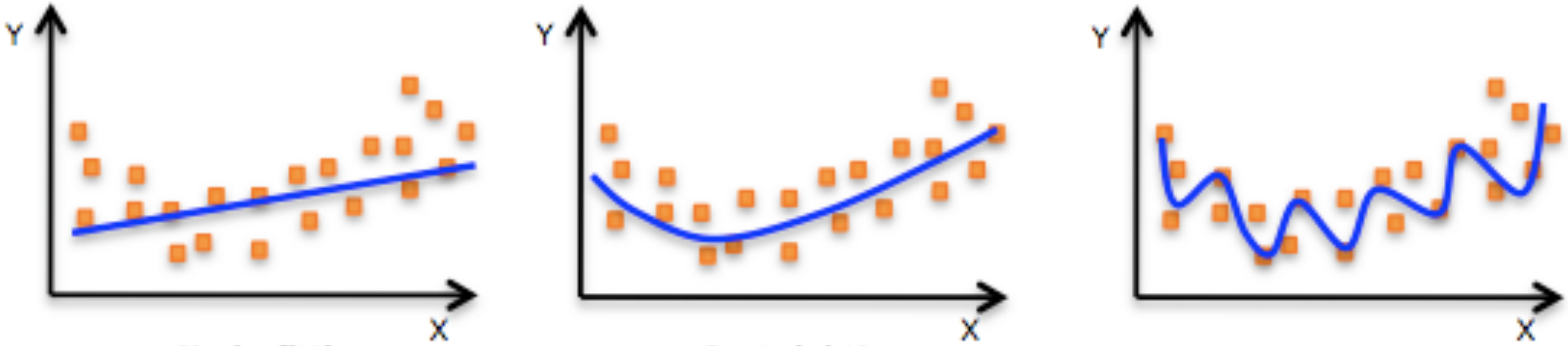
Test Set

- **Test set**: used only to assess the performances of a fully trained classifier.
- It is **never used** during the training process so that the error on the test set provides an unbiased estimate of the generalization error.



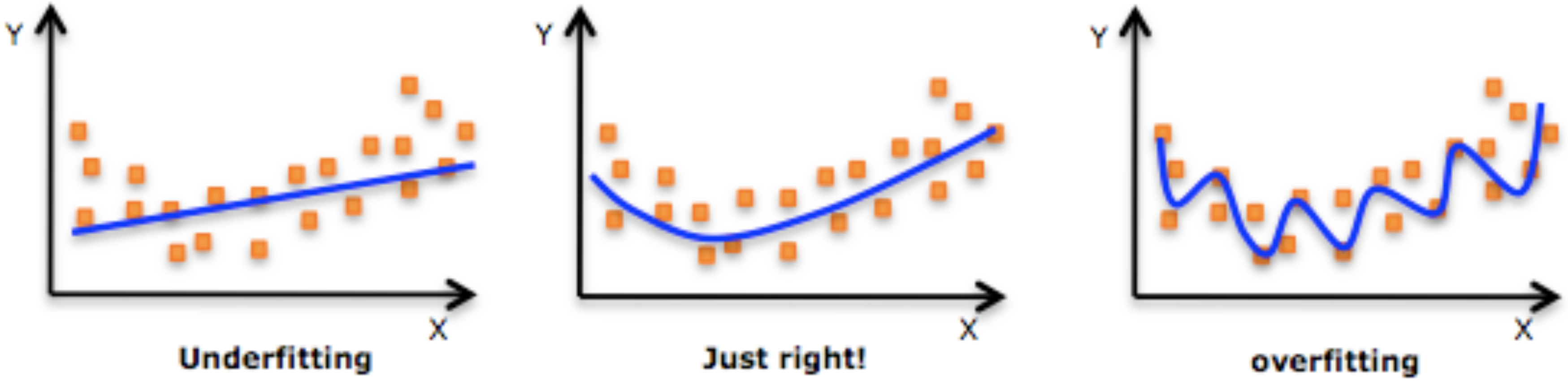
A common problem: OVERFITTING

- Model is not be able to generalize.
- Learn the “data” and not the underlying function.
- Performs well on the data used during the training and poorly with new data.
- How to avoid: cross-validation, early stopping, regularization, Bayesian priors, model comparison.



A common problem: OVERFITTING

- Model is not be able to generalize.
- Learn the “data” and not the underlying function.
- Performs well on the data used during the training and poorly with new data.
- How to avoid: cross-validation, early stopping, regularization, Bayesian priors, model comparison.



Underfitting

Just right!

overfitting

Overfitting in supervised learning

- Example: overfitting in supervised learning.
- Blu is the training error, red the validation error, over time.
- If the validation error increase while the training error decrease it is a warning sign for overfitting.

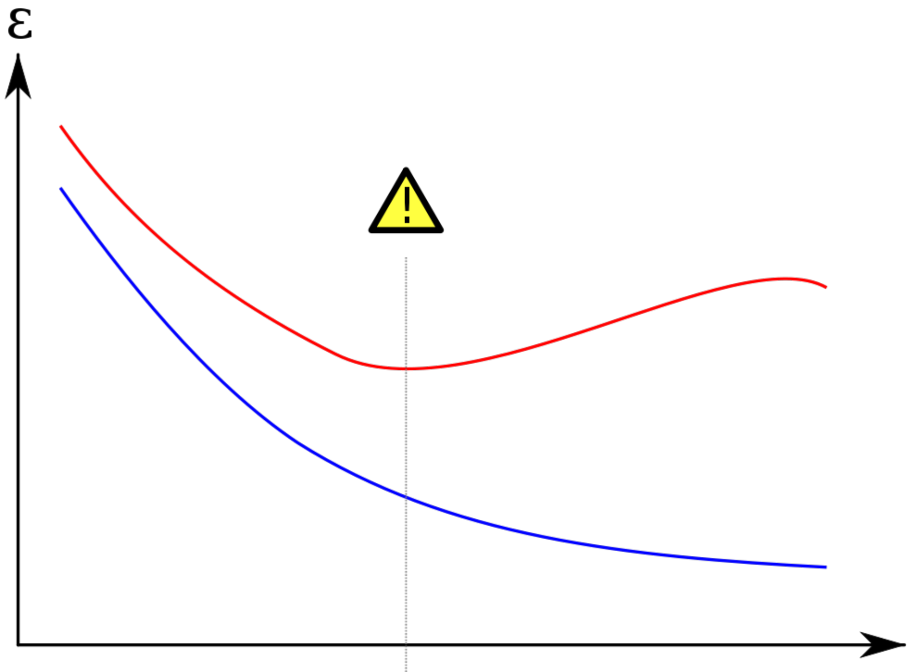


Image credit: Wikipedia.

Summary

- Supervised Learning: general concepts
- Target function and hypothesis
- Training, Validation and Test Set
- Different types of cross-validation
- Overfitting

[redacted]