Advanced Machine Learning

Lecture 7: Model Order Selection

Nora Ouzir : nora.ouzir@centralesupelec.fr
Lucca Guardiola : lucca.guardiola@centralesupelec.fr

Oct. - Nov. 2020

CentraleSupélec

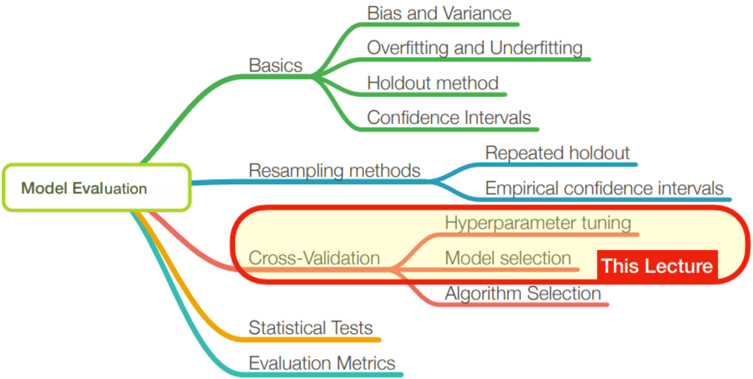
Content

1. Reminders on ML
2. Robust regression
3. Hierarchical clustering
4. Classification and supervised learning
5. Non-negative matrix factorization
6. Mixture models fitting
7. Model order selection
8. Dimension reduction and data visualization

Model Selection and Evaluation

- How to set an algorithm’s unknown parameters in practice?
- How to make choices about the model (*e.g., kernel type*)?
- How to choose the best algorithm for a particular problem?

Today’s Lecture



Today's Lecture

1. Introduction / Motivations

2. Cross-validation

3. Bayesian approaches and Information criteria
   1. The Bayesian viewpoint
   2. Information Criteria

4. Application example

Today's course

1. Introduction / Motivations
2. Cross-validation
3. Bayesian approaches and Information criteria
    1. The Bayesian viewpoint
    2. Information Criteria
4. Application example

Motivations

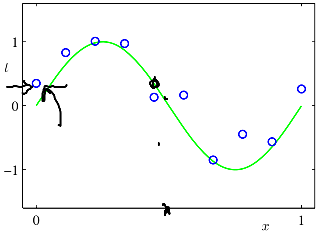
Goal: Make high-level decisions about the model we want to use.

Examples from the course

- Number of components in a mixture model K
- Target dimension for matrix factorization
- Density/minimum points parameters for density clustering
- Type of kernel in a support vector machine
- Degree of a polynomial in a regression problem

- But also a network architecture of (deep) neural networks, ...

Curve fitting example



True data generated from a sinusoid ($\sin(2\pi x)$) + (small) Gaussian noise (Bishop, 2006)

Goal: predict the value of t for some new value of x, without knowledge of the green curve → Model selection

Curve fitting example

We want to learn a prediction function y such that

$$y(x, \mathbf{w}) = \sum_{i=1}^{\boxed{M}} w_i x^i$$

- where $\boxed{M}$ is the polynomial order (unknown) and $\mathbf{w} = (w_0, \ldots, w_M)$ are the polynomial coefficients (to be learnt).

Recall:

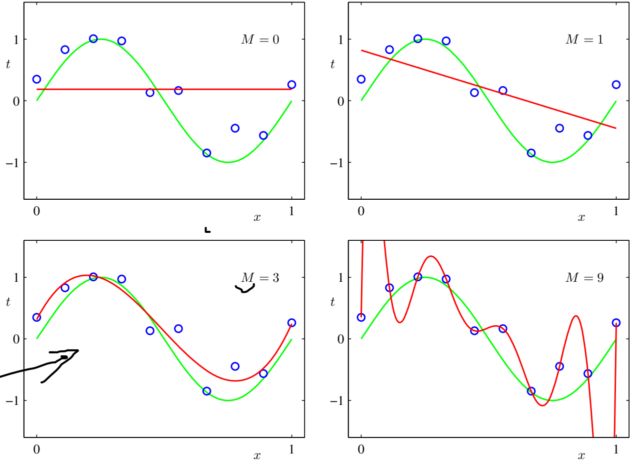
For $\underline{\mathbf{w}}$, one minimizes an *error function*, *e.g.*,

$$e(\mathbf{w}) = \sum_{n=1}^{N} \underline{\rho}\,(y(x_n, \mathbf{w}) - t_n)$$

- $\underline{N}$: number of observed data.
- $e(\mathbf{w})$ is a quadratic function w.r.t $\mathbf{w}$ ⇒ unique solution $\mathbf{w}^*$

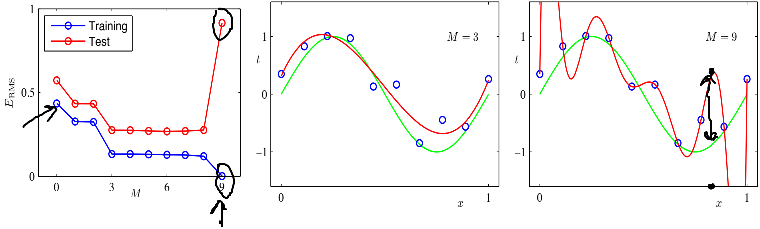
Problem: still need to choose *M*!!!

Curve fitting example



How to evaluate the “best model”?

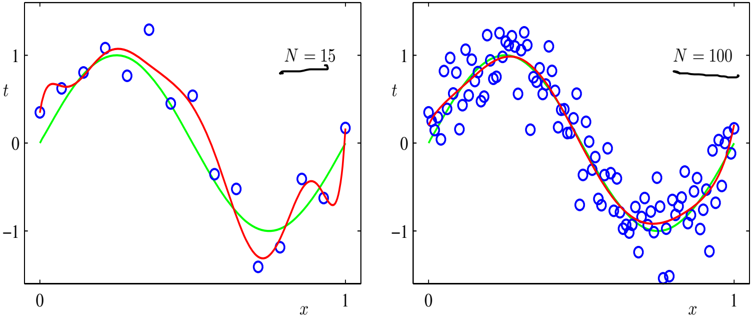
Curve fitting example: Test vs training data



Differences btw training and test datasets

- Model fits training data perfectly, but may not do well on test data: ($M = 9, e_{RMS} = 0$, but poor estimation of $\sin(2\pi x)$)) → Overfitting
- Training performance $\neq$ test performance, but we are mainly interested in test performance → Generalization
- Need mechanisms for assessing how a model generalizes to unseen test data → Model selection

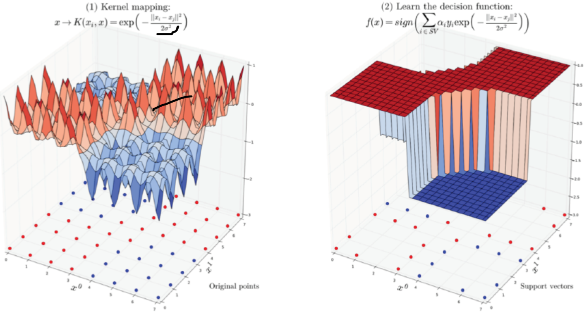
Curve fitting example: Sample size



Solutions for $\underline{M = 9}$ with different numbers of data points

Increasing the size of the data set reduces over-fitting

Example: SVM with Gaussian Kernel

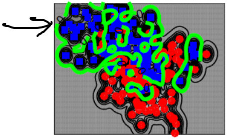


(1) Kernel mapping:
$x \to K(x_i, x) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

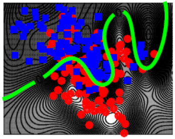
Original points

(2) Learn the decision function:
$f(x) = sign\left(\sum_{i \in SV} \alpha_i y_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right)$

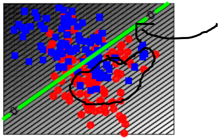
Support vectors



σ too small

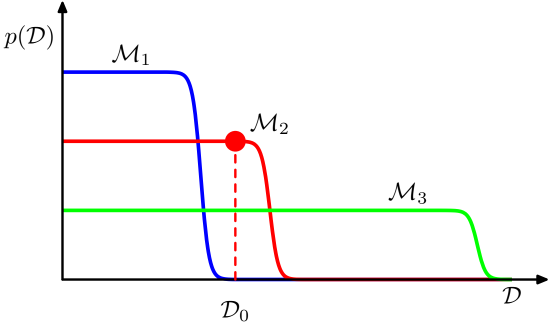


nice σ



σ too large

Model choice - Occam's (Ockham) Razor



(Normalized) dist. of data sets for three models of different complexity, in which $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex - $\mathcal{D}_0$: observed dataset - $\mathcal{M}_2$ with intermediate complexity has the largest evidence (Bishop, 2006)

Key Idea: choose the simplest model that explains "reasonably" well the data

Model selection and evaluation

Raised issues

- Model evaluation : what measure(s) of performance?
- Estimation of the generalisation capacity of the model
- Practical model selection procedures

True vs Empirical Risk

$$\rightarrow R_N(y) = \frac{1}{N}\sum_{i=1}^{N} \rho(t_i - y(x_i)) \quad / \quad \text{dataset: } \mathcal{D}$$

Expected error (generalization ability)

- $R(y) = \mathbb{E}_{x,T}\, \rho(t, y(X)) = \int_{x,T} \rho(t, y(x)) \underbrace{p(x,t)}_{} \, dx\, dt$

$\rightarrow$ min $R(y)$

$\rightarrow$ Approximation to $\boxed{R}$?

R_N : complex $\uparrow$: $R_N \rightarrow 0$ $\quad R \uparrow\uparrow$

True vs Empirical Risk

True vs Empirical Risk

Supremum on generalization error

$$R(y) \leq \overbrace{\frac{1}{N}\sum_{i=1}^{N}\rho(y(x_i),t_i)} + \overbrace{term(N,h(\mathcal{M}))}$$

$\mathcal{O}(M(\theta))$

w

$\underline{M}$

where h stands for the complexity of the model $\mathcal{M}$.

- The empirical risk is not sufficient for estimating the true error R
- If h increases → overfitting
- The bigger the N the better!

Generalization

We are looking at $\overbrace{R(\mathcal{D}_\infty, y)}$ the theoretical performance of y on all possible future data

→ Idea: test on data other than those used for training

Today’s course

1. Introduction / Motivations

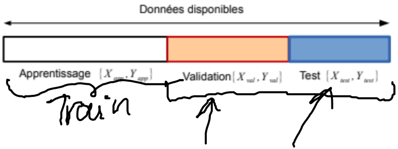
2. Cross-validation

3. Bayesian approaches and Information criteria
   1. The Bayesian viewpoint
   2. Information Criteria

4. Application example

Validation set

Key Idea: Choose the best model without testing on $\mathcal{D}$~~test~~ train

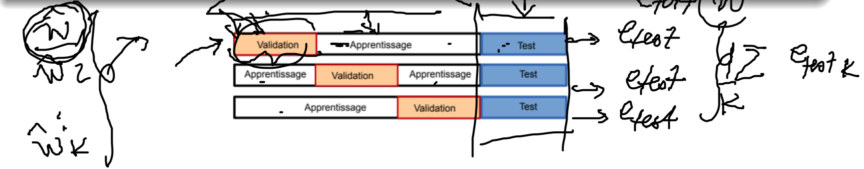

Données disponibles
Apprentissage $\{X_{app}, Y_{app}\}$
Validation $\{X_{val}, Y_{val}\}$
Test $\{X_{test}, Y_{test}\}$
Train

1. Randomly split $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{test}$
2. Train all possible models on $\mathcal{D}_{train}$
3. Evaluate the performance on $\mathcal{D}_{val}$
4. Select the model with the best performance on $\mathcal{D}_{val}$
5. Test the selected model on $\mathcal{D}_{test}$ (used only once)

If data is limited we are wasting training data...

K-fold Cross-validation

Key Idea: give an accurate estimate of the true error without wasting too much data



Validation | Apprentissage | Test

Apprentissage | Validation | Apprentissage | Test

Apprentissage | Validation | Test

1. Randomly split the data into $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$
2. Randomly split $\mathcal{D}_{train} = \mathcal{D}_1 \cup ... \cup mathcalD_K$ into K subsets $\mathcal{D}_K$
3. For $k = 1, ..., K$ $\mathcal{D}_K$
   - 3.1 Train the model on $K - 1$ sets
   - 3.2 Evaluate performance on the remaining set $\mathcal{D}_k$
4. Average the K measures of performance

N.B: To reduce variability, multiple rounds of cross-validation are performed using different partitions then averaged.

K-fold Cross-validation

The averaged error can be expressed as

$$\underbrace{R_{CV}}_{} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}\underbrace{\rho(t_i^k, y^{k'}(x_i^k))}_{}$$

where $y^{k'}$ is learnt using the $K-1$ partitions exept the k-th one.

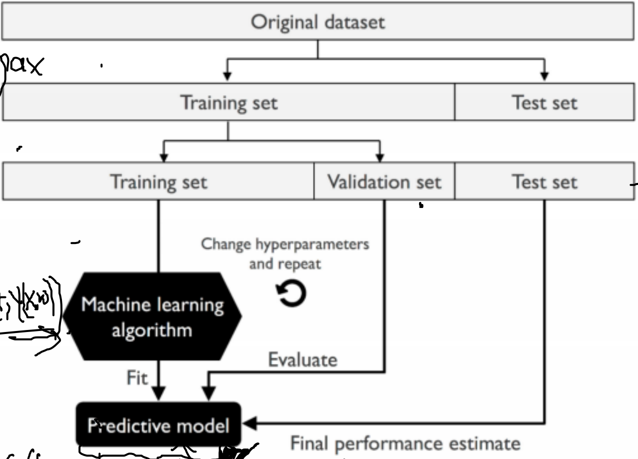
↓ ↓

- if $K = N$, cross-validation is approximately an unbiased estimator of the generalization error → too expensive!
- Potential high variance
- Typically we choose $\overline{K = 5}$ or $\overline{K = 10}$ for a satisfactory bias-variance trade-off.

Cross-validation: Practical procedure

$M = M_{min}, \dots, M_{max}$

$= 1, \dots, 10$



Original dataset

Training set

Test set

Training set

Validation set

Test set

Change hyperparameters and repeat

Machine learning algorithm

Fit

Evaluate

Predictive model

Final performance estimate

$k=1$

$\hookrightarrow w(k, M)$

$w_{1,1}$

$w_{1,2}$

$\vdots$

$w_{1,10}$

$LS(t, y(x, w))$

$M^* = \arg\min_{M} LS(t, y(x_k, w_{k,M}))$

$R_{CV}(M=1)$

$R_{CV}(M=2 \dots)$

$\rho(x_{test}, \hat{M}_1)$

$\hat{M}_{1}$

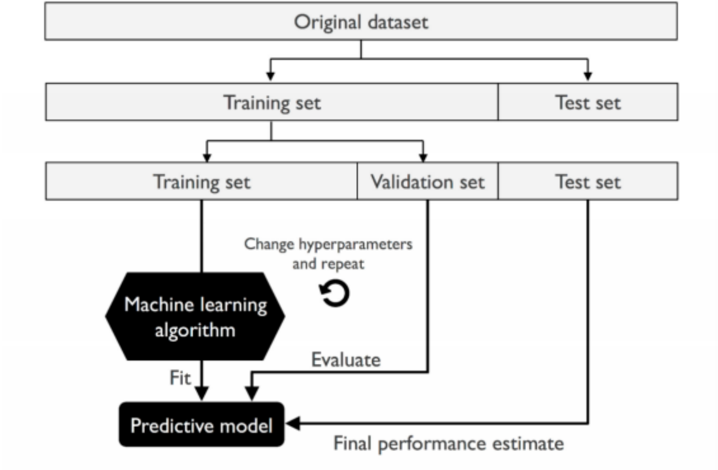
$\hat{M}_{2}\ \rho(x_{test}, \hat{M}_2)$

$\vdots$

$\hat{M}_{10}\ \rho(\dots, M)$

$\rightarrow \frac{1}{K} \sum \rho(\dots)$

Cross-validation: Practical procedure



Other performance measures to evaluate the model...

Today's course

1. Introduction / Motivations

2. Cross-validation

3. Bayesian approaches and Information criteria
    1. The Bayesian viewpoint
    2. Information Criteria

4. Application example

Today's course

1. Introduction / Motivations

2. Cross-validation

3. Bayesian approaches and Information criteria
    1. The Bayesian viewpoint
    2. Information Criteria

4. Application example

The Bayesian viewpoint

Key Idea: Place a prior $p(\mathcal{M})$ on the class of models

i=1, ... L

Ingredients

- Given a training set $\mathcal{D}$, the posterior distribution over models is

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$$

- Model evidence (marginal likelihood):

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\theta_i)\,p(\theta_i|\mathcal{M}_i)\,d\theta_i$$

Highest evidence: If we don't have any prior knowledge of the model → uniform $p(\mathcal{M})$ and

choose $\mathcal{M}_1$ over $\mathcal{M}_2$ $\iff$ $p(\mathcal{D}|\mathcal{M}_1) > p(\mathcal{D}|\mathcal{M}_2)$

Powerful rule but the integral is intractable in practice...

Bayesian Model Averaging

Key Idea: Do not choose a model but make predictions using average over several models with weights given by the posterior probability of each model given the data

- Place a prior $\underline{p(\mathcal{M})}$ on the class of models
- Instead of selecting the "best" model, integrate out the corresponding model parameters $\theta_{\mathcal{M}}$ and average over all models $\mathcal{M}_i, i = 1, \ldots, L$

$$
\begin{aligned}
\underline{p(\mathcal{D})} &= \int \sum_{i=1}^{L} \underline{p(\mathcal{M}_i)}\, \underline{p(\mathcal{D}|\theta_i)\, p(\theta_i|\mathcal{M}_i)}\, d\theta_i \\
&= \sum_{i=1}^{L} \underline{p(\mathcal{M}_i)\, \boxed{p(\mathcal{D}|\mathcal{M}_i)}}
\end{aligned}
$$

- Generally gives better answers than a single model
- Computationally expensive and integral often intractable

Today’s course

1. Introduction / Motivations

2. Cross-validation

3. Bayesian approaches and Information criteria
    1. The Bayesian viewpoint
    2. Information Criteria

4. Application example

Information criteria

Maximized

Key Idea: Correct for the bias of MLE by adding a penalty term to compensate for the overfitting of more complex models (large nbr of parameters)

Let M_j be the number of unknown parameters for model $\mathcal{M}_j$ and N the data size

→ Akaike Information Criterion (AIC)

$$\boxed{AIC(j)} = \overline{\ln(p(\mathbf{x}|\hat{\theta}_{ML}))} - \boxed{M_j}$$

Information criteria

Key Idea: Correct for the bias of MLE by adding a penalty term to compensate for the overfitting of more complex models (large nbr of parameters)

Let M_j be the number of unknown parameters for model $\mathcal{M}_j$ and N the data size

Akaike Information Criterion (AIC)

$$AIC(j) = \ln(p(\mathbf{x}|\hat{\theta}_{ML})) - M_j$$

Bayesian Information Criterion (BIC) / Minimum Description Length (MDL)

$$BIC(j) = \overbrace{\ln(p(\mathbf{x}|\hat{\theta}_{ML}))} - \frac{1}{2}\boxed{M_j}N \longleftarrow$$

Information criteria

Key Idea: Correct for the bias of MLE by adding a penalty term to compensate for the overfitting of more complex models (large nbr of parameters)

Let M_j be the number of unknown parameters for model $\mathcal{M}_j$ and N the data size

Akaike Information Criterion (AIC)

$$AIC(j) = \ln(p(\mathbf{x}|\hat{\theta}_{ML})) - M_j$$

Bayesian Information Criterion (BIC) / Minimum Description Length (MDL)

$$BIC(j) = \ln(p(\mathbf{x}|\hat{\theta}_{ML})) - \frac{1}{2} M_j \overset{\log(N)}{\cancel{N}}$$

No integrals or posteriors involved → Much easier to compute!!

Information Criteria: Practical guidelines

No clear choice between AIC and BIC, but generally:

- BIC penalizes model complexity more heavily than AIC
  - → For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity

Information Criteria: Practical guidelines

No clear choice between AIC and BIC, but generally:

- BIC penalizes model complexity more heavily than AIC
  - → For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity
- BIC usually works better if the true model is among $\mathcal{M}_1, \ldots, \mathcal{M}_L$
  - → BIC is asymptotically consistent
  - → AIC tends to choose models which are too complex as $N \rightarrow \infty$

Information Criteria: Practical guidelines

No clear choice between AIC and BIC, but generally:

- BIC penalizes model complexity more heavily than AIC
  - → For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity
- BIC usually works better if the true model is among $\mathcal{M}_1, \ldots, \mathcal{M}_L$
  - → BIC is asymptotically consistent
  - → AIC tends to choose models which are too complex as $N \to \infty$
- AIC is usually preferable otherwise (“all models are wrong”)

Information Criteria: Practical guidelines

No clear choice between AIC and BIC, but generally:

- BIC penalizes model complexity more heavily than AIC

  → For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity

- BIC usually works better if the true model is among $\mathcal{M}_1, ..., \mathcal{M}_L$

  → BIC is asymptotically consistent

  → AIC tends to choose models which are too complex as $N \to \infty$

- AIC is usually preferable otherwise (“all models are wrong”)

- Both are mainly suited to N much larger than p (typically $d << \sqrt{N}$)

Other techniques and criteria

Many others techniques:

- Minimum Message Length (see application - Bayesian criterion)
- Modified AIC accounting for small sample size:

$$mAIC(j) = \ln(p(\mathbf{x}|\hat{\theta}_{ML})) - M_j - \frac{M_j(M_j+1)}{N - M_j - 1}$$

- Hypothesis testing vs Bayesian model comparison, ...

For mixture models

- All previous techniques
- Split and merge[1](see applications + lab)
- Reversible jump[2]

[1] Zhang, Z. et al., (2003). EM algorithms for Gaussian mixtures with split-and-merge operation. Pattern recognition, 36(9), 1973-1983.

[2] Zhang, Z. et al., (2004). Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Statistics and Computing, 14(4), 343-355.

Today’s course

1. Introduction / Motivations

2. Cross-validation

3. Bayesian approaches and Information criteria
    1. The Bayesian viewpoint
    2. Information Criteria

4. Application example