An introduction to statistical learning theory (2nd part)

Gianluca Bontempi

Machine Learning Group

Département d'Informatique

Boulevard de Triomphe - CP 212

http://www.ulb.ac.be/di

Algorithm-based assessment

Consider

1. an assessment criterion the Mean Integrated Squared Error (MISE)

$$\text{MISE}(N) = E_{\mathbf{D}_N}[R(\alpha_N)]$$

2. a parametric class of *hypothesis* functions $h(\cdot, \alpha)$ with $\alpha \in \Lambda$. Let us consider the case where the set Λ is structured into a nested sequence of sets:

$$\Lambda_1 \subseteq \cdots \subseteq \Lambda_s \subseteq \ldots \Lambda_S = \Lambda \tag{1}$$

where s is an *index of complexity*.

Model selection

The learning problem can then be decomposed in two steps:

1. the estimation of MISE. This can be performed either in an analytical or resampling way.
2. the selection of the optimal complexity on the basis of the estimate. The algorithm of model selection uses the estimates the generalization error of the learned hypothesis function $h(\cdot, \alpha_{D_N}^s)$ for $s = 1, \ldots, S$ and returns the one expected to have the lowest one.

Analytical estimation of MISE

We will consider two methods:

Final prediction error: this method holds for linear targets and linear hypothesis functions.

Akaike criterion: this method holds when the target and the hypothesis belong to the same parametric family.

Expectation of the empirical error

The expectation of the residual sum of squares can be written as

$$E_{\mathbf{D}_N}[R_{\text{emp}}] = E_{\mathbf{D}_N}[\mathbf{e}^T\mathbf{e}] = E_{\mathbf{D}_N}[\mathbf{Y}^T P\mathbf{Y}] = \sigma^2_{\mathbf{w}}\text{tr}(P) + E[\mathbf{Y}^T]\ P\ E[\mathbf{Y}]$$

Since $\text{tr}(ABC) = \text{tr}(CAB)$

$$\text{tr}(P) = \text{tr}(I - H) = N - \text{tr}(X(X^TX)^{-1}X^T) =$$
$$= N - \text{tr}(X^TX(X^TX)^{-1}) = N - \text{tr}(I_p) = N - p$$

we have

$$E[\mathbf{e}^T\mathbf{e}] = (N-p)\sigma^2_{\mathbf{w}} + (X\beta)^T P(X\beta) =$$
$$= (N-p)\sigma^2_{\mathbf{w}} + \beta^T X^T(I - X(X^TX)^{-1}X^T)X\beta = (N-p)\sigma^2_{\mathbf{w}}$$

This is the expectation of the error made by a linear model trained on D_N to predict the value of the output in D_N.

Bisedness of the empirical error

It can be shown that in the linear case

$$E_{\mathbf{D}_N}[R_{\text{emp}}] = E_{\mathbf{D}_N}[\mathbf{e}^T\mathbf{e}] \neq \text{MISE}$$

As a consequence, if we replace R_{emp} with

$$\mathbf{e}^T\mathbf{e} + 2\sigma^2_{\mathbf{w}}p$$

we obtain an unbiased estimator.

Nevertheless, this estimator requires an estimate of the noise variance.

Final prediction error

- Given an a priori estimate $\hat{\sigma}^2_{\mathbf{w}}$ we have the **Predicted Square Error (PSE)** criterion

$$\text{PSE} = R_{\text{emp}}(\alpha_N) + 2\hat{\sigma}^2_{\mathbf{w}} p$$

where $\alpha_N = \hat{\beta}$ and $p = n + 1$.

- Taking as estimate of $\sigma^2_{\mathbf{w}}$

$$\hat{\sigma}^2_{\mathbf{w}} = \frac{1}{N-p} R_{\text{emp}}(\alpha_N)$$

we have the **Final Prediction Error (FPE)**

$$\text{FPE} = \frac{1 + p/N}{1 - p/N} R_{\text{emp}}(\alpha_N)$$

Maximum likelihood formulation of learning

- The derivation of the Akaike criterion is based on a *maximum likelihood* formulation of the learning problem.
- Consider a training set D_N generated according to the probability distribution $F_{\mathbf{z}}(\cdot)$ and a parametric class of density functions $g(z, \theta)$, with $\theta \in \Theta$, which approximates the probability $F_{\mathbf{z}}(\cdot)$.
- the maximum likelihood approach consists in returning an estimate of θ on the basis of the training set D_N. The estimate is

$$\theta_N = \arg \max_{\theta \in \Theta} L_{\text{emp}}(\theta)$$

where $L_{\text{emp}}(\theta)$ is the empirical log-likelihood

$$L_{\text{emp}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ln g(z_i, \theta)$$

- We define, in analogy with the functional risk, the quantity $L(\theta)$, i.e. the expected log-likelihood of the distribution $g(\cdot, \theta)$

$$L(\theta) = \int_Z p(z) \ln g(z, \theta) dz \tag{2}$$

The quantity $L(\theta)$ is negative and represents the accuracy of $g(z, \theta)$ as estimator of $p_{\mathbf{z}}(z)$. Therefore, the larger the value of $L(\theta)$ the better is the approximation of $p(z)$ returned by $g(z, \theta)$.

- We define with θ_0 the parameter of class θ which maximizes the expected log-likelihood:

$$\theta_0 = \arg \max_{\theta \in \Theta} L(\theta)$$

- The maximum likelihood analogous of the MISE is the *mean expected log-likelihood*

$$L_N = E_{\mathbf{D}_N}[L(\boldsymbol{\theta}_N)] = \int_{Z^N} l(\theta_N) dP^N(D_N) \tag{3}$$

that is the average over all possible realizations of a dataset of size N.

The Akaike's information criterion

- The derivation of the Akaike's criterion is based on the following assumption: *there exists a value* $\theta^* \in \Theta$ *so that the probability model* $g(z, \theta^*)$ *is equal to the underlying distribution* $P(z)$.
- The Akaike's criterion states that the quantity

$$\text{AIC} = l_{emp}(\theta_N) - \frac{n}{N}$$

  is an unbiased estimate of the mean expected log-likelihood.

Resampling techniques for MISE estimation

The most known techniques are

Cross-validation: it is known to provide a nearly unbiased estimate of the MISE. However, the low bias of cros-validation is often paid by high variability.

Bootstrap: they aim to reduce the variability of MISE predictions.

Cross-validation

- The basic idea of cross-validation is that one builds a model from one part of the data and then uses that model to predict the rest of the data.
- The dataset D_N is split k times in a training and a test subset, the first containing N_{tr} samples, the second containing $N_{ts} = N - N_{tr}$ samples.
- Each time, N_{tr} samples are used by the parametric identification algorithm $\mathcal{L}$ to select a hypothesis $\alpha^i_{N_{tr}}$, $i = 1, \ldots, k$, from Λ and the remaining N_{ts} samples are used to estimate the error of $h(\cdot, \alpha^i_{N_{tr}})$
- A common form of cross-validation is the "leave-one-out" (l-o-o). Let $D_{(i)}$ be the training set with z_i removed, and $h(x, \alpha_{N(i)})$ be the corresponding prediction rule. The l-o-o cross-validated error estimate is

$$\widehat{\mathrm{MISE}} = \frac{1}{N} \sum^{N}$$

Leave-one-out

- A common form of cross-validation is the "leave-one-out" (l-o-o). Let $D_{(i)}$ be the training set with z_i removed, and $h(x, \alpha_{N(i)})$ be the corresponding prediction rule. The l-o-o cross-validated error estimate is

$$\widehat{\mathrm{MISE}}_{\mathrm{LOO}} = \frac{1}{N} \sum_{i=1}^{N} C\left(y_i, h(x_i, \alpha_{N(i)})\right)$$

Leave-one-out unbiasedness

The almost unbiasedness of leave-one-out can be shown in the following manner

Theorem 1 (Devroye et al. 1996).

$$E_{\mathbf{D}_N}[\widehat{MISE}_{LOO}] = E_{\mathbf{D}_{N-1}}[R(\boldsymbol{\alpha}_{N-1})] = MISE(N-1)$$

Proof.

$$E_{\mathbf{D}_N}[\widehat{\mathrm{MISE}}_{\mathrm{LOO}}] = E_{\mathbf{D}_N}\left[\frac{1}{N}\sum_{i=1}^{N} C\left(y_i, h(x_i, \alpha_{N(i)})\right)\right] =$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_{\mathbf{x},\mathbf{y},\mathbf{D}_{N-1}}\left[C\left(\mathbf{y}, h(\mathbf{x}, \alpha_{N-1})\right)\right] = E_{\mathbf{D}_{N-1}}\left[R(\boldsymbol{\alpha}_{N-1})\right] = \mathrm{MISE}(N-1)$$

□

Other estimation techniques

- Cross-validation provides a nearly unbiased estimate of MISE but the low bias is often paid by high variability.
- Bootstrap estimates can be thought of as smoothed versions of cross-validation.
- The bootstrap has another important advantage: it also provides a direct assessment of the variability of the estimates of
  1. the parameters α_N,
  2. the prediction error MISE.

Boostrap estimates of prediction error (1)

- The simplest bootstrap approach generates B bootstrap samples $D_{(b)}$, estimates the model $h(\cdot, \alpha_{(b)})$ on each of them, and then applies each fitted model to the original sample D_N to give B estimates

$$\widehat{\mathrm{MISE}}_{\mathrm{BOO}}^{(b)} = \sum_{i=1}^{N} (y_i - h(x_i, \alpha_{(b)}))^2$$

  of the prediction error.

- The overall bootstrap estimate of prediction error is the average of these B estimates.

$$\widehat{\mathrm{MISE}}_{\mathrm{BOO}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathrm{MISE}}_{\mathrm{BOO}}^{(b)} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N} \sum_{i=1}^{N} (y_i - h(x_i, \alpha_{(b)}))^2$$

Boostrap estimates of prediction error (2)

- This simple bootstrap approach turns out not to work very well. A second way to employ the bootstrap pardigm is to estimate the bias (or *o*ptimism) of the empirical risk.

$$\widehat{\text{Bias}}^{\text{MSE}_{\text{emp}}} = \text{MISE} - E_{\mathbf{D}_N}[R_{\text{emp}}]$$

- The bootstrap estimate of this quantity is obtained by generating B bootstrap samples $D_{(b)}$ estimating the model $h(\cdot, \alpha_{(b)})$ for each of them and calculating the difference between the MISE on D_N and the empirical risk on $D_{(b)}$.

$$\widehat{\text{Bias}}^{\text{MSE}_{\text{emp}}}_{\text{BOO}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\text{MISE}}_{\text{BOO}} - \frac{1}{B}\sum_{b=1}^{B} R^{(b)}_{\text{emp}}$$

- The final estimate is then

$$\widehat{\text{MISE}}^{2}_{\text{BOO}} = R_{\text{emp}} + \text{Bias}^{R_{\text{emp}}}_{\text{BOO}}$$

Other bootstrap estimates

- Let us consider the simple bootstrap estimator $\widehat{\text{MISE}}_{\text{BOO}}$. This is obtained by calculating the prediction error of $\alpha_{(b)}$ for each elements of D_N.
- One problem with this estimate is that we have points belonging to the training set $D_{(b)}$ and test set D_N. In particular it can n be shown that the percentage of points belonging to both is 63.2%.
- In order to remedy to this problem, an idea could be to consider as test samples only the ones that do not belong to $D_{(b)}$.

Other bootstrap estimates (II)

- We will define by $\widehat{\mathrm{MISE}}^{0}_{\mathrm{BOO}}$ the MISE computed only on the samples that do not belong to $D_{(b)}$.

$$\widehat{\mathrm{MISE}}^{0}_{\mathrm{BOO}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B_i}\sum_{b\in C_i}(y_i - h(x_i, \alpha_{(b)}))^2$$

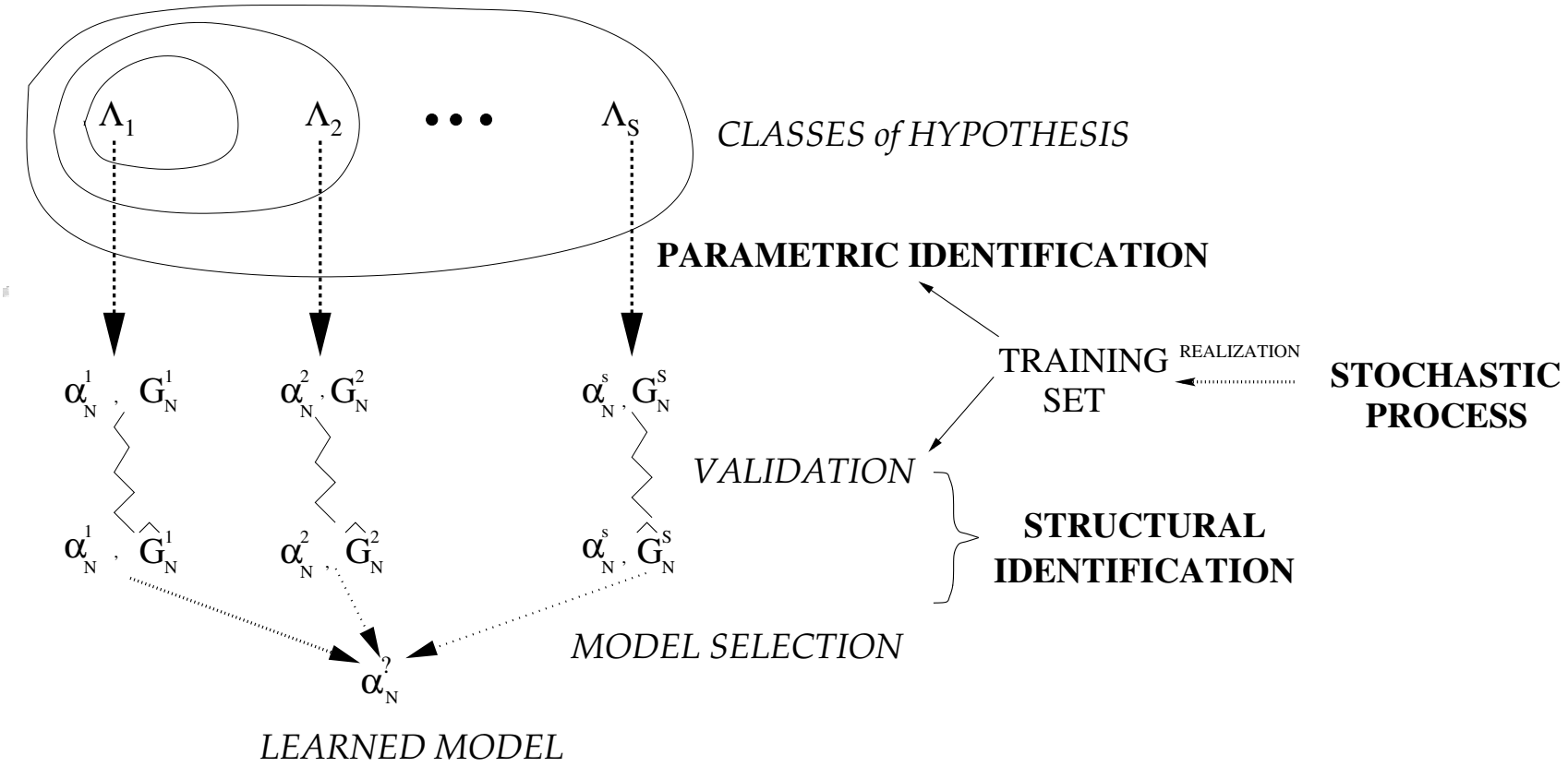
where C_i is the set of indices of the bootstrap samples $D_{(b)}$ that do NOT contain i and B_i is the number of such bootstrap samples $D_{(b)}$.

The .632 estimator

- However, it can be shown that the samples used to compute $\widehat{\text{MISE}}^{0}_{\text{BOO}}$ are particularly hard test cases (too far from the training set) and that consequently $\widehat{\text{MISE}}^{0}_{\text{BOO}}$ is a pessimistic estimate of MISE.
- On the other side, the test cases used in $\widehat{\text{MSE}}_{\text{emp}}$ are too easy (too close to the test set) and consequently $\widehat{\text{MSE}}_{\text{emp}}$ is an optimistic estimate of MISE.
- A more reliable estimator is provided by the weighted average of the two quantities

$$\widehat{\text{MISE}}^{.632}_{\text{BOO}} = 0.368 * \widehat{\text{MSE}}_{\text{emp}} + 0.632 * \widehat{\text{MISE}}^{0}_{\text{BOO}}$$

Supervised learning procedure



The supervised learning procedure

1. A nested sequence of classes of hypotheses

$$\Lambda_1 \subseteq \cdots \subseteq \Lambda_s \subseteq \ldots \Lambda_S$$

   is defined so that $\Lambda^* = \cup_{s=1}^{S} \Lambda_s$.

2. An hypothesis α_N^s, $s = 1, \ldots, S$, is selected by minimizing the empirical risk (*parametric identification*).

3. A *validation* procedure returns $\hat{G}_N^s$ which estimates the generalization error G_N^s of the hypothesis α_N^s.

4. The hypothesis $\alpha_N^{\bar{s}}$ with $\bar{s} = \arg\min_s \hat{G}_N^s$ is returned as the final outcome (*model selection*).