Machine Learning

TSIA-SD 210 - P3

Lecture 2 - 2. Methodology

Florence d'Alché-Buc

Contact: florence.dalche@telecom-paristech.fr,
2A Filière SD, Télécom ParisTech,Université of Paris-Saclay, France

Table of contents

Outline

Statistical learning: a methodology

- Three main problems to be solved :
  - **Representation problem**: determine in which representation space the data will be encoded and determine which family of mathematical functions will be used
  - **Optimization problem (focus of the course)**: formulate the learning problem as an optimization problem, develop an optimization algorithm
  - **Evaluation problem**: provide a performance estimate

Statistical learning for supervised classification

Two main family of approaches:

1. Discriminant approaches : just find a classifier which does not estimate the Bayes classifier
2. Generative probabilistic approaches that are built to model $h(x) = \hat{P}(Y = 1|x)$ using $\hat{p}(x|Y = 1)$, $\hat{p}(x|Y = -1)$ and prior probabilities.

Outline

Model selection or model evaluation

Evaluation metrics

- Define an evaluation metric for performance measure: for instance classification error, Area under the ROC Curve end so on...
- It may happen that the evaluation metric is different from the loss function: why ? the target loss is not convex, you use a surrogate loss

Model Selection:

- Choose the "complexity" of the model you want to use for learning
- How ? Selection of the hyperparameter value on validation set (never on test set)

Model Evaluation:

- Once the hyperparameter is chosen, the learning algorithm is applied, we get a classifier
- This classifier must be evaluated on a test set

First example

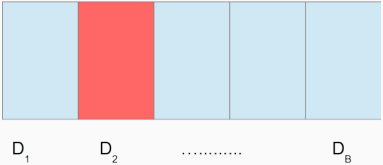
A linear classifier in 2D space:

$$h(x) = \text{signe}(\beta_1 x_1 + \beta_2 x_2 + \beta_0) \tag{1}$$

Learning h_β by minimizing $\sum_{i=1}^{n} L(y_i, h(x_i)) + \lambda \left\|\beta\right\|_2^2$
OR $\sum_{i=1}^{n} L(y_i, h(x_i)) + \lambda \left\|\beta\right\|_1$

Cross-validation 1

Nota Bene : $\mathcal{S}$ dataset devoted to the model construction: either training dataset itself or a validation set
In any case, NEVER on the TEST set !



Divide $\mathcal{S}$ into B folds of the (approximatively) B equal folds $S_{b=1}, \ldots, D_{b=B}$ avec $|D_b| = n/B$. Data are uniformly drawn without replacement.

Cross-validation

1. For a given λ
2. For $b \in \{1, \ldots, B\}$:
   - Train the model parameterized by λ on all the data except **sauf** D_b to get an estimator $\hat{h}^b_{\lambda,n}$
   - Compute **on the remaining data** i.e. D_b (test) the empirical risk with the chosen evaluation metric

$$R_{n,b}(\lambda) = \frac{1}{n/B} \sum_{j \in D_b} L(x_j, y_j, \hat{h}^b_{\lambda,n})$$

3. Compute the Cross-validation risk associated to the model with λvalue

$$R^B_{n,CV}(\lambda) = \frac{1}{B} \sum_{b=1}^{k} R_{n,b}(\lambda)$$

How to find the value of λ ?

Repeat this procedure on all $\lambda \in \Lambda$ taken on a discrete grid and eventually choose:

$$\hat{\lambda}_{n,B} = \arg\min_{\lambda \in \Lambda} R_{n,CV}(\lambda). \tag{2}$$

Selection: training error versus test error

First do model selection either on a specific validation set or on the training set itself Then, learn from $\mathcal{S}_{train}$ using $\tilde{\lambda}$, selected by CV
Eventually, test on $\mathcal{S}_{test}$
$\mathrm{Err}_{CV,val}$ tells us how much generalization error we can wait for, when learning on a dataste of sier errornous dit à quel type d'erreur en généralisation nous attendre en $n_{val} - n_{val}/B$. $\mathrm{Err}_{\mathcal{S}_{app}}$ tells us to which extent the classifier has succeeded on training data
$\mathrm{Err}_{\mathcal{S}_{test}}$ tells us to which extent the classifier is able to succeed on test data

Evaluation and model selection in scientific papers

Usually, many people take $\mathcal{S}_{val} = \mathcal{S}_{train}$ and do Cross-validation selection on the training set it self

Many splits Train/Test (usually 10): empirical mean and error bars are presented

Outline

Introduction

Model Evaluation and selection

References

References

- Performance evaluation / model selection
  - Chapter 4 and 7 of Elements of statistical learning, Hastie, TIbshirani and Friedman.
  - A.-L. Boulesteix, Ten rules for reducing overoptimistic reporting in methodological computational research, `http://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1004191&type=printable`