Introduction to Supervised Learning

1. What is supervised learning?

2. The ERM principle

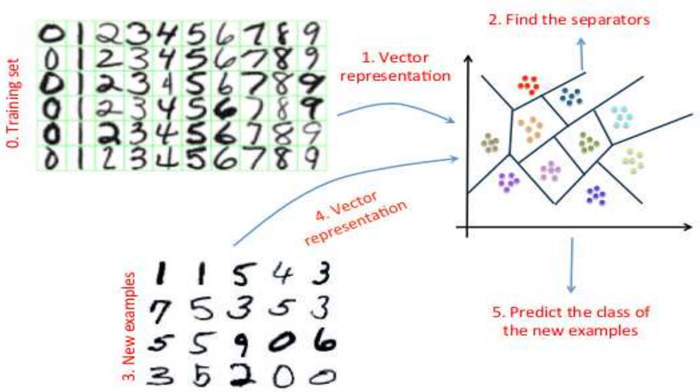
What is supervised learning?

Setting

- Starting point : the data $\mathcal{D} = \{(x_i, y_i), i = 1, \cdots, n\}$
  - $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is the feature vector which describes the object i
  - $y_i \in \mathcal{Y}$ its associated label/response.
- Represent in a relevant way $x_1, \cdots, x_n$?
- Predict the class (classification) or the response (regression) of a new observation in an automatic manner?
- Define a function f which associates to each x_i, its corresponding response $f(x_i) \in \mathcal{Y}$

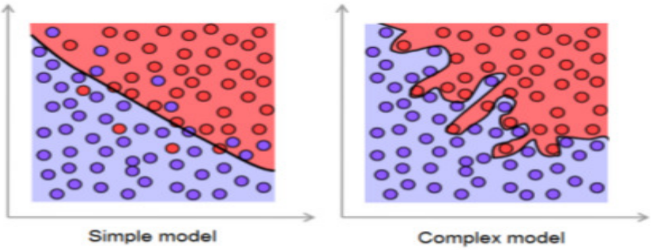
What is supervised learning?

A classical pipeline in ML



What is supervised learning?

- Always possible to define a function f fitting exactly the data
- Not always reasonable!
- Compromise to do between prediction ability of the model and its complexity



What is supervised learning?

Some additional questions

- How can we learn a parametric model f solving the supervised learning problem ?
- Classical models : linear/non linear, white-box/black-box models.
- How shall we evaluate the prediction properties of a given parametric model? The key choice of the loss function/metric

What is supervised learning?

Roadmap

- Give ideas explaining this compromise between prediction ability and complexity
- Present some parametric models to solve a classification/regression problem
- Additional topics : explore features importance, give confidence intervals for the prediction

The ERM principle

The ingredients of a supervised learning problem?

- Dataset $\mathcal{D}_n = \{(x_i, y_i), \imath = 1, \cdots, n\}$. Usually, the (x_i, y_i) are assumed to be i.i.d. realisations of an unknown distribution $\mathbb{P}_{(X,Y)}$
- Hypothesis class $\mathcal{H}$: shape of the classifier/regressor f.
- Loss ℓ : evaluation of the error of f on a data point

Question : how can we learn f?

The ERM principle

- This function f should minimize over $\mathcal{H}$ the risk

$$R(f) := \mathbb{E}[\ell(f(X), Y)]$$

- Since the distribution of the dataset $\mathbb{P}_{(X,Y)}$ is unknown, one cannot estimate this risk
- In practice, the theoretical risk is replaced with the empirical one

$$\widehat{R}(f, \mathcal{D}_n) = \frac{1}{n} \sum_i \ell(f(x_i), y_i)$$

- The Empirical Risk Minimization principle consists in minimizing the empirical risk on $\mathcal{H}$ on $\mathcal{D}_\backslash$ to find f

The ERM principle

- Is this approach theoretically grounded?
- What about practical evaluation?

The ERM principle

The ERM principle in practice

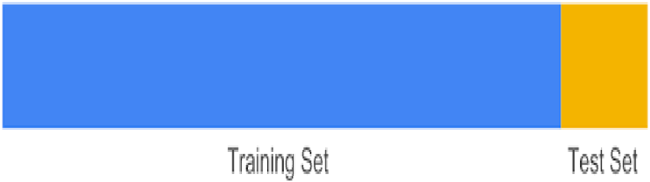
More on model's errors

- In practice, we want to have good predictions on new observations, not included in the initial dataset used to learn f.
- We may then distinguish between
  - training error : measure of how accurately an algorithm is able to predict outcomes values on $\mathcal{D}_\backslash$
  - generalisation error : measure of how accurately an algorithm is able to predict outcome values for previously unseen data
- How can we estimate the generalisation error?

The ERM principle

The ERM principle in practice

Usual evaluation procedure : split $\mathcal{D}_n$ into train and test set and evaluate on each one!



The ERM principle

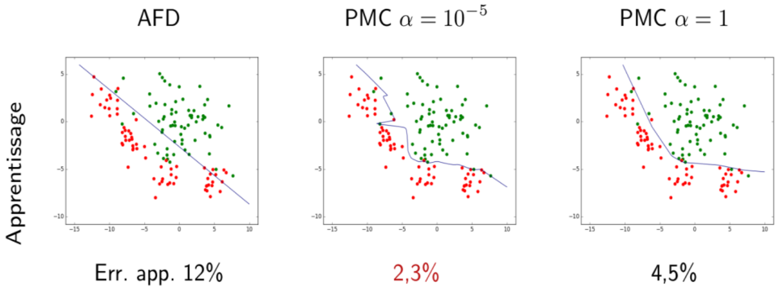
- The error of the model can be tested on the test set = generalisation error
- If the model is too complex, the generalisation error will be too large

The ERM principle

How shall we select the best model?

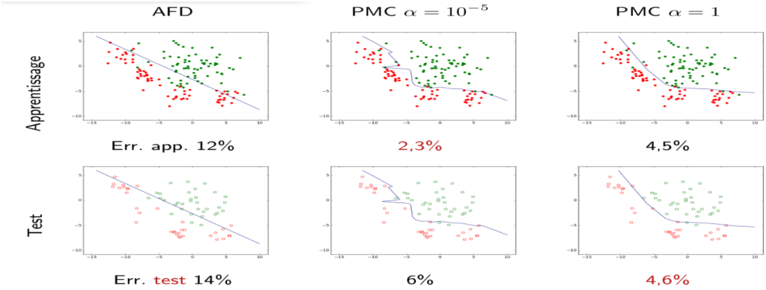
- Training error can be estimated since the data are available
- We assume that the distribution of future data is the same that this of the training set !
- Minimise the training error is it sufficient to minimize the generalisation error?
- Comparison between three models

Training errors



Example extracted from the course of N. Thome (CNAM)

Test errors



Example extracted from the course of N. Thome (CNAM)

Facts

- The model which has the lowest training error does not have the lowest test error
- In whole generality, the test error is larger than the training one
- The difference between these two errors depends on the family of models

Facts

- We cannot measure generalisation error, we estimate it using the test set
- We can also use a theoretical upper bound on the difference between generalisation error and training error of the form: generalisation error $\leq$ training error + bound

Facts

Some difficulties

- If we split the dataset and keep observations for the test we have less data to learn
- This estimation of the generalisation error has a high variance

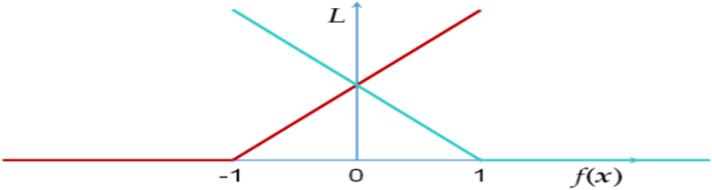
Facts

Alternative approach : cross-validation

More on Generalisation bounds

To simplify, we consider only the classification setting (for all i, $y_i \in \{0, 1\}$)

- Back to assumptions : $\mathcal{D}_n = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \imath = 1, \cdots, n\}$ i.i.d. realisations of $(X, Y) \sim \mathbb{P}_{X,Y}$.
- We are given $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ a bounded loss function.
- Some examples
  - Loss 0–1 : $L(y_1, y_2) = 1_{\{y_1 \neq y_2\}}$.
  - Hinge loss : $L(y_1, y_2) = \max(0, 1 - y_1 y_2)$



  - Quadratic loss $L(y_1, y_2) = (y_1 - y_2)^2$

More on Generalisation bounds

- The classifier f has to minimize the generalisation error

$$R(f) = \mathbb{E}[\ell(f(X), Y)] = \int_{\mathcal{X}\times\mathcal{Y}} L(f(x), y) d\mathbb{P}_{X,Y}(x, y) ,$$

  and $f \in \mathcal{H}$, $\mathcal{H}$ known class of functions.
- This class of function could be parametric

$$\mathcal{F} = \{f_\theta, \theta \in \Theta\} .$$

- Problem $\mathbb{P}_{X,Y}$ is unknown!

More on Generalisation bounds

We replace the generalisation error with

$$\widehat{R}(f, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i)$$

Consistance of ERM principle

The ERM principle is said to be consistent if

$$\widehat{R}(f_n, \mathcal{D}_n) - R(f_n) \overset{(p)}{\to} 0$$

and

$$\widehat{R}(f_n, \mathcal{D}_n) \overset{(p)}{\to} \inf_{g \in \mathcal{F}} R(g)$$

More on Generalisation bounds

Theorem (Vapnik, 1981)

The ERM principle is consistent iff for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{g\in\mathcal{F}} \left|R(g) - \widehat{R}(g, \mathcal{D}_n)\right| > \varepsilon\right) \to 0 \; .$$

(convergence in probability).

More on Generalisation bounds

A special case

If the hypothesis class $\mathcal{H}$ is finite, for any $g \in \mathcal{H}$

- by definition

$$\widehat{R}(g, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(X_i), Y_i)$$

- the r.v. $Z_i = \ell(g(X_i), Y_i)$ are i.i.d., integrables with expectation $R(g)$

we can use the weak law of large numbers.

In whole generality much more complicated :

Vapnik, V. N., & Chervonenkis, A. Y. (1982). Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theory of Probability Its Applications, 26(3), 532-553.

More on Generalisation bounds

- Beyond consistency of ERM principle?
- Generalisation bounds of the form : with probability greater than $1 - \delta$

$$\forall f \in \mathcal{F},\ R(f) \leq \widehat{R}(f, \mathcal{D}_n) + v_n\ .$$

- v_n : depends on δ, n and on the complexity of class $\mathcal{H}$
- The relationship between δ and v_n yields the convergence rate.

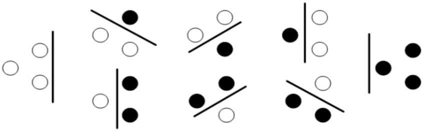
More on Generalisation bounds

Several notions of complexity

- Vapnik-Chervonenkis complexity
- Rademacher complexity

Vapnik-Chervonenkis (VC) dimension

- The class $\mathcal{H}$ shatters $\mathcal{D}_n = \{(x_i, y_i),\ i = 1, \cdots, n\}$ if for all assignments of labels to $x_1, \cdots, x_n$, there exists $f \in \mathcal{H}$ makes no errors when evaluating that set of data points



Shattering of 3 points by the family of linear classifiers

Vapnik-Chervonenkis (VC) dimension

VC-dimension

- Let $\mathcal{E}(\mathcal{H}, \mathcal{D}_n) = \{(x_1, f(x_1)), \cdots, (x_n, f(x_n)), f \in \mathcal{H}\}$ and $C(\mathcal{H}, n) = \max_{|\mathcal{D}_n|=n} |\mathcal{E}(\mathcal{F}, \mathcal{D}_n)|$.
- If $\mathcal{H}$ is a class of functions from $\mathcal{X}$ onto $\{-1, 1\}$ one defines the VC dimension of $\mathcal{H}$ as

$$\mathcal{V} = \max\{v,\ C(\mathcal{H}, v) = 2^v\}\ .$$

Example : for the linear classifier the VC dimension is 3.

An example of generalisation bound

Proposition (Vapnik 1981)

Let $\delta \in (0, 1)$ and $\mathcal{H}$ a class of functions with finite VC dimension $\mathcal{V}$. With probability greater than $1 - \delta$

$$\forall f \in \mathcal{F},\ R(f) \leq \widehat{R}(f, \mathcal{D}_n) + \sqrt{\frac{8\mathcal{V}\ln(2en/\mathcal{V}) + 8\ln(4/\delta)}{n}}\,.$$

Here one has

$$v_n = \sqrt{\frac{8\mathcal{V}\ln(2en/\mathcal{V}) + 8\ln(4/\delta)}{n}}$$

fast rate of convergence!

Comments

- One has

$$R(f) \leq \widehat{R}(f, \mathcal{D}_n) + \text{ generalisation error}$$

- $\widehat{R}(f, \mathcal{D}_n)$ is the error of f on the test set
- The more $\mathcal{H}$ is a complex family, the more the generalisation error is large
- Existence of other complexity measures as Rademacher complexity that can be estimated on data

Comments