Machine Learning

Marc Sebban & Amaury Habrard
LaHC

Contents

1 What is Machine Learning? 2

2 Supervised learning problem 2

2.1 Notations . . . 2
2.2 Curse of dimensionality - Overfitting - Underfitting . . . 3
2.3 Regularized Risk Minimization . . . 4
2.4 Bias/Variance trade-of . . . 4
2.5 Statistical learning theory . . . 4

1 What is Machine Learning?

Machine Learning aims at knowing how to make algorithms that can *learn* from data. They are divided in two category:

- Supervised learning, subdivided into
  - Classification: predict a yes/no answer
  - Regression: predict a continuous value, such as the price of a house
  - Ranking: output the "most relevant" data

  The aim is to predict fro labelled data

- Unsupervised learning, subdivided into
  - Clustering
  - Dimensionality Reduction

  The aim is to find the underlying structure of unlabelled data

Possible Applications Computer Vision, Robotics, Speech Recognition, Artificial Intelligence

Required Skills

- Convex Optimization
- Algorithm: Asymptotic behaviour

We will mainly use SVM (Support Vector Machine), that deals with classification problems. They use the *kernel trick*, which is projection of the data on a high-dimensional space (potentially infinite) where the data becomes linearly separable.

2 Supervised learning problem

2.1 Notations

- Let $S = \{z_i = (\mathbf{x_i}, y_i)\}_{i=1}^m$ be a set of m training examples i.i.d. from an unknown joint distribution $\mathcal{D}_\mathcal{Z}$ over a space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- The $\mathbf{x_i}$ values ($\mathbf{x_i} \in \mathcal{X}$) are typically vectors in $\mathbb{R}^d$ whose components are usually called *features.*
- The y values ($y \in \mathcal{Y}$) are drawn from a discrete set of *classes/labels* (typically $\mathcal{Y} = \{-1, +1\}$ in *binary classification*) or are continuous values (*regression*)
- We assume that there exists a *target function* f such that $y = f(\mathbf{x})$, $\forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$.

Definition 1. *A supervised learning algorithm L automatically outputs from S a model or a classifier (or a hypothesis) $h \in \mathcal{H}$ as close to f as possible.*

2.2 Curse of dimensionality - Overfitting - Underfitting

The number of training example is very important! Sadly, as the number of features or dimension grows, the amount of data (i.e. examples necessary to learn) grows exponentially: it is the *curse of dimensionality.*

To avoid this problem, we can:

- pre-process the data into a lower dimensional space
- regularize the underlying optimization problem at running time

This issue is very closed to *overfitting.*

Definition 2 (Overfitting)**.** *In statistics,* overfitting *occurs when a model is excessively complex, such as having too many degrees of freedom (e.g. polynomial of high order) with respect to the amount of data available* $\rightarrow$ *use a* regularization*.*

Definition 3 (Underfitting)**.** Underfitting *occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data.*

To pick the best hypothesis h^*, we need a criterion to assess the quality of h. Given a non-negative loss function $\updownarrow : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}^+$ measuring the degree of agreement between $h(\mathbf{x})$ and y, we can define the *tree risk.*

Definition 4 (True Risk)**.** *The true risk $\mathcal{R}^\ell(h)$ (also called* generalization error*) of a hypothesis h with respect to a loss function ℓ corresponds to the expected loss suffered by h over the distribution $\mathcal{D}_\mathcal{Z}$.*

$$\mathcal{R}^\ell(h) = \mathbb{E}_{\mathcal{Z} \sim \mathcal{D}_\mathcal{Z}} \ell(h, z)$$

Unfortunately, $\mathcal{R}^\ell(h)$ cannot be computed as $\mathcal{D}_\mathcal{Z}$ is unknown, so we try to minimise the *empirical risk* $\hat{\mathcal{R}}^\ell$, a statistical measure of the true risk over S.

Definition 5 (Empirical Risk)**.** *Let $S = \{z_i = (\mathbf{x_i}, y_i))\}_{i=1}^m$ be a training sample. The empirical risk $\hat{\mathcal{R}}^\ell$ (also called empirical error) of a hypothesis $h \in \mathcal{H}$ with respect to a loss function ‘ corresponds to the expected loss suffered by h on S.*

$$\hat{\mathcal{R}}_\ell(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Definition 6 (0/1 loss)**.** *The most natural loss function for binary classification is the 0/1 loss (also called classification error)*

$$\ell_{0/1}(h, z) = 1 \qquad \textit{if } yh(x) < 0 \textit{ and 0 otherwise}$$

$\mathcal{R}^{\ell_{0/1}}$ then corresponds to the proportion of correct predictions.

Warning Due to the non convexity and non differentiability of the 0/1 loss, minimizing the empirical risk is NP-hard. For this reason, we use surrogate loss functions such that:

- *Hinge loss* (used in SVM): $\ell_{hinge}(h, z) = \max(0, 1 - yh(x))$
- *Exponential loss* (used in boosting): $\ell_{exp}(h, z) = e^{-yh(x)}$
- *Logistic loss* (used in logistic regression): $\ell_{log}(h, z) = \ln(1 + e^{-yh(x)})$

2.3 Regularized Risk Minimization

To prevent the algorithm from overfitting, a supervised learning problem often take the following regularized form:

$$\min_{h\in\mathcal{H}} \hat{\mathcal{R}}^{\ell}(h) + \lambda ||h||_p$$

Where λ is a constant penalizing "too complex" models, and $||.||_p$ a ℓ_p-norm over the classifier h.

Definition 7 (ℓ_p-norm)**.** *If θ is a d-dimensional vector:*

$$||\theta||_p = \left(\sum_{i=1}^{d} |\theta_i|^p\right)^{\frac{1}{p}}$$

The ℓ_2-norm is used to reduce the risk of overfitting (it decreases the large values of the model), and the ℓ_1 also allows the induction of sparse models - i.e. with less features (example: LASSO or ℓ_1-SVM).

Remark Increasing θ with the ℓ_1-norm causes more and more of the parameters θ_j to be driven to zero. The gradient on the ℓ_1-norm is constant w.r.t. the magnitude of each vector component.

Downside The l_1-norm is not differentiable.

2.4 Bias/Variance trade-of

There are three sources of error between $h \in \mathcal{H}$ and the target function $f \in \mathcal{F}$:

1. The inductive bias: nothing guarantees the equality between the target concept space $\mathcal{F}$ and the selected class of hypotheses $\mathcal{H}$, even if the learner is able to provide an optimal hypothesis h^* from $\mathcal{H}$.
2. The variance: since the training set S is finite and randomly drawn from $\mathcal{D}_{\mathcal{Z}}$, the learner usually does not provide the optimal hypothesis h^*.
3. The presence of noise: some training examples can be mislabelled. The learner receives a training set of a "noisy" function $f_b = f + \varepsilon$.

The Bias/Variance trade-off comes from the Mean Square Error (MSE), in statistics:

Definition 8 (MSE)**.** *Let θ a theoretical parameter ($\mathcal{R}(h)$ in our case) and $\hat{\theta}$ an estimate of θ ($\hat{\mathcal{R}}(h)$ in our case). Let $B = \mathbb{E}(\theta) - \theta$ be the bias of $\hat{\theta}$ w.r.t. θ. The MSE assesses the quality of θ in terms of its variation and unbiasedness. It is the expected value of the square loss between $\hat{\theta}$ and θ.*

$$\begin{aligned} MSE &= \mathbb{E}_z[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}_z[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2] \\ &= \mathbb{E}_z[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + B)^2] \\ &= \mathbb{V}(\hat{\theta}) + B^2 \end{aligned}$$

2.5 Statistical learning theory

Definition 9 (Empirical Risk Minimization)**.** *The ERM principle rests on the fact that if h works well on the training set S it might also work well on new examples.*

Definition 10 (Probably Approximately Correct (PAC) Condition)**.** *[Valiant 1984] The ERM principle is valid if the true risk of the hypothesis $h \in \mathcal{H}$ induced from S is closed to the true risk of the optimal hypothesis $h^* \in \mathcal{H}$*

$$h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i)$$
$$h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i)$$

Condition of validity of the ERM principle:

$$\forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 1, \forall \delta \leq 1, \mathbb{P}(|\mathcal{R}(h) = \mathcal{R}(h^*)| \geq \gamma) \leq \delta$$

Definition 11 (Bayesian error)**.** *The bayesian error ϵ^* is the lowest possible error rate (or irreducible error) for any hypothesis h.*

$$\epsilon^8 = \int_{x \in R_i \ s.t. \ y \neq C_i} \mathbb{P}(C_i|x)\mathbb{P}(x)dx$$

where x is an instance, y its corresponding label, R_i is the area/region that a classifier function h classifies as C_i.

Remark In many application, $\epsilon^* > 0$, and as S is finite, selecting the optimal h does not imply getting the optimal hypothesis h^*.