Ecological Modelling 240 (2012) 113–122

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Review

A review of supervised machine learning algorithms and their applications to ecological data

C. Crisci [a], B. Ghattas [b,*], G. Perera [c]

[a] UMR, 6540, CNRS, Université de la Méditerranée, DIMAR, Centre d'Océanologie de Marseille, Station Marine d'Endoume, Chemin de la Batterie des Lions, Marseille 13007, France
[b] Université de la Méditerranée, Département de Mathématiques, Case 901, 163 avenue de Luminy, Marseille 13009, France
[c] Facultad de Ingenieria de la Universidad de la República, Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 11 July 2011
Received in revised form 8 January 2012
Accepted 1 March 2012
Available online 16 June 2012

Keywords:
Machine learning
Ecological data
Regression analysis
Classification rules
Prediction
Mass mortality events
Coastal rocky benthic communities
Positive thermal anomalies

ABSTRACT

In this paper we present a general overview of several supervised machine learning (ML) algorithms and illustrate their use for the prediction of mass mortality events in the coastal rocky benthic communities of the NW Mediterranean Sea. In the first part of the paper we present, in a conceptual way, the general framework of ML and explain the basis of the underlying theory. In the second part we describe some outstanding ML techniques to treat ecological data. In the third part we present our ecological problem and we illustrate exposed ML techniques with our data. Finally, we briefly summarize some extensions of several methods for multi-class output prediction.

Contents

* Corresponding author.
E-mail addresses: carolina.crisci@univmed.fr (C. Crisci), ghattas@univmed.fr (B. Ghattas), gperera@fing.edu.uy (G. Perera).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2012.03.001

1. Introduction

Ecological systems rarely require simple statistical analysis. For example, much of the data collected by ecologists often exhibit *unusual* distributions, non-linearity, multiple missing values, complex data interactions, dependence on the observations, etc. (Fielding, 1999; De'ath, 2007; Guisan et al., 2002; Cutler et al., 2007). The size of datasets is another very hard and frequent problem. Many of the ecological databases are very large and some are continuously expanding. The problems related to a large number of cases or variables are likely to become more severe as more biodiversity data are accumulated and remotely sensored data are increasingly used (Fielding, 1999). Bellman's "curse of dimensionality", often invoked by statisticians to refer to how difficult it may be to deal with a huge amount of data, appears nowadays frequently in Ecology.

Machine learning (ML) techniques are not and will never be the solution to all the problems risen by ecological data. However, these techniques provide a powerful set of tools that deserves a serious attention to deal with some relevant ecological problems. As a first point, let us remark that ML concerns are nothing but the most ancient, classical and widely studied statistical problems: classification, regression, decision, clustering, density estimation, etc. However, what makes ML a particular field is not precisely its goals and problems but its tools, techniques and strategies, characterized by the massive use of algorithms and computational resources to deal with large sets of data, high number of variables and complex data structures.

ML approaches are intensively applied in different areas and there is no doubt that Ecology is today one of the most relevant areas of ML application (Flach, 2001). This is reflected in the large number of publications that appeared in the last years in which diverse ML techniques are applied to solve a wide variety of problems. Studies of the relationship between organisms (species presence/absence, population and community attributes) and habitat characteristics applying ML techniques are well documented in terrestrial (e.g. Ryder and Irwin, 1987; Franklin, 1998; Shan et al., 2006; Cutler et al., 2007), fresh water (e.g. Lek and Guégan, 1999; Džeroski, 2001; Kocev et al., 2010) and marine ecosystems (e.g. De'ath and Fabricius, 2000; Defeo and Gómez, 2005; Merckx et al., 2009; Knudby et al., 2010; Volf et al., 2011). Some particular applications of these techniques in Ecology are the prediction of algal blooms (Ribeiro and Torgo, 2008), fish recruitment (Fernandes et al., 2010), habitat suitability of tree species (Benito Garzón et al., 2006), organism identification (Morris et al., 2001) and determination of factors affecting dispersal of marine species (Pontin et al., 2011).

Previous reviews of ML methods expose in detail a few number of techniques (Recknagel, 2001). Some of them illustrate some ML techniques through case studies but without expanding on theoretical basis (Džeroski, 2001). Further articles present the theoretical basis of some specific technique in a more or less formal manner, illustrating them with ecological examples (Lek and Guégan, 1999; De'ath and Fabricius, 2000; Guisan et al., 2002; De'ath, 2007).

Here we intend to present a comprehensive view of ML techniques giving a brief overview of the theory underlying these approaches. We restrict our study to situations where we wish to model the effect of a set of explanatory variables (X) on a target variable (Y). This is the context of *Supervised Learning* (SL) said otherwise, *Regression* or *Classification* depending on the nature of Y. We have selected eight methods among the large panel of available approaches. Theses methods gave rise since the 1995th to extensive research in the machine learning community as they have numerous advantages among which being non-parametric and free from any distribution assumption. Besides, most of these methods offer a lot of extensions and may be used to model multidimensional outputs or functional outputs. Finally, they may be mathematically unified, as they be expressed as linear or convex combinations of non-parametric functions.

We illustrate these methods with an ecological example, the prediction of mass mortality events in the NW Mediterranean coastal rocky benthic communities.

The paper is organized as follows. In Section 2 the general framework of SL is presented with a glance of its theoretical basis but explained in a conceptual manner. In Section 3 we present a wide panel of SL techniques and in Section 4 we illustrate the exposed techniques with our application. Finally, in Section 5 we summarize some technical extensions useful in particular ecological problems such as prediction of multi-class and functional outputs.

2. Supervised Learning (SL): general framework and fundamentals

The main purpose of SL techniques is to learn how to predict a random variable $Y \in \mathcal{Y}$ based on a set of explicative random variables denoted by $X \in \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ depend on the problem at hand but may be thought to be respectively $\mathbb{R}$ and $\mathbb{R}^d$ for example. We will often call X the *input* and Y the *output.* As a leading example, one may think about a variable Y that represents the presence/absence of a rare lichen species (Cutler et al., 2007), and a set of variables X that consists on elevation, aspect and slope. The main problem is to find a *predictor*:

$$f: \quad \mathcal{X} \to \mathcal{Y}$$

$$X \to f(X)$$

chosen among the set of all functions $\mathcal{F} = \{f: \ \mathcal{X} \to \mathcal{Y}\}$. To build a "good" predictor we have to define a performance criterion that is, a *loss function* denoted L which depends namely on f, X and Y. We thus say that predictor f is better than predictor g if $L(f, X, Y) < L(g, X, Y)$. To simplify the notation we will omit the dependence of L on X and Y.

Suppose that there exists a unique predictor $f^* \in \mathcal{F}$ which minimizes the loss function L, called the *optimal predictor*,

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} L(f, X, Y).$$

In general however, it is not possible to minimize L over the whole set of possible functions $\mathcal{F}$ (that may be a very large set) but only over a given class of predictors C that corresponds to a set of practically computable predictors (for instance the class of linear models). In such case one obtains a predictor f^{**} satisfying

$$f^{**} = \underset{f \in \mathcal{C}}{\operatorname{argmin}} L(f, X, Y).$$

The predictor f^{**} may be different of the globally optimal predictor f^* (which may be not included in C). Besides, the predictor

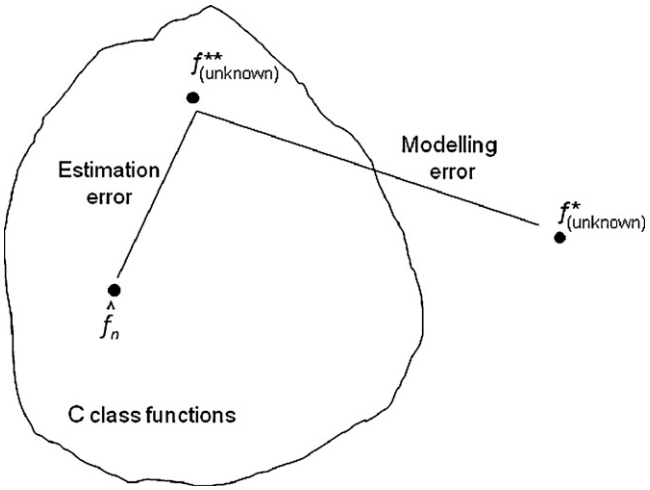


Fig. 1. Representation of best of all possible predictors (f^*), best of all possible predictors inside the chosen class C (f^{**}) and the empirical predictor ($\hat{f}_n$). Adapted from Devroye et al. (1996).

f^{**} is not available neither in practice, since one is not able to minimize the *loss function* on C but only an empirical version of it $\hat{L}_n$ based on a sample of size n. This empirical estimation corresponds to the predictor used in practice. It is denoted by $\hat{f}_n$ in the sequel, and satisfies

$$\hat{f}_n = \underset{f \in \mathcal{C}}{\operatorname{argmin}}\, \hat{L}_n(f, X, Y).$$

The loss of performance due to the difference between f^* and f^{**} is of modelistic nature and depends on how relevant is our choice of C (in other words, which model we choose). If a bad choice of C is done, no further sampling will allow to balance this loss of performance. This is why the difference $L(f^*) - L(f^{**})$ is often called *approximation error* or *modeling error*. On the other hand, a second loss of performance, due to the difference between f^{**} and $\hat{f}_n$ is of purely statistical nature. If a very large sample is available (i.e. if n tends to infinity), $\hat{f}_n$ will converge in some sense to f^{**} under suitable hypothesis on our model. This explains why the difference $L(f^{**}) - L(\hat{f}_n)$ is often called *estimation error* (Fig. 1).

Statistical learning theory aims to give the necessary and sufficient conditions on the class of functions C that would guarantee the consistency of the estimator $\hat{f}_n$, that means its convergence to the optimal predictor f^*.

2.1. The optimal model

The loss function used to estimate the model f depends on the nature of Y. In regression one generally seeks to minimize the quadratic error:

$$L(f) = E[(Y - f(X))^2]$$

where E is the expectation with respect to the joint distribution of (X, Y). In this case the optimal model is the conditional expectation of Y over X

$$f^*(x) = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, L(f) = E(Y|X = x)$$

and the predictor used minimizes the empirical version of the quadratic error,

$$\hat{f}_n = \underset{f \in \mathcal{C}}{\operatorname{argmin}}\, \hat{L}_n(f) = \underset{f \in \mathcal{C}}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(X_i)]^2$$

where $(X_i, Y_i)_{i=1\ldots,n}$ are n independent and identically distributed (*iid*) realizations of (X, Y).

In classification ($Y \in \{1, \ldots, J\}$) the classical loss function is the misclassification error:

$$L(f) = P[Y \neq f(X)]$$

The optimal model is based on Bayes rule for classification:

$$f^*(x) = \underset{j \in \{1,\ldots,J\}}{\operatorname{argmax}}\, P[y = j|X = x]$$

where $P[Y = j|X = x]$ is the posterior conditional probability of having level j for Y. Note that in the binary case where $Y \in \{0,1\}$ we have that $P[Y = 1|X = x] = E[Y|X = x]$ and thus the optimal model is given by the conditional expectation just as for regression.

In both cases, the choice of class C consists of choosing the mathematical form for f.

2.2. Estimating the model loss

We will generally build the estimator $\hat{f}_n$ of f from an *iid training sample* $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The performance of $\hat{f}_n$ is estimated by means of a new sample $(X_1^v, Y_1^v), \ldots, (X_m^v, Y_m^v)$ called the *evaluation sample*, which is another *iid* sample, independent from the training sample. This method intends to prevent from "overfitting". Indeed, $\hat{f}_n$ has a low error over the learning sample, while its performance may be poor when applied to new data.

Generally, data are given without any precision about the evaluation sample. In practice we randomly split the data at hand into two subsamples (the learning sample and the evaluation or test sample). The model is estimated over the training set and its performance is studied using the evaluation sample. To reduce any bias due to the random choice of the evaluation sample, the loss is averaged over several random splits of the data (see Fig. 2).

Other types of performance estimations, based on well-known procedures such as *cross-validation*, leave one out, bootstrap and other resampling techniques, may be used in practice to give unbiased and numerically efficient estimations. We refer to Hastie et al. (2003) for readers interested in going deeper into the theoretical basis of ML.

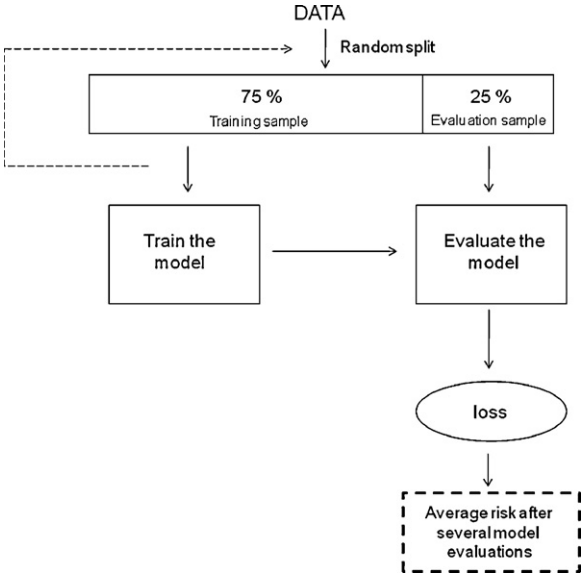


Fig. 2. Random data splits for model training and evaluation. To avoid bias estimation due to one random split, several splits are performed and the loss is averaged over the different random test samples.

3. Some particular machine learning tools with ecological examples

In what follows we present a short description of some well-known and widely used learning algorithms, with references to ecological examples for some of them. Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ and $Y \in \mathcal{Y}$. In case of regression $\mathcal{Y} = \mathbb{R}$ and in case of classification $\mathcal{Y} = \{1, \ldots, J\}$, $J \in \mathbb{N}$.

Suppose we have at hand an *iid* sample $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathcal{Y}$. We denote $x = (x_{\bullet 1}, \ldots, x_{\bullet d}) \in \mathbb{R}^d$ a generic realization of X. We present different family of predictors f. The presented methods are all non-parametric, that is, the estimated function cannot be described using a finite number of parameters.

3.1. Generalized additive models

A relevant statistical development of the last thirty years is the advance in regression analysis provided by generalized linear models (GLMs, Nelder and Wedderburn, 1972) and generalized additive models (GAMs, Hastie and Tibshirani, 1990). These methods provide powerful tools in many scientific research fields among whose ecological research is a good example (Flach, 2001; Guisan et al., 2002). The traditional linear model, where

$$f(x) = E(Y|X = x) = \alpha_0 + \sum_{k=1}^{d} \alpha_k x_{\bullet k}$$

often fails in explaining ecological data, because in real life, effects are seldom linear. Non-linear regression models are more suitable and GAMs are among the most practical of them.

GAMs may be seen as an extension of GLMs which are the analog of linear models for discrete outputs $Y \in \mathcal{Y} = \{1, \ldots, J\}$. If Y is binary $Y \in \{0, 1\}$ and $p(x) = P[Y = 1|X = x]$ the GLM has the form:

$$g(p(x)) = \alpha_0 + \sum_{k=1}^{d} \alpha_k x_{\bullet k}$$

where g is a link function which belongs to an exponential family, the most common being the logit function $g(p) = \log(p/1 - p)$. For the general case where Y has J levels, we estimate one such model per level and $p(x)$ in the left hand side is replaced by $p_j(x) = P[Y = j|X = x]$.

Additive models generalize the LM by taking the form

$$f(x) = E(Y|X = x) = \alpha_0 + \sum_{k=1}^{d} fk(x_{\bullet k})$$

where f_j are typically non-parametric univariate smooth models such as *Kernel predictors*.[1]

Finally, GAMs are the analogous to additive models for discrete outputs:

$$g(p(x)) = \alpha_0 + \sum_{k=1}^{d} f_k(x_{\bullet k})$$

The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and the set of explanatory variables (Guisan et al., 2002). The d non-parametric functions are simultaneously estimated using efficient algorithms like the *backfitting*. Note finally that each component f_k may have different form (kernel, splines, etc.).

3.2. Classification and regression trees

Classification and regression trees (CART) is the most popular of the tree based methods introduced by Breiman et al. (1984). Several extensions of the original method have been proposed (see Loh and Vanichsetakul, 1988; Chaudhuri et al., 1995). The basic idea behind CART is to split the space of explicative variables into multidimensional rectangles, and to use on each of them a very simple local predictor (a constant on each rectangle for instance).

The development of a tree structure comes from the recursive partitioning of the original data set in two subsets that are more *homogeneous* than the first one, leading to a branching structure. The underlying idea is that, as the ramification increases, the *homogeneity* in each node increases too.

The split of the current data is done via a binary rule over the explanatory variables. For continuous variables the split has the form $x < s$ where s is a threshold over the variable x. When x is discrete, the binary rule takes the form $x \in \mathcal{L}$, where $\mathcal{L}$ is a subset of the possible levels of x. At each step of the partitioning an exhaustive search among all the splits is done. The chosen split is the one that generates the most homogeneous subsamples. Homogeneity is measured with respect to the target variable Y. When Y is continuous the criterion is the *deviance*, and we obtain a *regression tree*. When Y is discrete we deal with *classification trees* and the most used criteria are Entropy and Gini index. In each new subsample (*leaf*), a value of Y is assigned. For regression trees, it is the mean value of Y over the leaf observations. For classification trees the most frequent level of Y in the leaf is used.

Let us now pay attention to the question of when should we stop the ramification. If the number of iterations is small, there will be a lot of data in each terminal node, resulting in low statistical error, but high modeling error. If the number of terminal nodes is large, the contrary happens. Therefore, the optimal number of terminal nodes is an intermediate one (Bell, 1996), and can be obtained by adding to the homogeneity measure of nodes a term that linearly depends on the number of leaves, penalizing the complexity of the tree. This procedure is known as "pruning" and leads to very efficient results.

Finally, let us remark some of the advantages of CART:

- The obtained results are very easy to understand.
- CART is robust to the effects of outliers in the output. Such observations are often isolated into nodes where they do not affect the rest of the tree.
- Trees can handle mixed variable types and missing values (Bell, 1996).
- CART gives an importance index for each explanatory variable introduced in the model.
- CART can deal with data sets with complex structures and they are powerful compared with alternative methods as the set complexity increases.
- CART may also be used in the context of functional data analysis, when X is a time series or a signal.

This method has been used in a high number of ecological studies, e.g. species–habitat associations (Bell, 1996; De'ath and Fabricius, 2000; Ryder and Irwin, 1987), species distributions (Huttmann and Lock, 1997; Ribic and Ainley, 1997), response of species to human impacts (Grubb and Bowerman, 1997).

[1] A Kernel predictor assigns to each value of the input variable x a weighted mean of the output variable over all the observations in the sample. Weights are computed using a kernel function which gives low weights for observations which are far from x and higher weights for observations close to x.

3.3. Ensemble classifiers

A weak point of CART is its *unstability* with regard to changes in the training sample. A slight perturbation of the learning set may induce important changes in the model structure. Ensemble learning or aggregation techniques arose as a solution to this problem. The underlying idea is to consider a linear combination of several models $f_1, \ldots, f_k, f_1, \ldots, f_k$,

$$f^{(a)}(x) = \sum_{k=1}^{K} \alpha_k f_k(x)$$

where α_k are real coefficients. If $\sum_{k=1}^{K} \alpha_k = 1, f^{(a)}$ is a convex combination. When the output variable Y is discrete $Y \in \{1, \ldots, J\}$ the aggregation of $f_1, \ldots, f_k$ is generally done using weighted majority vote.

Aggregation is generally efficient when the models $\{f_k\}_{k=1,\ldots,K}$ are unstable like for instance regression and classification trees. The aggregated model is more stable than any model taken from the ensemble and has generally a lower loss. In what follows we will describe some of these algorithms.

3.3.1. Bagging

Bagging (Breiman, 1996a, 1996b) is based on a well-known statistical method called Bootstrap. For a given data set of size n, a bootstrap sample is obtained by resampling with replacement n observations of this dataset (Davison and Hinkley, 1997; Efron and Tibshirani, 1994). In Bagging, we generate M bootstrap samples of the training set and perform a CART model on each of them. The final model is obtained by combining the M models. In regression, the bagging predictor is defined as the mean of the M models. In classification, it is defined as the simple majority vote over the M models.

Bagging systematically outperforms simple CART models and the gain is generally very significant in regression.

3.3.2. Random forests

The last and very remarkable contribution of Leo Breiman to the development of ML is the random forest technique, based on the use of a large series of low-dimensional regression trees. Its theoretical development is presented in Breiman (2001). Let us briefly explain the main features of random forests in the case of classification.

Random forest performs several classification trees. For a new observation the prediction given by a random forest is a majority vote over the predictions given by the trees in the forest. The trees in the forest are performed as follows:

- If the size of the training sample is n, a bootstrap sample of size n will be used to build each tree. This is done in the same way as for bagging.
- No pruning is used: each tree is developed to its largest extent.
- At each node of a tree, the best split is selected among all the splits on only $m \ll d$ randomly chosen variables. This avoids the curse of dimensionality and highly increases the speed of the algorithm.

Random forest presents several noticeable properties: it is probably one of the most efficient learning algorithms in terms of prediction accuracy and it runs fast and efficiently over very large data bases. Besides, it offers an intuitive approach to assess the importance of each explanatory variable used in the model.

A detailed example of random forests' application to an ecological problem may be found in Cutler et al. (2007).

3.3.3. Boosting

Boosting is an iterative technique allowing to obtain powerful algorithms based on weak predictors or classifiers (see Freund, 1995). More precisely, one starts with a classifier whose accuracy (probability of success) is just slightly over 0.5, and implement an iterative procedure that requires more efficiency where the previous predictor had the worst performance. In the long run, one will obtain an efficient predictor. There are many variants of this method; the first Boosting algorithm, known as AdaBoost is very clearly exposed in Hastie et al. (2003). AdaBoost was designed for the binary classification problem $Y \in \{0, 1\}$. There is a large literature and an extensive work to adapt this algorithm for the regression and to extend it to the multi-class case $Y \in \{1, \ldots, J\}$. Among the latest works about multi-class extension of boosting algorithm Zhu et al. (2006) gives a very direct and simple algorithm. For an ecological application of this method see De'ath (2007).

3.4. Support Vector Machines

Support Vector Machines (SVM) provide another method of data classification and regression. We will present here only the basic idea of the method, in the context of classification.

Assume that we have a training sample with inputs in $\mathbb{R}^d$ and a binary output $Y \in \{-1, 1\}$. Assume that the data are *linearly separable* i.e. there exists at least one hyperplane in $\mathbb{R}^d$ which perfectly separates the two subgroups corresponding to each level of Y. We want to select the hyperplane that best separates the classes and which is the farthest possible from all the cases.

Consider an hyperplane H defined by $f(x) = 0$ with

$$f(x) = \langle w, x \rangle + b$$

where $w \in \mathbb{R}^d$ corresponds to the normal vector to the hyperplane, and $\langle , \rangle$ denotes the dot product in $\mathbb{R}^d$. Note that the sign of $f(x)$ indicates the position of the point x with respect to the hyperplane H in $\mathbb{R}^d$. The *margin* of observation i with respect to H is defined by:

$$\gamma_i = y_i f(x_i)$$

The margin may be seen as the signed distance to H. For an observation i, $\gamma_i > 0$ if and only if that observation is well classified by H. The Hyperplane margin with respect to the sample S is defined by $\gamma_H = \min_{i=1,\ldots,n} \gamma_i$. It corresponds to the margin of the nearest point to H.

SVM aim to find among all the hyperplanes in $\mathbb{R}^d$ the one that maximizes the margin,

$$H^* = \underset{H}{\operatorname{argmax}}\, \gamma_H$$

In the linearly separable case the optimal hyperplane H^* is found using numerical optimization where γ_H is maximized under the constraints that all the observations in the sample are well classified, i.e. $\gamma_i > 0$ for all $i = 1, \ldots, n$. This hyperplane is unique (Fig. 3).

For the non-linearly separable case the data are embedded into a space of higher dimension called the feature space using a *Kernel* function. The underlying idea is that in higher dimension classes may be separated more easily (Fig. 4). The Kernel characterizes the transformations of the data. Typically, one uses a Gaussian Kernel:

$$K(x, y) = \exp^{-\delta ||x-y||^2}$$

where $\delta > 0$.

For a detailed exposition of SVM see Hastie et al. (2003) and Vapnik (1998).

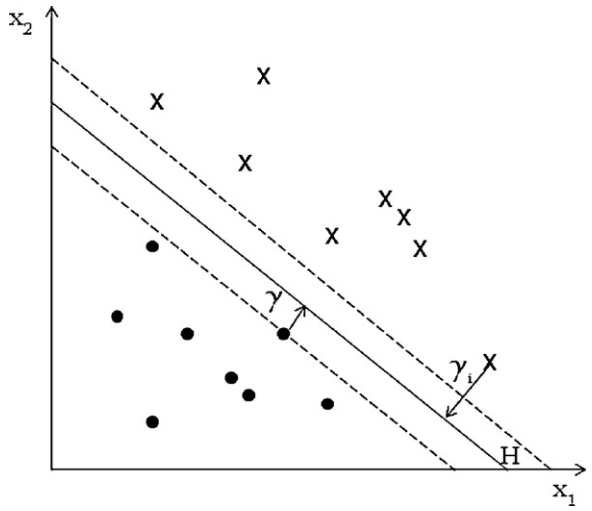


Fig. 3. Two linearly separable classes in $\mathbb{R}^2$. γ_i is the margin of observation i, γ is the margin of the hyperplane with respect to the sample.

3.5. Projection pursuit (in particular, neural networks)

The projection pursuit regression (PPR) model may be written as follows:

$$f(x) = \sum_{m=1}^{M} g_m(\langle w_m, x\rangle)$$

where $g_1, \ldots, g_M$ are unspecified smooth functions, and w_m are d-dimensional vectors.

Observe that $\langle w_m, x\rangle$ corresponds to the projection of x into the direction w_m and the model is specified once the M directions and functions are estimated. This explains the name of the method: one looks for the best directions where the more significant features of the input are revealed, and in that direction, a general non-linear predictor is adjusted (Friedman and Stuetzle, 1981). The function $g_m(\langle w_m, x\rangle)$ is usually called a *ridge function*.

From a practical point of view, PPR models are fitted via an iterative procedure described in Hastie et al. (2003). When instead the functions g_m are taken within a particular class of parametric functions, PPR reduces to neural networks, probably one of the most widely known learning algorithms (Rosenblatt, 1962; Taylor, 2006). Very few achievements exist for a classification version of the projection pursuit approach.

A quite good description of projection pursuit type approach (ANNs) in the context of Ecology may be found in Lek and Guégan (1999).

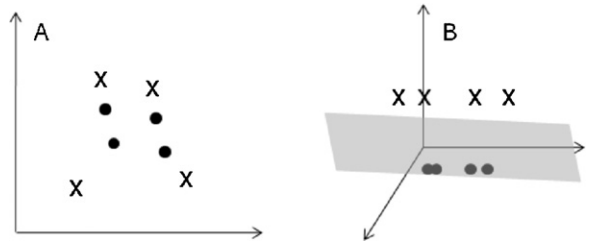


Fig. 4. SVM for binary classification when the two classes are not linearly separable (A). Embedding the data into a three dimensional space using a non-linear transformation, gives rise to a linearly separable configuration in a higher space (B).

3.6. Nearest neighbors

Nearest neighbors approaches are among the simplest and most intuitive methods. To compute the predictor $\hat{f}(x)$ at point x, we have to define a neighborhood $N_k(x)$ corresponding to the set of the k closest observations to x among the learning sample in $\mathbb{R}^d$. For regression, the predictor is the average of the output over the k nearest neighbors,

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} Y_i$$

In classification the mean is replaced by the majority vote.This is clearly a non-smooth method. In binary classification the border line between the two classes is in general a very irregular curve (Hastie et al., 2003). The k-nearest neighbors ($k$nn) is robust against outliers. The only unknown parameter is k.

4. A case study example: modeling mortality events in the NW Mediterranean coastal rocky benthic communities

In this section we study an ecological problem by means of different ML techniques exposed in the previous section. First, we describe the data and the problem that we want to address. Next, we show how the different models are fit to the data using some R packages. Finally, we present the results obtained from each model for the regression case and for classification.

4.1. The data

In the last years, mass mortality events of unprecedented extension and severity were registered in the coastal rocky benthic communities of the North West (NW) Mediterranean Sea during the summer period. Long lived species such as gorgoninas were within the most affected organisms (Cerrano et al., 2000; Perez et al., 2000; Garrabou et al., 2009). These mortality events were related to positive thermal anomalies occurring during summer (Bensoussan et al., 2010; Crisci et al., 2011). Since warming trends were detected for the NW Mediterranean coastal waters (Romano et al., 2010), future mass mortality events related to positive thermal anomalies are likely to occur. For the first time, the availability of biological surveys and hourly temperature records offers the opportunity to study the statistical relation between temperature and mortality.

The main objective here is to give insights in the role of temperature in the mortality events and to generate predictive tools of these events.

Data were collected from four regions of the NW Mediterranean sea:

- Parc Natural del Montgrí, Illes Medes i Baix Ter, L'Estartit, Spain,
- Riou, Marseille, France,
- Parc National de Port-Cros, France,
- Reserve Naturelle de Scandola, Corsica, France.

On one side, biological surveys were carried out at each region for the same period from shallow (≈5 m) to deep waters (≈40 m). Several sites were surveyed after summer inside each of the four study regions, the number of sites varies from 2 to 14 depending on the region and the year. Inside each site, the mortality of a minimum of 70 colonies was registered for 1–3 gorgonian species (*Corallium rubrum*, *Eunicella cavolinii*, *Paramuricea clavata*). These species are important structural and biomass contributors of the coralligenous communities, one of the most diverse communities of the Mediterranean (Ballesteros, 2006). A total of more than 20,000 colonies were surveyed. For each colony, mortality was computed

as the proportion of the colony affected either by recent epibiosis or denuded axis (skeleton).

On the other side, at each region, thermometers were placed from 5 to 40 m depth every 5 m to give hourly measurements. Data are available from 1999 to 2006 at Riou and Port-Cros, but measurements began on 2002 for Illes Medes and on 2004 for Scandola.

The two greatest mortality events were observed during summers 1999 and 2003 and a less severe one was observed during summer 2006. Hence temperature data for each of these events are available on at least two regions.

Data concern only summer period from the 1st July to the 30th September.

The dependent variable, *nec*, is continuous and corresponds to the percentage of affected colonies (colonies affected with more than ten percent of necrosis). This variable showed to be very useful to quantify mortality and to identify interregional patterns during the 2003 mortality event (Garrabou et al., 2009). To illustrate the use of the different models in the context classification we use a discretized version of the dependent variable in two levels and denote it *nec.d*. Level 0 corresponds to observations for which less than 10% of the colonies are affected while the level 1 indicates sites with 10 or higher percent of affected colonies.

The explanatory variables are temperature descriptors and retain information on the magnitude and variability of temperature anomalies over different length periods of time (see Table 1).

In total 102 observations are available, two dependent variables, one continuous and the other discrete, and five explanatory variables.

Table 1
Temperature variables used to fit the models. For a detailed description of these variables see Crisci et al. (2011).

Variable label	Description
Max	Summer maximum *T*
*MeanT*_15	Mean *T* of the 15 summer consecutive days with the highest mean *T*
*CV*15	*CV* of the 15 summer consecutive days with the highest mean *T*
*MeanT*_30	Mean *T* of the 30 summer consecutive days with the highest mean *T*
*CV*30	*CV* of the 30 summer consecutive days with the highest mean *T*

Table 2
R packages and functions used to run models described in Section 3. Default function parameters were considered, optimized by cross validation whenever possible. For boosting, different packages were chosen to treat the regression and the classification problem. The R package for projection pursuit in a supervised classification frame is no more available.

Model	R package	R function
GAM	mgcv	mgcv
CART	rpart	rpart
Random forest	randomForest	randomForest
Bagging	adabag	bagging
Boosting regression	gbm	gbm
Boosting classification	adabag	adaboost.M1
SVM	e1071	svm
Projection pursuit	stats	ppr
Nearest neighbors	class	*k*nn

4.2. Fitting the different models

We performed all models presented in previous section over our ecological example. The models were run using R software (version 2.12.2). The packages necessary for each model and the functions used to build the models are summarized in Table 2.

Two of the used models give interpretable graphical outputs. Fig. 5 shows result from the GAMs for the regression case (for classification, the figure is omitted as it is very similar as for regression). For sake of clarity we omitted the variable *CV*30, as it was the less

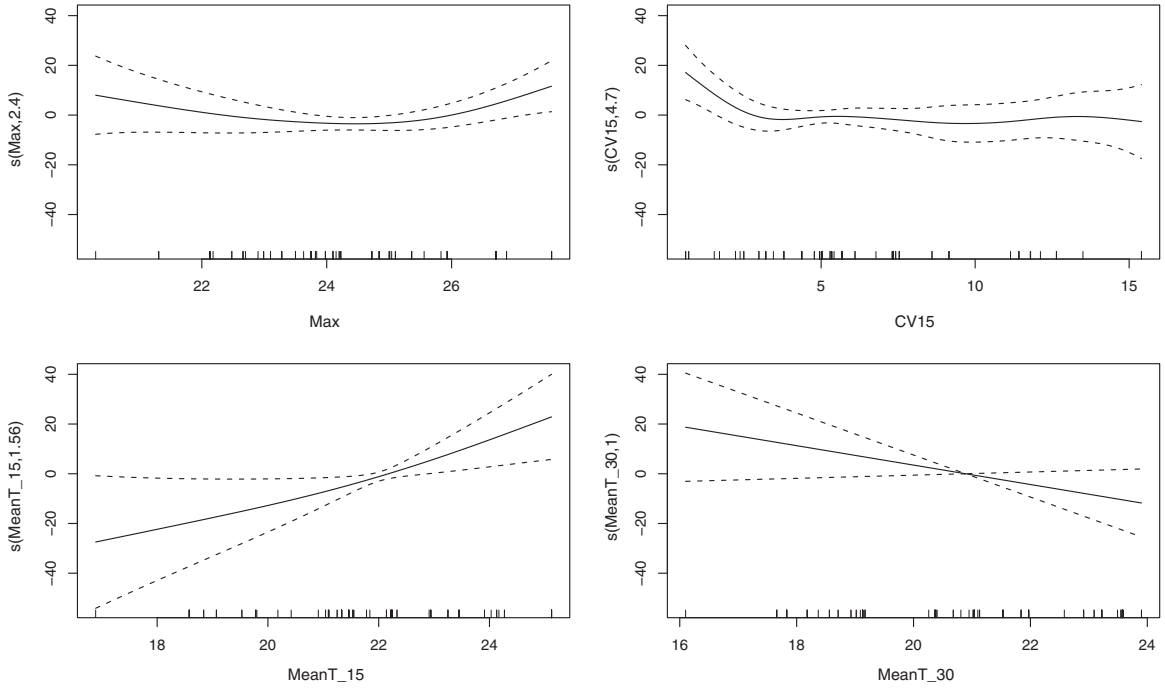


Fig. 5. Graphical output of GAM. Each curve corresponds to the contribution of each explanatory variable to the percentage of affected colonies: maximum summer temperature (*Max*, top left), Mean temperature of the 15 consecutive days with the highest mean T (*MeanT*_15, bottom left), Coefficient of variation of the 15 consecutive days with the highest mean temperature (*CV*15, top right), and mean temperature of the 30 consecutive days with the highest mean temperature (*MeanT*_30, bottom right). Dashed lines represent confidence intervals of each curve. The *y*-axis label indicates the covariate name and the estimated degrees of freedom of the smooth curve.

significant statistically. The *Max* and *CV*15 variables have non-linear relations with the percentage of affected colonies. In the first case, the highest values of mortality were associated with the lowest and highest values of the Max variable (top left panel). In the second case, low values of *CV*15 were associated with the highest values of mortality, but for higher *CV*15 values, mortality remained constant (top right panel). *MeanT_*15 and *MeanT_*30 presented more linear relations (positive or negative) with the dependent variable (bottom left and right panel).

Fig. 6 shows the obtained regression tree (a) and classification tree (b) when using *nec* and *nec.d* as dependent variable respectively. In (a), at each node, the average value of the dependent variable (percentage of affected colonies) and the number of observations is indicated. Each branch is labeled with the corresponding binary rule over one of the explanatory variables. In (b), at every node, the number of cases in each of the two categories is indicated, and also, the label of the node, which corresponds to the most frequent category in the node.

For the case of regression, highest values of mortality were related with high values of the mean temperature of the 30 consecutive days with highest mean temperature (*MeanT_*30) and also with not very high values of *MeanT_*30 but with high Maximum summer temperature (*Max*) values. Within cases with not very high *MeanT_*30 and *Max*, the low variation of the thermal regime in periods of 15 days (*CV*15) seemed to be determinant to produce relative high values of mortality.

4.2.1. Models' performances

Data was randomly split using 3/4 to train and 1/4 to test the models. The models accuracy was averaged over 100 such splittings. Each model is used with the continuous output *nec* as a regression model, and with its discretized version *nec.d* as a classifier. For regression we report the mean squared error averages over the 100 test samples, and for classification we present the mean misclassification error. results are given in Table 3. For the case of projection pursuit only regression models were performed as they are not available for classification.

Random forest, followed by Bagging were the models that presented the lowest error rates both for regression and classification, while Boosting presented the highest error rates for regression and SVM for classification. Although it is difficult to explain the differences among the models' performances, it may be noted that ensemble methods often have higher performances than CART or GAM. SVM has outperforms ensemble methods generally in very high dimensions, with very few observations. *k*nn for regression is very difficult to tune and its results depend on the choice of the distances used to define the neighbors of each observation.

Table 3
Mean squared errors for regression (*nec* variable) and mean misclassification error for classification(*nec.d* variable) using the different models. PPR has been tested only in regression.

Model	*nec*	*nec.d*
GAM	11.96	0.1546
CART	12.06	0.1777
Bagging	11.33	0.1358
Boosting	13.89	0.1346
Random forest	11.20	0.1365
SVM	12.08	0.1958
PPR	11.89	–
*k*nn	14.98	0.1638

5. Extending learning algorithms to specific or general cases

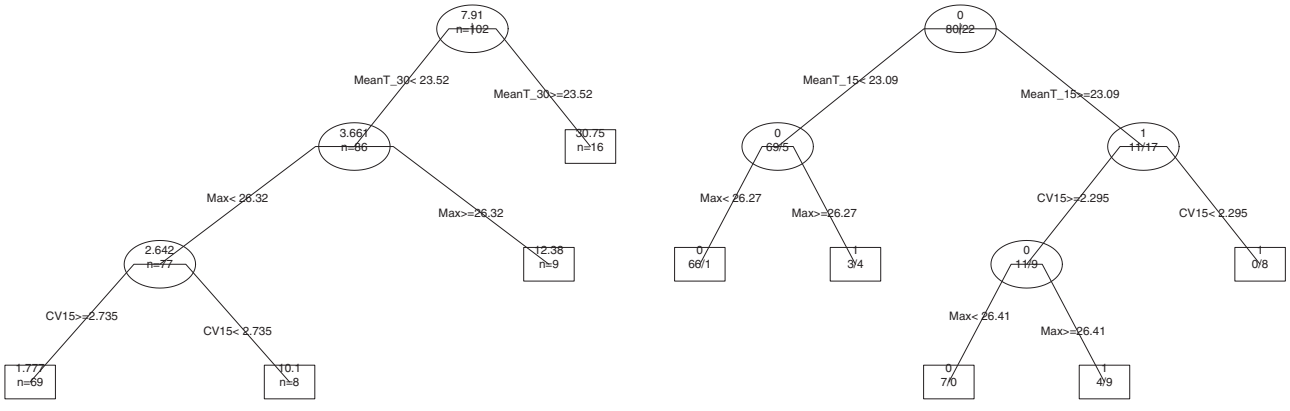
We mention here some of the most challenging extensions for the algorithms described above.

5.1. Multi-class extensions

It may be noted that some learning algorithms are initially designed for specific cases like SVM or boosting whose first versions where created for the binary supervised case, e.g. the dependent variable may have only two possible values. An extensive research has been dedicated to extend these algorithms to the *multi-class* case. For SVM we may cite Phetkaew et al. (2003) and for Boosting we refer to Guruswami and Sahai (1999) and Zhu et al. (2006). Theoretical problems concerning these approaches are still unresolved. The most common approaches generally try to decompose the multi-class problem into several binary problems, resolve each of these simple problems and aggregate their solutions.

5.2. Multivariate discrete or continuous output

We may often be concerned by the prediction of a multivariate output $Y \in \mathbb{R}^q$. For instance, ozone observations may be available simultaneously at q measurement sites. The different sites may share or not the same independent variables. We would like to use a model to predict the future values of Y. One solution would be to adjust one model per site. This approach is in general not reliable especially when the observations among sites are dependent.



(a) Regression tree (b) Classification tree

Fig. 6. Regression and Classification tree for the *nec* (a) and the *nec.d* (b) variables respectively. In (a), at each node, the average value of the dependent variable (percentage of affected colonies) and the number of observations is indicated. In (b), the frequencies of the two classes are given at each node, as well as the most frequent class.

Very few statistical tools provide a direct modeling issue for this situation. For both discrete and continuous multivariate output, there exist extensions of CART model to handle these cases (Zhang, 1998; Segal, 1992).

These extensions focus mainly on adapting the splitting criterion of CART to the multivariate case, and finding convenient estimators for the criterion at each step of the tree construction. For ecological applications of these approaches we refer to De'ath (2002) and Kocev et al. (2009).

5.3. Functional data

When either the output variable Y or one of the explanatory variables may be seen as discretized versions of a continuous curve, on may need functional data analysis (FDA) tools (Ramsay and Silverman, 2002). FDA approaches have appeared in the early 90s and have extended several classical data analysis techniques (Principal Component Analysis, Regression, Discriminant Analysis, Analysis of Variance, ...) to the functional case. The main idea of these approaches is to construct a mathematical representation of each observed curve in a functional space using a common basis such as splines or wavelets.

FDA approaches have been integrated to some Supervised Learning techniques. For instance, Nerini and Ghattas (2007) extended CART to the case where Y is a density curve discretized at q points. The extension they proposed may handle also the case where the discretization grid is different among the observed curves. Their application concerned the prediction of the zooplankton size distributions using wind and temperature measurements.

For independent functional data, very few works appear in the literature. In particular there exist an extension of CART to the case where the independent variable is supposed to be a time series (Yuu et al., 2003). In this work FDA approaches are not used. The underlying idea consists on adapting the splitting rule which has initially the form $x_j < s$. The splitting rule becomes $D(x(t), x_0(t)) < s$ where $x(t)$ is one explanatory functional independent variable (observed temperatures over one day for example), $x_0(t)$ is one of the observations present in the learning sample, s is a threshold and D is a dissimilarity measure between two time series. The dissimilarity measure is based on *Dynamic Time Warping* and computed directly over the original data.

Acknowledgments

The authors gratefully acknowledge the funding support given by the Alβan Program (C. Crisci Doctoral Fellowship) as well as C. Deniau and L. Reboul for their important comments and remarks on the manuscript. Finally, we want to thank to J. Garrabou (Institut de Cincies del Mar, Barcelona), C. Linares (Universitat de Barcelona) and R. Coma (Centre d'Estudis Avançats de Blanes) for the mortality data.

References

Ballesteros, E., 2006. Mediterranean coralligenous assemblages: a synthesis of present knowledge. Oceanography and Marine Biology 44, 123–195.
Bell, J.F., 1996. Application of classification trees to habitat preference of upland birds. Journal of Applied Statistics 23, 349–359.
Benito Garzón, M., Blazek, R., Neteler, M., Sánchez de Dios, R., Sainz Ollero, H., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian peninsula. Ecological Modelling 197, 383–393.
Bensoussan, N., Romano, J.C., Harmelin, J.G., Garrabou, J., 2010. High resolution characterization of northwest Mediterranean coastal waters thermal regimes: to better understand responses of benthic communities to climate change. Estuarine, Coastal and Shelf Science 87, 431–441.
Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.
Breiman, L., 1996b. Heuristic of instability and stabilization in model selection. The Annals of Statistics 24 (6), 2350–2383.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Chapman & Hall/CRC.
Cerrano, C., Bavestrello, G., Bianchi, C., Cattaneo-vietti, R., Bava, S., Morganti, C., Morri, C., Picco, P., Sara, G., Shiaparelli, S., Siccardi, A., Sponga, F., 2000. A catastrophic mass mortality episode of gorgonians and other organisms in the Ligurian sea (North-Western Mediterranean), summer 1999. Ecology Letters 3, 284–293.
Chaudhuri, P., Lo, W., Loh, W., Yang, C.C., 1995. Generalized regression trees. Statistica Sinica 17 (3), 641–666.
Crisci, C., Besoussan, N., Romano, J., Garrabou, J., 2011. Temperature anomalies and mortality events in marine communities: insights on factors behind differential mortality impacts in the NW Mediterranean. PLoS ONE 6, e23814, http://dx.doi.org/10.1371%2Fjournal.pone.0023814.
Cutler, D.R., Edwards, T.C.J., Beard, K., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forest for classification in ecology. Ecology 88 (11), 2783–2792.
Davison, A., Hinkley, D., 1997. Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 1. Cambridge University Press.
De'ath, G., 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83 (4), 1105–1117.
De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243–251.
De'ath, G., Fabricius, K.E., 2000. Classification and regression tress: a powerful yet simple technique for ecological data analyses. Ecology 81 (11), 3178–3192.
Defeo, O., Gómez, J., 2005. Morphodynamics and habitat safety in sandy beaches: life-history adaptations in a supralittoral amphipod. Marine Ecology Progress Series 293, 143–153.
Džeroski, S., 2001. Application of symbolic machine learning to ecological modelling. Ecological Modelling 146, 263–273.
Devroye, L., Gyorfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. Springer.
Efron, B., Tibshirani, R.J., May 1994. An Introduction to the Bootstrap. Chapman & Hall/CRC.
Fernandes, J., Irigoien, X., Goikoetxea, N., Lozano, J.A., Inza, I., Pérez, A., Bode, A., 2010. Fish recruitment prediction, using robust supervised classification methods. Ecological Modelling 221, 338–352.
Fielding, A., 1999. Machine Learning Methods for Ecological Applications. Springer.
Flach, P., 2001. On the state of the art in machine learning: a personal review. Artificial Intelligence 13 (1), 199–222.
Franklin, J., 1998. Predicting the distribution of shrub species in Southern California from climate and terrain-derived variables. Journal of Vegetation Science 9, 733–748.
Freund, Y., 1995. Boosting a weak learning algorithm by majority. Information and Computation 121 (2), 256–285.
Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. Journal of the American Statistical Association 76 (376), 817–823.
Garrabou, J., Coma, R., Bensousson, N., Chevaldonné, P., Cigliano, M., Diaz, D., Harmelin, J.G., Gambi, M.C., Kersting, D.K., Ledoux, J.B., Lejeusne, C., Linares, C., Marschal, C., Perez, T., Ribes, M., Romano, J.C., Serrano, E., Teixido, N., Torrents, O., Zabala, M., Zuberer, F., Cerrano, C., 2009. Mass mortality in NW Mediterranean rocky benthic communities: effects of the 2003 heat wave. Global Change Biology 15, 1090–1103.
Grubb, T., Bowerman, W., 1997. Variations in breeding bald eagle responses to jets, light planes and helicopters. Journal of Raptor Research 31, 213–222.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157 (2), 89–100.
Guruswami, V., Sahai, A.,1999. Multiclass learning, boosting, and error-correcting codes. In: COLT'99: Proceedings of the Twelfth Annual Conference on Computational Learning Theory. ACM Press, New York, NY, USA, pp. 145–155.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall/CRC.
Hastie, T., Tibshirani, R., Friedman, J.H., 2003. The Elements of Statistical Learning. Springer.
Huttmann, F., Lock, A., 1997. A new software system for the PIROP database: data flow and an approach for a seabird-depth analysis. ICES Journal of Marine Science 54 (4), 518–523.
Knudby, A., Brenning, A., LeDrew, E., 2010. New approaches to modelling fish–habitat relationships. Ecological Modelling 221, 503–511.
Kocev, D., Džeroski, S., White, M., Newell, G., Griffioen, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling 220, 1159–1168.
Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., Džeroski, S., 2010. Learning habitat models for the diatom community in lake Prespa. Ecological Modelling 221, 330–337.
Lek, S., Guégan, J., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120, 65–73.
Loh, W., Vanichsetakul, N., 1988. Tree-structured classification via generalized discriminant analysis. Journal of the American Statistical Association 83 (403), 715–728.
Merckx, B., Goethals, P., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2009. Predictability of marine nematode biodiversity. Ecological Modelling 220, 1449–1458.
Morris, C.W., Autret, A., Boddy, L., 2001. Support vector machines for identifying organisms a comparison with strongly partitioned radial basis function networks. Ecological Modelling 146, 57–67.

Nelder, J., Wedderburn, R., 1972. Generalized linear models. Journal of the Royal Statistical Society 135 (3), 370–384.
Nerini, D., Ghattas, B., 2007. Classifying densities using functional regression trees: applications in oceanology. Computational Statistics & Data Analysis 51 (10), 4984–4993.
Perez, T., Garrabou, J., Sartoretto, S., Harmelin, J.G., Francour, P., Vacelet, J., 2000. Mortalité massive d'invertébres marins: un événement sans précédent en méditerranée nord-occidentale. Comptes Rendus de l'Académie des Sciences (Paris) – Series III 323, 853–865.
Phetkaew, T., Kijsirikul, B., Rivepiboon, W., 2003. Multiclass classification of support vector machines by reordering adaptive directed acyclic graph. In: International Workshop on Intelligent Systems.
Pontin, D., Schliebs, S., Worner, S., Watts, M., 2011. Determining factors that influence the dispersal of a pelagic species: a comparison between artificial neural networks and evolutionary algorithms. Ecological Modelling 222, 1657–1665.
Ramsay, J., Silverman, B., 2002. Applied functional data analysis: methods and case studies. Springer Series in Statistics, Springer.
Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecological Modelling 146, 303–310.
Ribeiro, R., Torgo, L., 2008. A comparative study on predicting algae blooms in Douro River, Portugal. Ecological Modelling 212, 86–91.
Ribic, C., Ainley, D., 1997. The relationships of seabird assemblages to physical habitat features in pacific equatorial waters during spring 1984–1991. ICES Journal of Marine Science 54 (4), 593–599.
Romano, J.C., Lugrezi, M.C., Durand, D., Durand-LeBreton, F., 2010. Serie du maré-graphe de marseille: mesures de témperatures de surface de la mer de 1895 á 1956: une correction. Comptes Rendus Geoscience 342 (12), 873–880.
Rosenblatt, F., 1962. Principles of Neurodynamics: Perceptron and Theory of Brain Mechanisms. Spartan Books.
Ryder, T.J., Irwin, L., 1987. Winter habitat relationships of pronghorns in Southcentral Wyoming. The Journal of Wildlife Management 51 (1), 79–85.
Segal, M.R., 1992. Tree-structured methods for longitudinal data. Journal of the American Statistical Association 87 (418), 407–418.
Shan, Y., Paull, D., McKay, R., 2006. Machine learning of poorly predictable ecological data. Ecological Modelling 195, 129–138.
Taylor, B.J., 2006. Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Springer.
Vapnik, V., 1998. Statistical Learning Theory. Wiley.
Volf, G., Atanasova, N., Kompare, B., Precali, R., Oani, N., 2011. Descriptive and prediction models of phytoplankton in the northern adriatic. Ecological Modelling 222, 2502–2511.
Yuu, Y., Einoshin, S., Hideto, Y., Katsuhiko, T., 2003. Decision-tree induction from time-series data based on a standard-example split test. In: Proceedings of the Twentieth International Conference on Machine Learning, Washington, DC.
Zhang, H., 1998. Classification trees for multiple binary responses. Journal of the American Statistical Association 93 (441), 180193.
Zhu, J., Rosset, S., Zou, H., Hastie, T., 2006. Multi-class AdaBoost. Technical Report. Standford University.