Supporting information S1 for Pichler & Hartig – Machine Learning and Deep Learning – A review for Ecologists

Pichler & Hartig, 2022

Table of contents

**Summary:** This document provides supporting information on Pichler & Hartig – Machine Learning and Deep Learning – A review for ecologists. The first section describes the methods for the trend analysis and the word clouds, the second section demnstrates the application of common supervised Machine learning and Deep learning algorithms in the R programming languages based on code examples.

## 1 Trend analysis

For the global trend analysis in Figure 1, we used the R package ‘europepmc’ (v0.4.1, Jahn (2021)) to search from 1920 to 2021 the PubMed and Medline NLM databases. We used the following queries ‘deep learning’, (‘machine learning’ OR ‘machine-learning’), and (‘p value’ OR ‘p-value’ OR ‘statistically significant) as representatives for Deep Learning, Machine Learning, and classical statistical approaches. The number of hits were normalized by total hits in each year. For the stream charts in Figure 2, we used the search queries Table S1 and added them to the query’ AND (“ecology” OR “ecolog\*” OR “evolution”) to restrict the queries to hits from the ecology and evolution field.

**Table** : Search queries and their corresponding ML and DL algorithm.

| **Queries** | **ML and DL algorithm** |
| --- | --- |
| ("artificial neural network" OR "deep neural network" OR   "multi-layer perceptron" OR "fully connected neural network") | Deep neural network (ANN) |
| ("convolutional neural network" OR "object detection") | Convolutional neural network (CNN) |
| ("recurrent neural network") | Recurrent neural network (RNN) |
| ("graph neural network" OR "graph convolutional") | Graph neural network (GNN) |
| ("random forest") | Random Forest (RF) |
| ("boosted regression tree" OR "boosted reg" OR   "gradient boosting" OR "adaboost") | Boosted Regression Trees (BRT) |
| ("k-nearest-neighbor") | k-nearest neighbor (kNN) |
| ("ridge regression" OR "lasso regression" OR   "elastic-net" OR "elastic net") | Ridge, lasso, or elastic-net regression |
| ("support vector machine" OR "support vector") | Support vector machine (SVM) |

For the word clouds in Table 1, we used again the ‘europepmc’ R package to search abstracts and titles within the ML and DL algorithm specific queries (Table 1) for the following ecological keywords: species distribution, species interaction, mortality, remote sensing, invasive, decision making, ecosystem, species identification, species detection, extinction, functional trait, ecological network, biodiversity, and camera trap.

We used the R packages ‘tm’ (Feinerer, Hornik, and Meyer (2008)), ‘wordcloud’ (Fellows (2018)) ,and ‘wordcloud2’ (Lang and Chien (2018)) to analyze and create the final word cloud plots.

## 2 Algorithms

The goal of this chapter is to provide short code examples for all common supervised ML algorithms. All examples are shown in the R programming language and are demonstrated at the example of the iris dataset (4 continuous and 1 nominal variables). We demonstrate a) how to apply the algorithms on classification tasks (response = nominal, three species) and how to generate class specific (quasi) probability predictions and b) how to apply the algorithms to regression tasks (response = continuous) and make continuous predictions.

### 2.1 Ridge, LASSO, and elastic-net regression

We can use the ‘glmnet’ R package (Simon et al. (2011)) for Ridge, LASSO, or elastic-net regularization. The ‘glmnet’ package supports different response families including ‘gaussian’, ‘binomial’ and ‘Poisson’. The strength of the regularization is set by the ‘lambda’ argument () and the weighting between Ridge and LASSO regularization by the ‘alpha’ parameter ():

Setting alpha = 0 turns off the LASSO and alpha = 1 the Ridge. Alphas between (0,1) will use both regularization types, turning the model into an elastic-net regularization.

When using regularization, it is important to scale all features otherwise effects for features that are on a larger scale are stronger regularized.

#### 2.1.1 Classification

Build models (for regularization it is important to scale the features):

##### 2.1.1.1 R

library(glmnet)  
X = scale(iris[,1:4])  
Y = iris$Species  
  
# Ridge:  
ridge = glmnet(X, Y, family = "multinomial", alpha = 0, lambda = 0.01)  
  
# LASSO:  
lasso = glmnet(X, Y, family = "multinomial", alpha = 1, lambda = 0.01)  
  
# Elastic-net:  
elastic = glmnet(X, Y, family = "multinomial", alpha = 0.5, lambda = 0.01)

Make predictions (class probabilities):

head(predict(lasso, newx = X, type = "response")[,,1], n = 3)

setosa versicolor virginica  
[1,] 0.9858987 0.01410131 3.438452e-09  
[2,] 0.9668897 0.03311031 1.397684e-08  
[3,] 0.9815369 0.01846312 5.279315e-09

##### 2.1.1.2 Python

In the sklearn implementation the regularization strength parameter ‘C’ corresponds to the lambda parameter from glmnet:

from sklearn.linear\_model import LogisticRegression  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
  
# Ridge:  
ridge = LogisticRegression(multi\_class='multinomial',   
 penalty = "l2",   
 C = 0.01,   
 solver="saga")  
ridge.fit(X, Y)  
  
# LASSO:

LogisticRegression(C=0.01, multi\_class='multinomial', solver='saga')

lasso = LogisticRegression(multi\_class='multinomial',   
 penalty = "l1",   
 C = 0.01,   
 solver="saga")  
lasso.fit(X, Y)  
  
# Elastic-net:

LogisticRegression(C=0.01, multi\_class='multinomial', penalty='l1',  
 solver='saga')

elastic = LogisticRegression(multi\_class='multinomial',   
 penalty = "elasticnet",   
 C = 0.01,   
 l1\_ratio=0.5,   
 solver="saga")  
elastic.fit(X, Y)  
  
# Make predictions (class probabilities):

LogisticRegression(C=0.01, l1\_ratio=0.5, multi\_class='multinomial',  
 penalty='elasticnet', solver='saga')

lasso.predict\_proba(X)[0:10,:]

array([[0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527],  
 [0.33912902, 0.31861571, 0.34225527]])

#### 2.1.2 Regression

##### 2.1.2.1 R

X = scale(iris[,2:4])  
Y = iris[,1]  
  
# Ridge:  
ridge = glmnet(X, Y, family = gaussian(), alpha = 0, lambda = 0.01)  
  
# LASSO:  
lasso = glmnet(X, Y, family = gaussian(), alpha = 1, lambda = 0.01)  
  
# Elastic-net:  
elastic = glmnet(X, Y, family = gaussian(), alpha = 0.5, lambda = 0.01)

Make predictions (class probabilities):

head(predict(lasso, newx = X), n = 3)

s0  
[1,] 5.006484  
[2,] 4.720600  
[3,] 4.781548

##### 2.1.2.2 Python

For regressions we can use the ElasticNet model class, here, however, lambda corresponds to alpha and l1\_ratio to the alpha parameter.

from sklearn.linear\_model import ElasticNet  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
  
# Ridge:  
ridge = ElasticNet(alpha = 0.01,  
 l1\_ratio = 0.011)  
ridge.fit(X, Y)  
  
# LASSO:

ElasticNet(alpha=0.01, l1\_ratio=0.011)

lasso = ElasticNet(alpha = 0.01,  
 l1\_ratio = 1.0)  
lasso.fit(X, Y)  
  
# Elastic-net:

ElasticNet(alpha=0.01, l1\_ratio=1.0)

elastic = ElasticNet(alpha = 0.01,  
 l1\_ratio = 0.5)  
elastic.fit(X, Y)  
  
# Make predictions:

ElasticNet(alpha=0.01)

lasso.predict(X)[0:10]

array([5.0064384 , 4.72032938, 4.78125162, 4.83107256, 5.06366021,  
 5.36149937, 4.93202142, 5.00273797, 4.66310758, 4.84826774])

### 2.2 Support Vector Machines

The support vector machine (SVM) algorithm estimates hyper-planes to separate our response species. In the following we use the ‘e1071’ package which supports a variety of different SVM algorithms (Meyer et al. (2022)).

#### 2.2.1 Classification

##### 2.2.1.1 R

library(e1071)  
X = scale(iris[,1:4])  
Y = iris$Species  
  
sv = svm(X, Y, probability = TRUE)   
summary(sv)

Call:  
svm.default(x = X, y = Y, probability = TRUE)  
  
  
Parameters:  
 SVM-Type: C-classification   
 SVM-Kernel: radial   
 cost: 1   
  
Number of Support Vectors: 51  
  
 ( 8 22 21 )  
  
  
Number of Classes: 3   
  
Levels:   
 setosa versicolor virginica

Make predictions (class probabilities):

head(attr(predict(sv, newdata = X, probability = TRUE), "probabilities"), n = 3)

setosa versicolor virginica  
1 0.9800342 0.01127612 0.008689710  
2 0.9726030 0.01808943 0.009307537  
3 0.9787203 0.01189528 0.009384438

##### 2.2.1.2 Python

from sklearn import svm  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
model = svm.SVC(probability=True).fit(X, Y)  
  
# Make predictions (class probabilities):  
  
model.predict\_proba(X)[0:10,:]

array([[0.97967217, 0.01156447, 0.00876336],  
 [0.97233838, 0.01826359, 0.00939803],  
 [0.97841386, 0.01213243, 0.00945372],  
 [0.97434554, 0.01551157, 0.0101429 ],  
 [0.97879051, 0.01192634, 0.00928315],  
 [0.97326027, 0.01709586, 0.00964386],  
 [0.97495475, 0.01379427, 0.01125098],  
 [0.97957726, 0.01167442, 0.00874832],  
 [0.96548202, 0.02268892, 0.01182905],  
 [0.97617521, 0.01428498, 0.00953981]])

#### 2.2.2 Regression

##### 2.2.2.1 R

library(e1071)  
X = scale(iris[,2:4])  
Y = iris[,1]  
  
sv = svm(X, Y)   
summary(sv)

Call:  
svm.default(x = X, y = Y)  
  
  
Parameters:  
 SVM-Type: eps-regression   
 SVM-Kernel: radial   
 cost: 1   
 gamma: 0.3333333   
 epsilon: 0.1   
  
  
Number of Support Vectors: 124

Make predictions (class probabilities):

head(predict(sv, newdata = X), n = 3)

1 2 3   
5.042085 4.711768 4.836291

##### 2.2.2.2 Python

from sklearn import svm  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
model = svm.SVR().fit(X, Y)  
  
# Make predictions:  
  
model.predict(X)[0:10]

array([5.03583855, 4.69496586, 4.81438855, 4.77951854, 5.10018373,  
 5.29981857, 4.97308737, 4.98199033, 4.63701656, 4.78431078])

### 2.3 k-nearest-neighbor

The k-nearest-neighbor algorithm doesn’t really learn from the data, predictions for new observations are made based on the class affiliation (or response value) of the nearest neighbors, e.g. by majority voting or averaging. The nearest neighbors are found by calculating the distance of the new observation to all observations in the train dataset.

In the following we use the ‘kknn’ package (Schliep and Hechenbichler (2016)). Different to other ML packages we can provide here already the test dataset in the fit function.

#### 2.3.1 Classification

##### 2.3.1.1 R

library(kknn)  
X = scale(iris[,1:4])  
Y = iris[,5,drop=FALSE]  
data = cbind(Y, X)  
  
knn = kknn(Species~., train = data, test = data)

Make predictions (class probabilities):

head(knn$prob, n = 3)

setosa versicolor virginica  
[1,] 1 0 0  
[2,] 1 0 0  
[3,] 1 0 0

##### 2.3.1.2 Python

from sklearn.neighbors import KNeighborsClassifier  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
model = KNeighborsClassifier().fit(X, Y)  
  
# Make predictions:  
  
model.predict\_proba(X)[0:10,:]

array([[1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.]])

#### 2.3.2 Regression

##### 2.3.2.1 R

library(e1071)  
X = scale(iris[,2:4])  
data = cbind(iris[,1,drop=FALSE], X)  
  
knn = kknn(Sepal.Length~., train = data, test = data)

Make predictions (class probabilities):

head(predict(knn), n = 3)

[1] 5.188492 4.739986 4.685332

##### 2.3.2.2 Python

from sklearn.neighbors import KNeighborsRegressor  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
model = KNeighborsRegressor().fit(X, Y)  
  
# Make predictions:  
  
model.predict(X)[0:10]

array([5.18, 4.78, 4.68, 4.76, 5.02, 5.34, 5.06, 5.1 , 4.7 , 4.8 ])

### 2.4 Random forest

The random forest (RF) algorithm is probably one of the most famous ML algorithms, and not without reason. Compared to other well performing algorithms, the RF algorithm has only a few hyper-parameters and because of the bagging and the random sampling of available variables in for the node splits, it has a well working internal complexity adaption.

In the following, we use the ‘ranger’ package (Wright and Ziegler (2017)).

#### 2.4.1 Classification

##### 2.4.1.1 R

library(ranger)  
X = iris[,1:4]  
Y = iris[,5,drop=FALSE]  
data = cbind(Y, X)  
  
rf = ranger(Species~., data = data, probability = TRUE, importance = "impurity")

Show feature importances:

importance(rf)

Sepal.Length Sepal.Width Petal.Length Petal.Width   
 9.280258 1.589384 43.076529 41.808611

Make predictions (class probabilities):

head(predict(rf, data = data)$predictions, n = 3)

setosa versicolor virginica  
[1,] 1.000 0.000 0  
[2,] 0.998 0.002 0  
[3,] 1.000 0.000 0

##### 2.4.1.2 Python

from sklearn.ensemble import RandomForestClassifier  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
model = RandomForestClassifier().fit(X, Y)

Feature importance

print(model.feature\_importances\_)

[0.1081721 0.02584575 0.48330769 0.38267446]

Make predictions:

model.predict\_proba(X)[0:10,:]

array([[1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.],  
 [1., 0., 0.]])

#### 2.4.2 Regression

##### 2.4.2.1 R

library(ranger)  
X = iris[,2:4]  
data = cbind(iris[,1,drop=FALSE], X)  
  
rf = ranger(Sepal.Length~., data = data, importance = "impurity")

Show feature importances:

importance(rf)

Sepal.Width Petal.Length Petal.Width   
 12.30838 45.11126 38.45659

Make predictions (class probabilities):

head(predict(rf, data = data)$predictions, n = 3)

[1] 5.118336 4.748380 4.646246

##### 2.4.2.2 Python

from sklearn.ensemble import RandomForestRegressor  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
model = RandomForestRegressor().fit(X, Y)

Feature importance:

print(model.feature\_importances\_)

[0.08109138 0.85628627 0.06262236]

Make predictions:

model.predict(X)[0:10]

array([5.106 , 4.83225 , 4.5762 , 4.78508571, 5.012 ,  
 5.446 , 4.7985 , 5.04970238, 4.53825 , 4.84353571])

### 2.5 Boosted gradient trees

Boosted gradient machines achieve currently state-of-the-art performance for structured (tabular) data which makes them probably one of the most important algorithms for E&E where structured data dominates the field.

In the following, we use the ‘xgboost’ package (Chen et al. (2022)).

#### 2.5.1 Classification

##### 2.5.1.1 R

library(xgboost)  
X = as.matrix(iris[,1:4])  
Y = as.integer(iris[,5]) - 1 # classes must be integers starting from 0  
  
xgdata = xgb.DMatrix(X, label = Y)  
  
# nrounds = number of trees in the ensemble  
brt = xgboost(data = xgdata,   
 objective="multi:softprob",   
 nrounds = 50,   
 num\_class = 3,  
 verbose = 0)

Show feature importances:

xgb.importance(model = brt)

Feature Gain Cover Frequency  
1: Petal.Length 0.671879438 0.57441039 0.3792049  
2: Petal.Width 0.311535837 0.29261084 0.3088685  
3: Sepal.Width 0.010177107 0.04910115 0.1162080  
4: Sepal.Length 0.006407618 0.08387763 0.1957187

Make predictions (class probabilities):

head(matrix(predict(brt, newdata = xgb.DMatrix(X)), ncol =3), n = 3)

[,1] [,2] [,3]  
[1,] 0.995287061 0.002195822 0.001027058  
[2,] 0.003323558 0.995396435 0.001592265  
[3,] 0.001389398 0.002407764 0.997380674

##### 2.5.1.2 Python

import xgboost as xgb  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
# Parameters:  
param = {  
 'max\_depth':2,   
 'eta':1,   
 'objective':'multi:softmax' }  
num\_round = 50  
  
model = xgb.XGBClassifier(param, num\_round, verbosity = 0).fit(X, Y)

/Users/maximilianpichler/Library/r-miniconda/envs/r-reticulate/lib/python3.6/site-packages/xgboost/core.py:433: FutureWarning: Pass `objective, use\_label\_encoder` as keyword args. Passing these as positional arguments will be considered as error in future releases.  
 format(", ".join(args\_msg)), FutureWarning)  
/Users/maximilianpichler/Library/r-miniconda/envs/r-reticulate/lib/python3.6/site-packages/xgboost/sklearn.py:1146: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use\_label\_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num\_class - 1].  
 warnings.warn(label\_encoder\_deprecation\_msg, UserWarning)

Feature importance

model.feature\_importances\_

array([0.00959796, 0.01645038, 0.67658573, 0.29736587], dtype=float32)

Make predictions:

model.predict\_proba(X)[0:10,:]

array([[9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9636227e-01, 2.3820533e-03, 1.2557388e-03],  
 [9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9679452e-01, 2.3830866e-03, 8.2237815e-04],  
 [9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9680281e-01, 2.3831066e-03, 8.1413286e-04],  
 [9.9636227e-01, 2.3820533e-03, 1.2557388e-03],  
 [9.9679452e-01, 2.3830866e-03, 8.2237815e-04]], dtype=float32)

#### 2.5.2 Regression

##### 2.5.2.1 R

library(xgboost)  
X = as.matrix(iris[,2:4])  
Y = iris[,1]  
  
xgdata = xgb.DMatrix(X, label = Y)  
  
# nrounds = number of trees in the ensemble  
brt = xgboost(data = xgdata,   
 objective="reg:squarederror",   
 nrounds = 50,   
 verbose = 0)

Show feature importances:

xgb.importance(model = brt)

Feature Gain Cover Frequency  
1: Petal.Length 0.86781219 0.4789538 0.3789062  
2: Petal.Width 0.06987880 0.2128402 0.2626953  
3: Sepal.Width 0.06230901 0.3082060 0.3583984

Make predictions:

head(predict(brt, newdata = xgb.DMatrix(X), n = 3))

[15:06:57] WARNING: amalgamation/../src/c\_api/c\_api.cc:785: `ntree\_limit` is deprecated, use `iteration\_range` instead.

[1] 3.506606 3.506606 3.506606 3.506606 3.506606 3.506606

##### 2.5.2.2 Python

import xgboost as xgb  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
# Parameters:  
model = xgb.XGBRegressor(  
 objective = 'reg:squarederror',  
 max\_depth = 2,   
 n\_estimators = 50,   
 verbosity = 0).fit(X, Y)

/Users/maximilianpichler/Library/r-miniconda/envs/r-reticulate/lib/python3.6/site-packages/xgboost/data.py:122: UserWarning: Use subset (sliced data) of np.ndarray is not recommended because it will generate extra copies and increase memory consumption  
 "memory consumption")

Feature importance:

print(model.feature\_importances\_)

[0.07703022 0.85646353 0.06650625]

Make predictions:

model.predict(X)[0:10]

array([5.107 , 4.8285 , 4.60348333, 4.79052143, 5.024 ,  
 5.406 , 4.79 , 5.06125 , 4.601 , 4.8383 ])

### 2.6 Deep neural networks

Deep neural networks, or more precisely here fully connected neural networks, can be flexibly built which makes their application more challenging than other ML algorithms.

In the following, we use the ‘keras’ (Allaire and Chollet (2022); Chollet et al. (2015)) package which is a higher level API on the python ‘tensorflow’ framework (Abadi et al. (2016)).

#### 2.6.1 Classification

##### 2.6.1.1 R

library(keras)  
X = scale(as.matrix(iris[,1:4]))  
Y = as.integer(iris$Species)  
# We need to one hot encode our response classes  
YT = k\_one\_hot(Y-1L, num\_classes = 3)  
  
DNN = keras\_model\_sequential() %>%   
 # first hidden layer  
 layer\_dense(input\_shape = ncol(X),   
 units = 10,   
 activation = "relu") %>%   
 # second hidden layer with regularization  
 layer\_dense(units = 20,   
 activation = "relu",  
 kernel\_regularizer = regularizer\_l1()) %>%   
 # output layer, 3 output neurons for our three classes  
 # and softmax activation to get quasi probabilities   
 # that sum up to 1 for each observation  
 layer\_dense(units = 3,   
 activation = "softmax")  
  
# print architecture  
summary(DNN)  
  
# add loss function and optimizer  
DNN %>%   
 compile(loss = loss\_categorical\_crossentropy,  
 optimizer = optimizer\_adamax(0.01))  
  
# train model  
DNN %>%   
 fit(X, YT, epochs = 50, verbose = 0)

Make predictions (class probabilities):

head(predict(DNN, X), n = 3)

##### 2.6.1.2 Python

from tensorflow import keras  
from tensorflow.keras.layers import \*  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
X = scale(iris.data)  
Y = iris.target  
  
# We need to one hot encode our response classes  
YT = keras.one\_hot(Y, num\_classes = 3)  
  
DNN = keras.model\_sequential()  
 # first hidden layer  
DNN.add(Dense(  
 input\_shape = X.shape[1],   
 units = 10,   
 activation = "relu"))   
 # second hidden layer with regularization  
DNN.add(Dense(  
 units = 20,   
 activation = "relu",  
 kernel\_regularizer = keras.regularizers.l1()))  
 # output layer, 3 output neurons for our three classes  
 # and softmax activation to get quasi probabilities   
 # that sum up to 1 for each observation  
DNN.add(Dense(  
 units = 3,   
 activation = "softmax"))  
  
# print architecture  
DNN.summary()  
  
# add loss function and optimizer  
DNN.compile(loss = keras.losses.categorical\_crossentropy,  
 optimizer = keras.optimizer.adamax(0.01))  
  
# train model  
DNN.fit(X, YT, epochs = 50, verbose = 0)

Make predictions:

DNN.predict(X)[0:10,:]

#### 2.6.2 Regression

##### 2.6.2.1 R

library(keras)  
X = scale(as.matrix(iris[,2:4]))  
Y = as.matrix(iris[,1,drop=FALSE])  
  
DNN = keras\_model\_sequential() %>%   
 # first hidden layer  
 layer\_dense(input\_shape = ncol(X),   
 units = 10,   
 activation = "relu") %>%   
 # second hidden layer with regularization  
 layer\_dense(units = 20,   
 activation = "relu",  
 kernel\_regularizer = regularizer\_l1()) %>%   
 # output layer, one output neuron for one response  
 # and no activation function  
 layer\_dense(units = 1)  
  
# print architecture  
summary(DNN)  
  
# add loss function and optimizer  
DNN %>%   
 compile(loss = loss\_mean\_squared\_error,  
 optimizer = optimizer\_adamax(0.01))  
  
# train model  
DNN %>%   
 fit(X, YT, epochs = 50, verbose = 0)

Make predictions:

head(predict(DNN, X), n = 3)

##### 2.6.2.2 Python

from tensorflow import keras  
from tensorflow.keras.layers import \*  
from sklearn import datasets  
from sklearn.preprocessing import scale  
iris = datasets.load\_iris()  
data = iris.data  
X = scale(data[:,1:4])  
Y = data[:,0]  
  
DNN = keras.model\_sequential()  
 # first hidden layer  
DNN.add(Dense(  
 input\_shape = X.shape[1],   
 units = 10,   
 activation = "relu"))   
 # second hidden layer with regularization  
DNN.add(Dense(  
 units = 20,   
 activation = "relu",  
 kernel\_regularizer = keras.regularizers.l1()))  
 # output layer, 3 output neurons for our three classes  
 # and softmax activation to get quasi probabilities   
 # that sum up to 1 for each observation  
DNN.add(Dense(  
 units = 1,   
 activation = "None"))  
  
# print architecture  
DNN.summary()  
  
# add loss function and optimizer  
DNN.compile(loss = keras.losses.mean\_squared\_error,  
 optimizer = keras.optimizer.adamax(0.01))  
  
# train model  
DNN.fit(X, YT, epochs = 50, verbose = 0)

Make predictions:

DNN.predict(X)[0:10]

### 2.7 Convolutional neural networks

Convolutional neural networks (CNN) are also deep neural networks but they are based on convolutional layers, which is a biologically inspired variation optimized to process image-based data (LeCun, Bengio, and Hinton (2015)). CNNs consist of two stages, in the first, the images are passed through convolutional layers and the models learns to detect edges and shapes in the images. In the second stage, the dimensions are dropped and fully-connected layers are used to classify the previously identified shapes.

In the following, we will use again the ‘keras’ package but we will not differentiate between classification and regression because the only difference would be to change the last layer and the loss function (see section ‘Deep neural networks’).

We will demonstrate the application of CNNs with the MNIST dataset which consists of handwritten digits. The objective of the CNNs is to classify the images. The MNIST dataset is one of the most famous benchmark dataset for image-based tasks (LeCun, Cortes, and Burges (2010)).

#### 2.7.1 R

library(keras)  
data = keras::dataset\_mnist()  
train = data$train  
X = train$x/255  
# we have to add a dimension that   
# informs the network about the channels  
# of the images  
X = array(X, dim = c(dim(X), 1))  
YT = k\_one\_hot(train$y, num\_classes = 10)  
  
  
  
CNN =   
 keras\_model\_sequential() %>%   
 # first hidden layer  
 layer\_conv\_2d(input\_shape = list(28, 28, 1),   
 filters = 16,  
 kernel\_size = c(2, 2),  
 activation = "relu") %>%  
 layer\_average\_pooling\_2d() %>%   
 layer\_conv\_2d(filters = 8,  
 kernel\_size = c(2, 2),  
 activation = "relu") %>%  
 # we use a normal DNN on top of the CNN:  
 # the layer flatten will remove the additional   
 # dimensions  
 layer\_flatten() %>%   
 layer\_dense(units = 20,   
 activation = "relu") %>%  
 # 10 output neurons for 10 classes  
 layer\_dense(units = 10,   
 activation = "softmax")  
  
  
# print architecture  
summary(CNN)  
  
# add loss function and optimizer  
CNN %>%   
 compile(loss = loss\_categorical\_crossentropy,  
 optimizer = optimizer\_adamax(0.01))  
  
CNN %>%   
 fit(X, YT, epochs = 3, verbose = 1, batch\_size = 125)

Make predictions (class probabilites):

head(predict(CNN, X[1:100,,,,drop=FALSE]), n = 3)

#### 2.7.2 Python

from tensorflow import keras  
from tensorflow.keras.layers import \*  
data = keras.datasets.mnist.load\_data()  
train = data.train[0][0]  
labels = data.train[0][1]  
  
# We need to one hot encode our response classes  
YT = keras.one\_hot(labels, num\_classes = 10)  
  
CNN = keras.model\_sequential()  
 # first hidden layer  
CNN.add(Conv2D(input\_shape = [28, 28, 1],   
 filters = 16,  
 kernel\_size = (2, 2),  
 activation = "relu"))  
CNN.add(AveragePooling2D())  
CNN.add(Conv2D(filters = 8,  
 kernel\_size = (2, 2),  
 activation = "relu"))  
 # we use a normal DNN on top of the CNN:  
 # the layer flatten will remove the additional   
 # dimensions  
CNN.add(Flatten())  
 # output layer, 3 output neurons for our three classes  
 # and softmax activation to get quasi probabilities   
 # that sum up to 1 for each observation  
CNN.add(Dense(  
 units = 10,   
 activation = "softmax"))  
  
# print architecture  
CNN.summary()  
  
# add loss function and optimizer  
CNN.compile(loss = keras.losses.categorical\_crossentropy,  
 optimizer = keras.optimizer.adamax(0.01))  
  
# train model  
CNN.fit(train, YT, epochs = 50, verbose = 0)

Make predictions:

CNN.predict(train[0:10,:,:])

### 2.8 Recurrent neural networks

Recurrent neural networks are also deep neural networks but use layers specialized to handle time-series. In the following, we will use again the ‘keras’ package but we will not differentiate between classification and regression because the only difference would be to change the last layer and the loss function (see section ‘Deep neural networks’).

About the data, we simulated in the following one time series from a simple ARIMA process, using the ‘arima.sim’ function. Our goal is to train a net which is able to predict the next 10 time points based on the previous 10 time points.

#### 2.8.1 R

## RNNs  
library(keras)  
data = as.matrix(arima.sim(n = 1000, list(ar = c(0.3, -0.7)) ))  
# We use here a simplified way to create X and Y   
# since the focus is on creating the RNNs  
data = matrix(data, ncol = 10L, byrow = TRUE)  
X = array(data[seq(1, 100, by = 2), ], dim = c(50, 10, 1))  
Y = data[seq(2, 100, by = 2), ]  
  
RNN =   
 keras\_model\_sequential() %>%   
 # first hidden layer  
 layer\_gru(input\_shape = list(10L, 1L),  
 units = 50,   
 activation = "relu") %>%  
 # we want to predict the next 10 time steps  
 layer\_dense(units = 10)  
  
  
# add loss function and optimizer  
RNN %>%   
 compile(loss = loss\_mean\_squared\_error,  
 optimizer = optimizer\_adamax(0.01))  
  
RNN %>%   
 fit(X, Y, epochs = 50, verbose = 0)

Make predictions:

head(predict(RNN, X), n = 3)

#### 2.8.2 Python

from tensorflow import keras  
from tensorflow.keras.layers import \*  
X = r.X # get data from R  
Y = r.Y   
  
RNN = keras.model\_sequential()  
 # first hidden layer  
RNN.add(GRU(input\_shape = [10, 1],units = 50, activation = "relu"))  
RNN.add(Dense(units = 10))  
  
RNN.summary()  
  
# add loss function and optimizer  
RNN.compile(loss = keras.losses.mean\_squared\_error,  
 optimizer = keras.optimizer.adamax(0.01))  
  
# train model  
RNN.fit(X, Y, epochs = 50, verbose = 0)

Make predictions:

RNN.predict(X)[0:10,:]

### 2.9 Graph (convolutional) neural networks

Graph neural networks (GNN) is a young representative of the deep neural network family but is receiving more and more attention in the last years because of their ability to process non-Euclidean data such as graphs.

Currently there is no R package for GNNs available. However, we can use the ‘reticulate’ package (Ushey, Allaire, and Tang (2022)) to use the python packages ‘torch’ and ‘torch\_geometric’ (Paszke et al. (2019), 2019; Fey and Lenssen (2019)).

The following example was mostly adapted from the ‘Node Classification with Graph Neural Networks’ example from the torch\_geometric documentation (<https://pytorch-geometric.readthedocs.io/en/latest/notes/colabs.html>).

The dataset is also provided by the ‘torch\_geometric’ package and consists of molecules presented as graphs and the task is to predict whether HIV virus replication is inhibited by the molecule or not (classification, binary classification).

#### 2.9.1 R

library(reticulate)  
# Load python packages torch and torch\_geometric via the reticulate R package  
torch = import("torch")   
torch\_geometric = import("torch\_geometric")  
  
# helper functions from the torch\_geometric modules  
GCNConv = torch\_geometric$nn$GCNConv  
global\_mean\_pool = torch\_geometric$nn$global\_mean\_pool  
  
  
# Download the MUTAG TUDataset  
dataset = torch\_geometric$datasets$TUDataset(root='data/TUDataset',   
 name='MUTAG')  
dataloader = torch\_geometric$loader$DataLoader(dataset,   
 batch\_size=64L,  
 shuffle=TRUE)  
  
# Create the model with a python class  
# There are two classes in the response variable  
GCN = PyClass(  
 "GCN",   
 inherit = torch$nn$Module,   
 defs = list(  
 `\_\_init\_\_` = function(self, hidden\_channels) {  
 super()$`\_\_init\_\_`()  
 torch$manual\_seed(42L)  
 self$conv = GCNConv(dataset$num\_node\_features, hidden\_channels)  
 self$linear = torch$nn$Linear(hidden\_channels, dataset$num\_classes)  
 NULL  
 },  
 forward = function(self, x, edge\_index, batch) {  
 x = self$conv(x, edge\_index)  
 x = x$relu()  
 x = global\_mean\_pool(x, batch)  
   
 x = torch$nn$functional$dropout(x, p = 0.5, training=self$training)  
 x = self$linear(x)  
 return(x)  
 }  
 ))

Training loop:

# create model object  
model = GCN(hidden\_channels = 64L)  
  
# get optimizer and loss function  
optimizer = torch$optim$Adamax(model$parameters(), lr = 0.01)  
loss\_func = torch$nn$CrossEntropyLoss()  
  
# set model into training mode (because of the dropout layer)  
model$train()  
  
# train model  
for(e in 1:50) {  
 iterator = reticulate::as\_iterator(dataloader)  
 coro::loop(for (b in iterator) {   
 pred = model(b$x, b$edge\_index, b$batch)  
 loss = loss\_func(pred, b$y)  
 loss$backward()  
 optimizer$step()  
 optimizer$zero\_grad()  
 })  
 if(e %% 10 ==0) cat(paste0("Epoch: ",e," Loss: ", round(loss$item()[1], 4), "\n"))  
}

Make predictions:

preds = list()  
test = torch\_geometric$loader$DataLoader(dataset, batch\_size=64L,shuffle=FALSE)  
iterator = reticulate::as\_iterator(test)  
model$eval()  
counter = 1  
coro::loop(for (b in iterator) {  
 preds[[counter]] = model(b$x, b$edge\_index, b$batch)  
 counter <<- counter + 1  
 })  
head(torch$concat(preds)$sigmoid()$data$cpu()$numpy(), n = 3)

#### 2.9.2 Python

# Load python packages torch and torch\_geometric via the reticulate R package  
import torch  
import torch\_geometric  
  
# helper functions from the torch\_geometric modules  
GCNConv = torch\_geometric.nn.GCNConv  
global\_mean\_pool = torch\_geometric.nn.global\_mean\_pool  
  
  
# Download the MUTAG TUDataset  
dataset = torch\_geometric.datasets.TUDataset(root='data/TUDataset',   
 name='MUTAG')  
dataloader = torch\_geometric.loader.DataLoader(dataset,   
 batch\_size=64,  
 shuffle=True)  
  
# Create the model with a python class  
# There are two classes in the response variable  
class GCN(torch.nn.Module):  
 def \_\_init\_\_(self, hidden\_channels):  
 super().\_\_init\_\_()  
 torch.manual\_seed(42)  
 self.conv = GCNConv(dataset.num\_node\_features, hidden\_channels)  
 self.linear = torch.nn.Linear(hidden\_channels, dataset.num\_classes)  
   
 def forward(self, x, edge\_index, batch):  
 x = self.conv(x, edge\_index)  
 x = x.relu()  
 x = global\_mean\_pool(x, batch)  
 x = torch.nn.functional$dropout(x, p = 0.5, training=self.training)  
 x = self.linear(x)  
 return(x)

Training loop:

# create model object  
model = GCN(hidden\_channels = 64)  
  
# get optimizer and loss function  
optimizer = torch.optim.Adamax(model.parameters(), lr = 0.01)  
loss\_func = torch.nn.CrossEntropyLoss()  
  
# set model into training mode (because of the dropout layer)  
model.train()  
  
# train model  
for e in 0:50:  
 iterator = reticulate::as\_iterator(dataloader)  
   
 for b in dataloader:  
   
 pred = model(b$x, b$edge\_index, b$batch)  
 loss = loss\_func(pred, b$y)  
 loss.backward()  
 optimizer.step()  
 optimizer.zero\_grad()  
   
 if e % 10 ==0:  
 print("Epoch: ", e," Loss: ", loss.item()[0], "\n"))

Make predictions:

preds = []  
test = torch\_geometric.loader.DataLoader(dataset, batch\_size=64,shuffle=False)  
model.eval()  
counter = 1  
for b in test:  
 preds.append( model(b.x, b.edge\_index, b.batch) )  
   
   
torch.concat(preds).sigmoid().data.cpu().numpy()[0:10]

## References

Abadi, Martı́n, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. “Tensorflow: A System for Large-Scale Machine Learning.” In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–83.

Allaire, JJ, and François Chollet. 2022. *Keras: R Interface to ’Keras’*. <https://CRAN.R-project.org/package=keras>.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2022. *Xgboost: Extreme Gradient Boosting*. <https://CRAN.R-project.org/package=xgboost>.

Chollet, Francois et al. 2015. “Keras.” GitHub. 2015. <https://github.com/fchollet/keras>.

Feinerer, Ingo, Kurt Hornik, and David Meyer. 2008. “Text Mining Infrastructure in r.” *Journal of Statistical Software* 25 (5): 1–54. <https://www.jstatsoft.org/v25/i05/>.

Fellows, Ian. 2018. *Wordcloud: Word Clouds*. <https://CRAN.R-project.org/package=wordcloud>.

Fey, Matthias, and Jan E. Lenssen. 2019. “Fast Graph Representation Learning with PyTorch Geometric.” In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Jahn, Najko. 2021. *Europepmc: R Interface to the Europe PubMed Central RESTful Web Service*. <https://CRAN.R-project.org/package=europepmc>.

Lang, Dawei, and Guan-tin Chien. 2018. *Wordcloud2: Create Word Cloud by ’Htmlwidget’*. <https://CRAN.R-project.org/package=wordcloud2>.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44.

LeCun, Yann, Corinna Cortes, and CJ Burges. 2010. “MNIST Handwritten Digit Database.” *ATT Labs [Online]. Available: Http://Yann.lecun.com/Exdb/Mnist* 2.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2022. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. <https://CRAN.R-project.org/package=e1071>.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems 32*, 8024–35. Curran Associates, Inc. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

Schliep, Klaus, and Klaus Hechenbichler. 2016. *Kknn: Weighted k-Nearest Neighbors*. <https://CRAN.R-project.org/package=kknn>.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent.” *Journal of Statistical Software* 39 (5): 1–13. <https://doi.org/10.18637/jss.v039.i05>.

Ushey, Kevin, JJ Allaire, and Yuan Tang. 2022. *Reticulate: Interface to ’Python’*. <https://CRAN.R-project.org/package=reticulate>.

Wright, Marvin N., and Andreas Ziegler. 2017. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77 (1): 1–17. <https://doi.org/10.18637/jss.v077.i01>.