Data mining and Machine Learning
Supervised and Unsupervised methodologies for environmental data.

- Elena Grimaccia

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

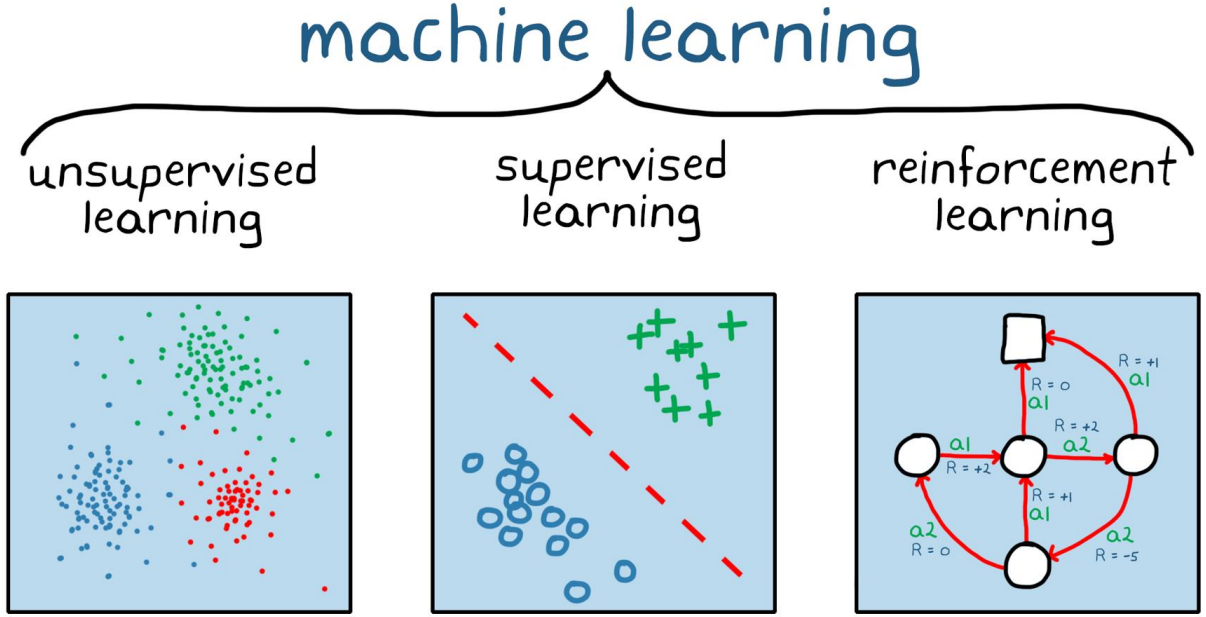
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Machine Learning Methods

Process of predicting Environmental output using machine/deep learning models. It involves data pre-processing, transformation into feature vectors, model training, and prediction using test data:

Step-by-Step Process

1. Environmental Dataset
2. Pre-processing:
   - Data Cleaning
   - Data Transformation
3. Pollutants/Environmental Characteristics → Feature Vector
4. Apply Desired Machine/Deep Learning Model
5. Train Prediction Model
6. Input: Environmental Test Data
7. Test Data → Feature Vector
8. Prediction Model (uses trained model + test feature vector)
9. Output: Outcome Predicted



The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite set (survived/died, digit 0-9, cancer class of tissue sample).
- We have data (x1; y1); : : : ; (xN; yN). These are observations (examples,examples, instances) of these measurements.

The Supervised Learning Problem

- Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective: find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- different from supervised learning, but can be useful as well.

Supervised Method: regression

The regression function f(x)

We can refer to the input vector as
$X = (X1, X2, ..., Xk)$
These are our features

Now we write our model as
$Y = f(X) + \varepsilon$
where ε captures measurement errors and other discrepancies.
Depending on the complexity of f, we may be able to understand how each component Xj of X affects Y .

Supervised Method: regression

The regression function f(x)

Is also defined for vector X; e.g.

$f(x) = f(x1,x2,x3) = E(Y|X1 = x1,X2 = x2,X3 = x3)$

Is the ideal or optimal predictor of $Y$with regard to mean-squared prediction error:

$f(x) = E(Y|X = x)$ is the function that minimizes

$$E(Y - f(X))^2 = E[f(X) + e - \hat{f}(X)]^2$$
$$= [f(X) - \hat{f}(X)]^2 + Var(e)$$

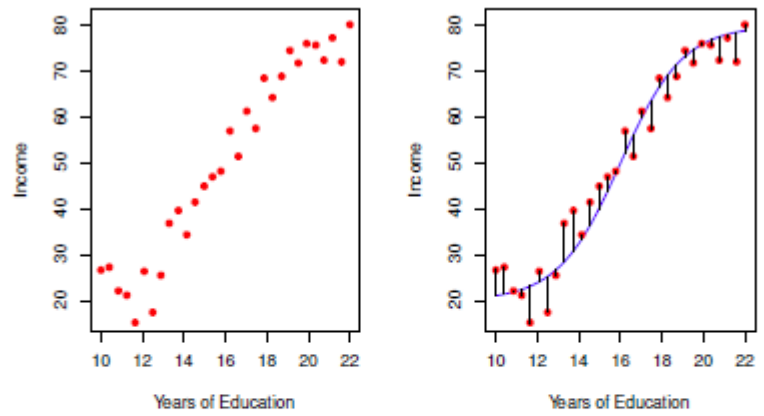


FIGURE 2.2. *The* Income *data set. Left: The red dots are the observed values of* income *(in tens of thousands of dollars) and* years of education *for* 30 *individuals. Right: The blue curve represents the true underlying relationship between* income *and* years of education*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

Supervised Method: linear regression

When the regression function f(x) is a line, we call this simple relationships between Y and X: linear regression.
Y = $\beta_0 + \beta_1 X$

We estimate β_0 *and* β_1 by miminasing the sum of residuals or, more precisely, thee Residual Sum of Squares,and we obtain:

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means. In other words, (3.4) defines the *least squares coefficient estimates* for simple linear regression.

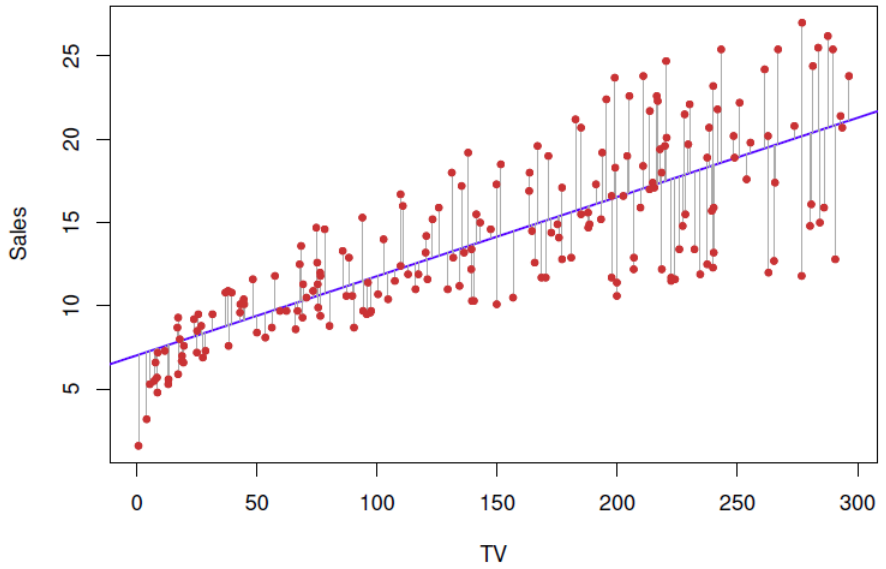


FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Accuracy of the model
Goodness of fit

We assess the accuracy of our estimates of the coefficients beta 0 and beta 1, employing standard errors, and p-values:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* $1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars.)*

Accuracy of the model
Goodness of fit

We assess the accuracy of our estimates of the coefficients beta 0 and beta 1, employing the

Mse= Mean Square Error = $\frac{1}{n}\ \sum(y_{i-}\ \widehat{f\ (x_i)}\ )^2$

Accuracy of the model
Goodness of fit

We assess the accuracy of our estimates with the R2 statistic.
It takes the form of a proportion — the proportion of variance explained — and so it always takes on a value between 0 and 1, and is independent of the scale of Y .

To calculate R^2, we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad (3.17)$$

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*, and RSS is defined $\text{RSS} = \sum(y_i - \hat{y}_i)^2$.

Supervised Method: Linear regression

Limits of linear regression:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

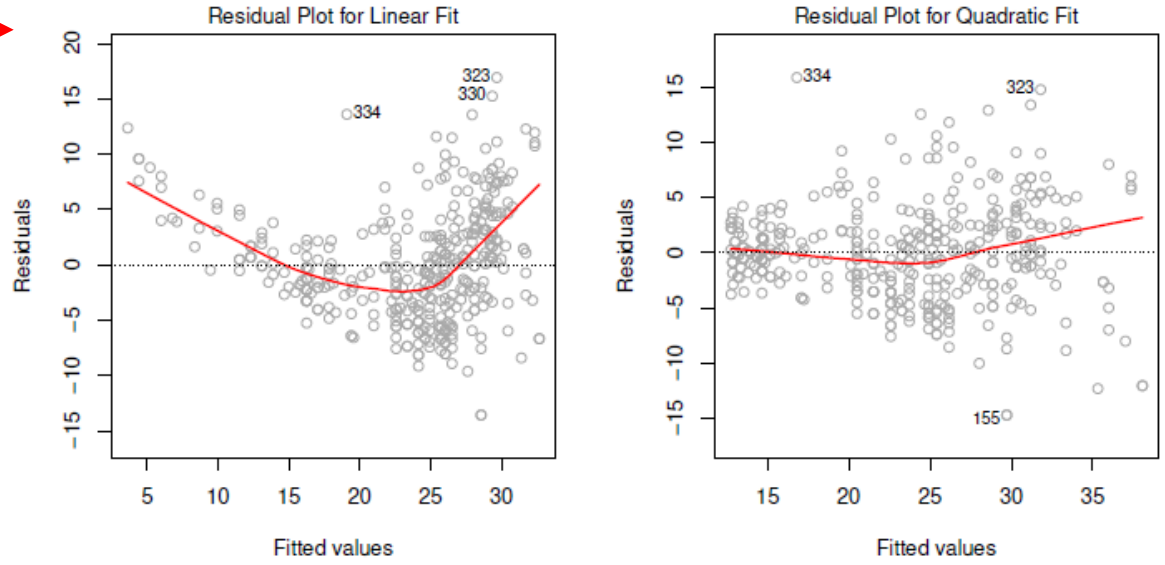


FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto` *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* `mpg` *on* `horsepower`. *A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* `mpg` *on* `horsepower` *and* $\texttt{horsepower}^2$. *There is little pattern in the residuals.*

Supervised Method: Linear regression

Limits of linear regression:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.

(Heteroskedasticity)

4. Outliers.
5. High-leverage points.
6. Collinearity.

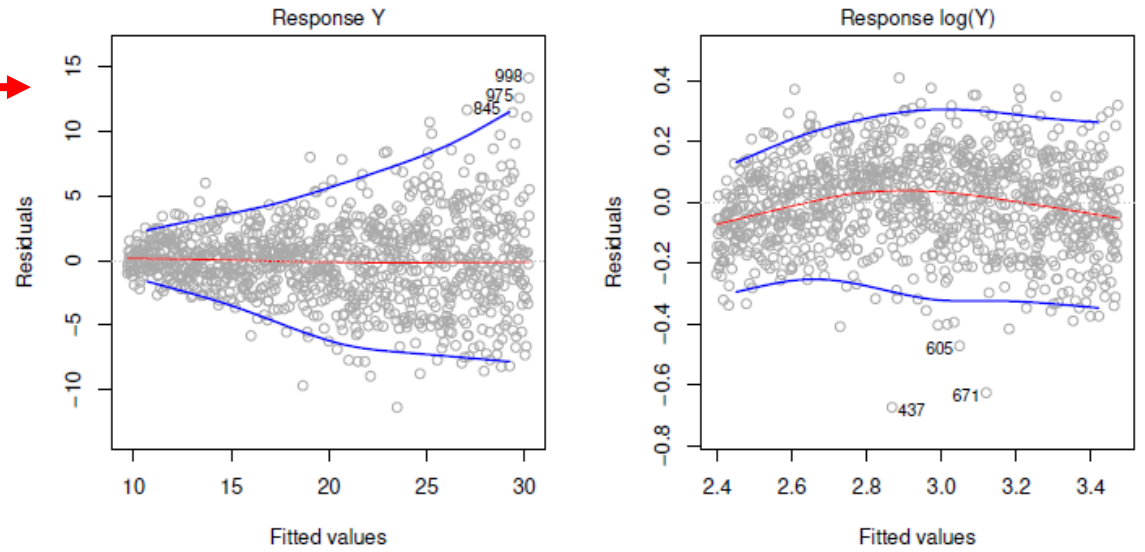


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.*

Supervised Method: Linear regression

Limits of linear regression:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error
4. Outliers.
5. High-leverage points.
6. Collinearity.

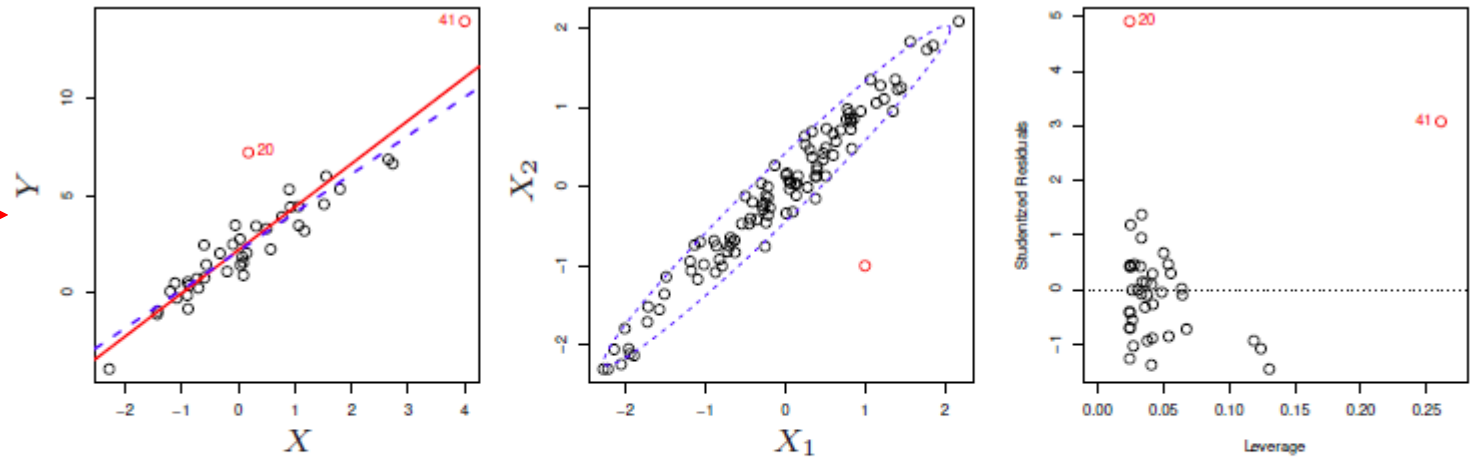


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation 41 has a high leverage and a high residual.*

Supervised Method: Linear regression extensions

We can also have:

- qualitative predictors (factors)
  - with only 2 levels (dummy variables)
  - with More than Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African Ameri} \end{cases} \quad (3.30)$$

Now β_0 can be interpreted as the average credit card balance for African Americans, β_1 can be interpreted as the difference in the average balance between the Asian and African American categories, and β_2 can be interpreted as the difference in the average balance between the Caucasian and African American categories. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the *baseline*.

Supervised Method: Linear regression extensions

We can also:

- remove the additive assumption: include interactions in the model: $Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X1X2 + \varepsilon$
  - hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
- Consider Non-Linear Relationships: $Y = \beta0 + \beta1X1 + \beta2X2 + \beta3(X2)^2 + \varepsilon$

Supervised Method: Multiple linear regression

We can extend our results to a multiple regression:

$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \ldots + \varepsilon$

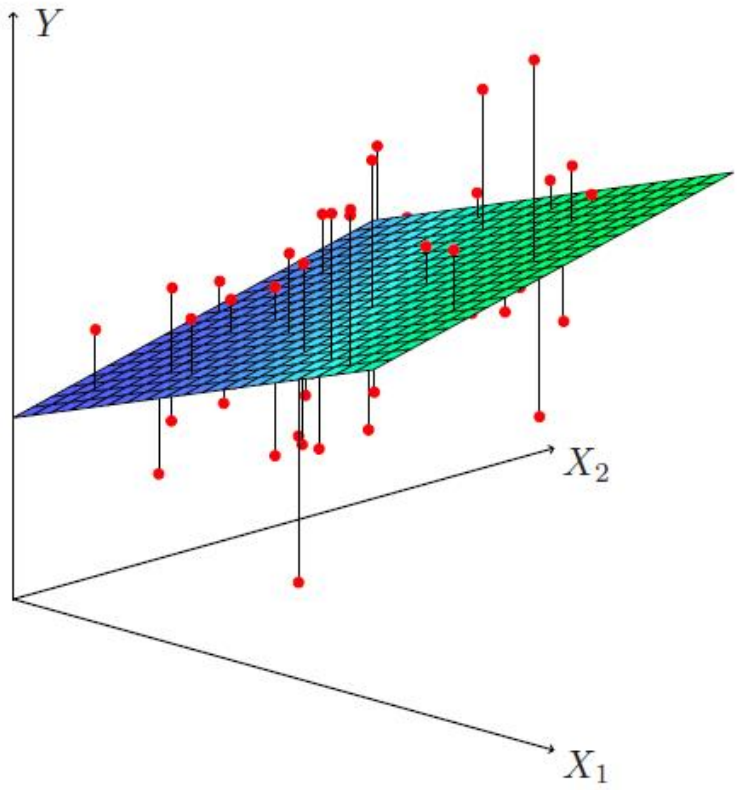


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

Supervised Method: Multiple Linear regression

Best Subset Selection
Forward stepwise selection
is a computationally efficient alternative to best
forward stepwise
subset selection.

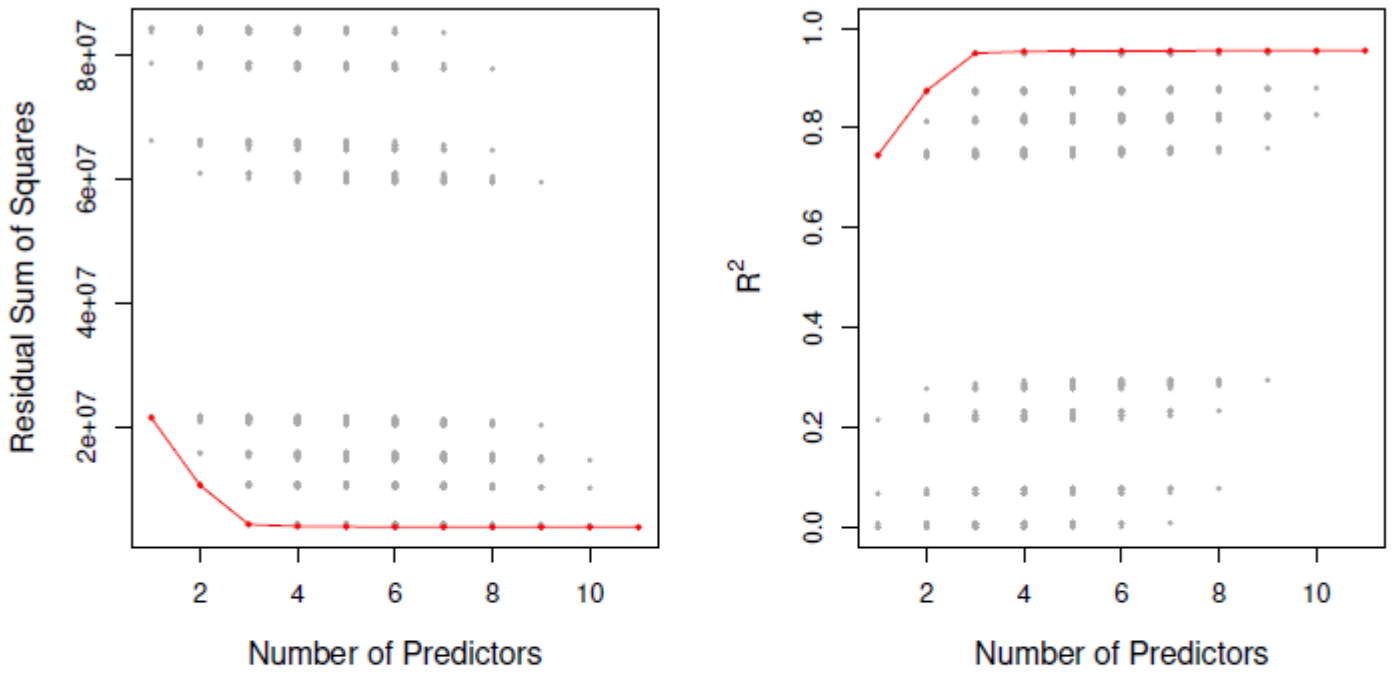


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Supervised Method: Multiple Linear regression

Subset Selection
Backwards stepwise selection

Algorithm 6.3 *Backward Stepwise Selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Supervised Method: Multiple Linear regression

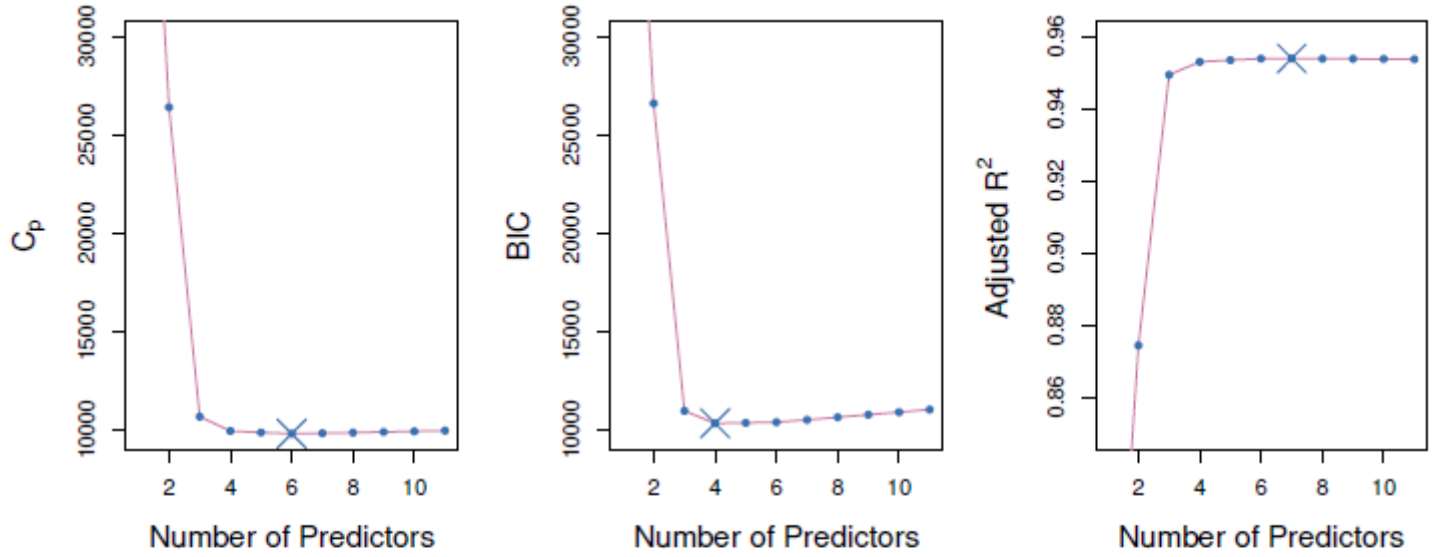


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

Supervised Method: Classification

The response variable in these models is instead qualitative.

- The most widely-used classifiers:
  - logistic regression,
  - linear discriminant analysis,
  - K-nearest neighbors.

Why not linear regression?

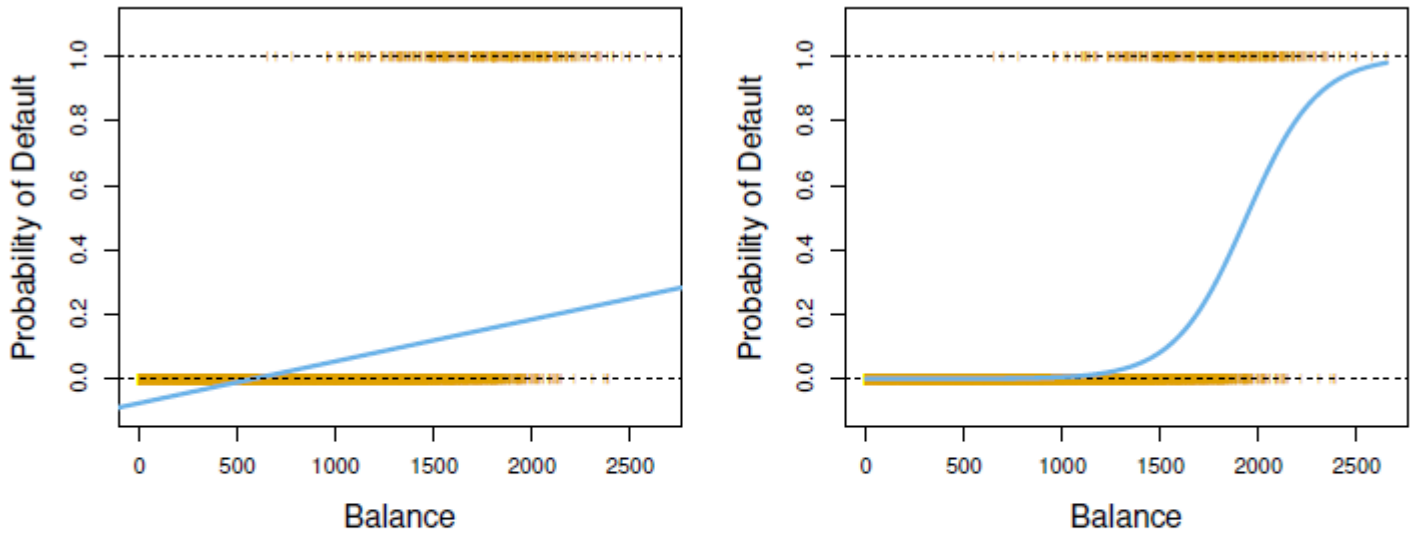


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic regression

By taking the logarithm of both sides of (4.3), we arrive at

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

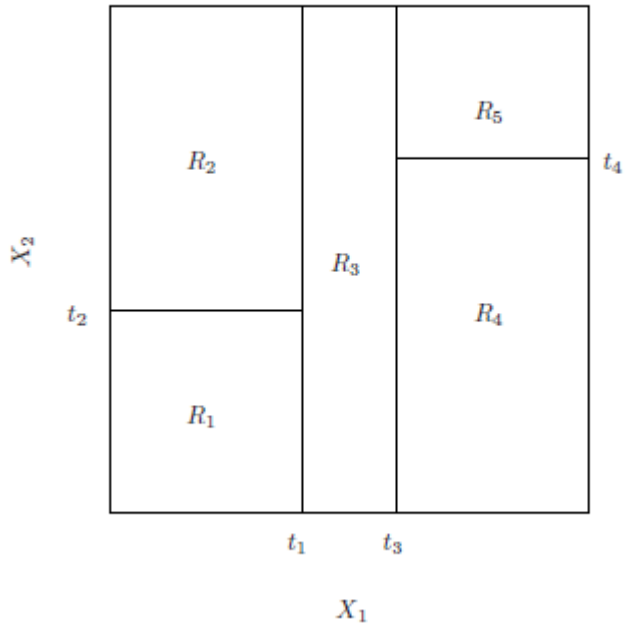
	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

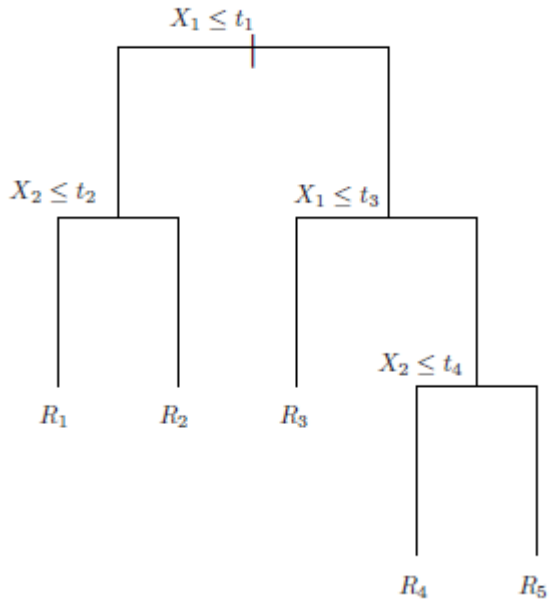
TABLE 4.3. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`, `income`, *and student status. Student status is encoded as a dummy variable* `student[Yes]`, *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* `income` *was measured in thousands of dollars.*

Tree – based analysis

The output of recursive binary splitting on a two-dimensional example.

A tree corresponding to the partition in the top right panel.





Algorithm 8.1 *Building a Regression Tree*

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. For each $k = 1, \ldots, K$:
   (a) Repeat Steps 1 and 2 on the $\frac{K-1}{K}$th fraction of the training data, excluding the kth fold.
   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for:

- Classification (most common),
- and sometimes regression tasks.

- It works by finding the optimal boundary (hyperplane) that best separates data points of different classes.
- Example: Two classes of points in 2D space:
  - Class A (e.g., "low obesity")
  - Class B (e.g., "high obesity")
- SVM tries to draw a line (or more generally, a hyperplane) that separates these two classes with the largest possible margin.

Support vector machine (SVM)

Key Terms:

- Hyperplane: The decision boundary that separates classes:
  - In 2D: a line.
  - In 3D: a plane.
  - In higher dimensions: a "hyperplane.“
- Margin: The distance between the hyperplane and the closest data points from each class. SVM maximizes this margin.
- Support Vectors: The data points that lie closest to the hyperplane. They are the "support" for the decision boundary — the SVM would change if you removed them.

Support vector machine (SVM)

SVM tries to find the best hyperplane that:

- Completely separates the classes (if possible),
- Maximizes the margin between them.

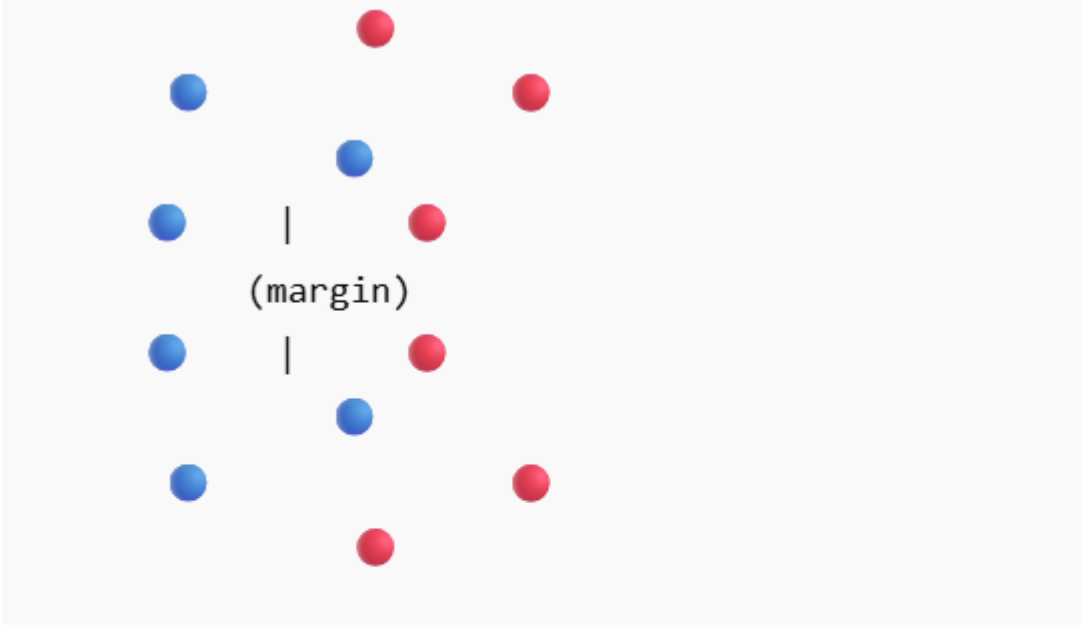
SVM solves a constrained optimization problem (quadratic programming) to do this.

- Once trained, it uses the hyperplane to classify new data points.
- If Data are Not Linearly Separable: the classes are mixed in complex ways: SVM uses a kernel function to project the data into a higher-dimensional space, where it is separable:
  - Linear kernel: no transformation; used when data is linearly separable.
  - Polynomial kernel: for curved boundaries.
  - Radial Basis Function (RBF) / Gaussian kernel: most common; captures complex, nonlinear relationships.
  - Sigmoid, similar to neural networks.

Support vector machine (SVM)

Output:

- After training, an SVM model can:
  - Predict the class of a new observation,
  - Estimate probabilities (if enabled),
  - Provide decision values (confidence of classification),
  - Highlight support vectors (important observations).



The vertical line is the **hyperplane**.

The horizontal space around it is the **margin**.

The closest 🔵 and 🔴 define the **support vectors**.

References

James, Witten, Hastie, Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer 2013
Here you can download the book: https://www.statlearning.com/
With applications in python too

Hastie, Tibshirani, and Friedman (2009)
The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)
Free PDF: https://web.stanford.edu/~hastie/ElemStatLearn/

Richard S. Sutton and Andrew G. Barton (2014, 2015)
Reinforcement Learning: An Introduction
Second edition, in progress
A Bradford Book
The MIT Press
Free PDF: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

More in detail:

- Based on what we have discussed, which methodology you imagine to employ on your data?

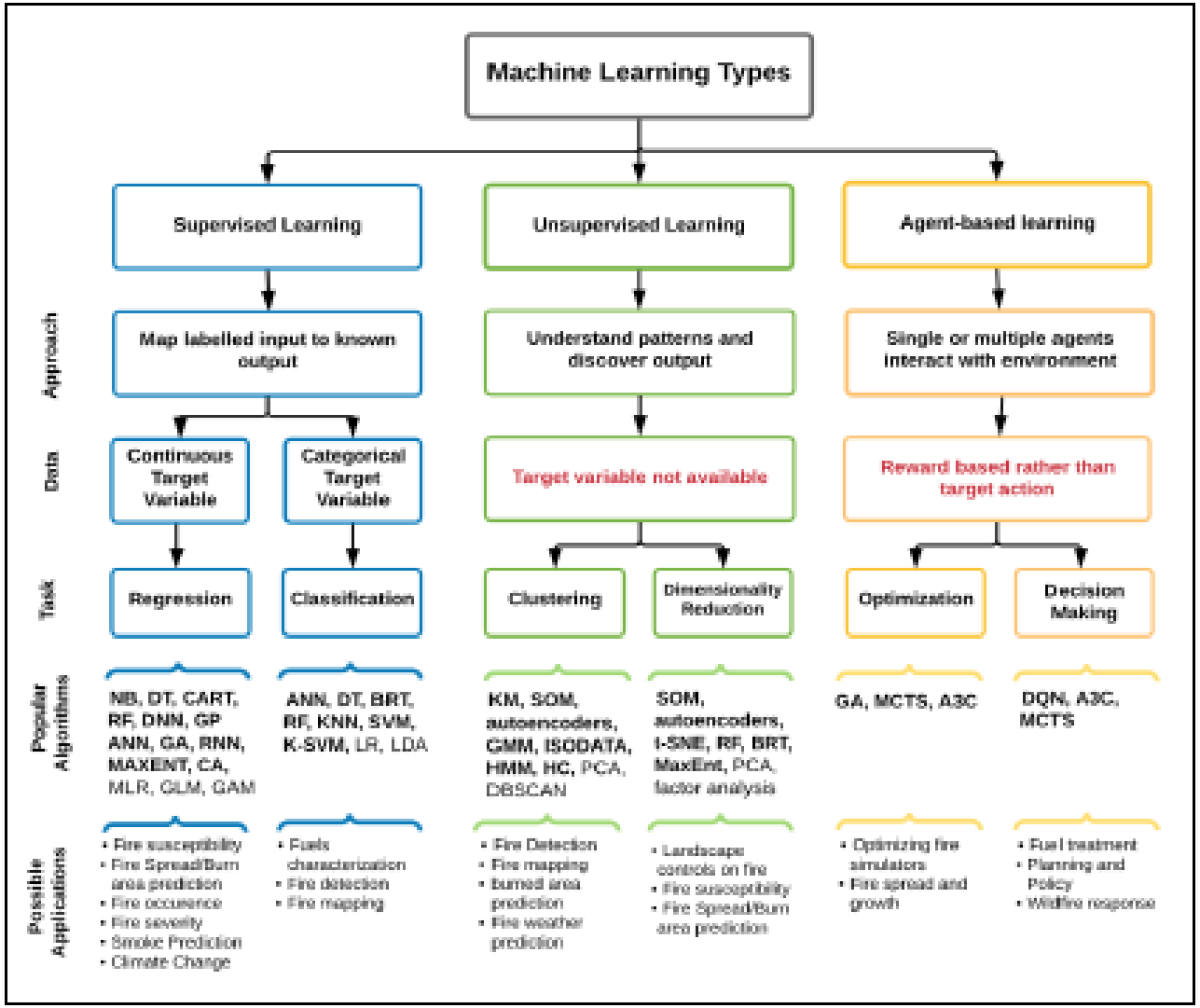


Fig 5: *diagram depicting the primary types of machine learning, data types, and modeling tasks, highlighting their associations with widely used algorithms and applications in wildfire science and management. Algorithms in bold indicate core ML methods, whereas non-bolded algorithms are generally not classified as core ML (Piyush et al., 2020).*

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani