Machine Learning

Emanuele Aliverti
University of Padova, Department of Statistical Sciences

emanuelealiverti.github.io
aliverti@stat.unipd.it

INFN School of Statistics 2024

Università degli Studi di Padova

Main References

Università degli Studi di Padova

AS2012 Azzalini A., Scarpa B. (2012). *Data analysis and data mining*. Oxford University Press.

ISL2023 James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer Nature.

HTF2011 Hastie T., Tibshirani R., Friedman J. (2011). *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag, Heidelberg & New York.

Introduction - supervised learning, machine learning

	References
AS2012	Chapter 1,2
ISL2023	Chapter 1, 2.1

Supervised learning

- Response variable y (also called *dependent variable*, *outcome measurement*, *target*)
- Explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ (also called *predictor measurements*, *regressor*, *covariate*, *input*, *feature*, *independent variable*)
- In the regression problem, y is quantitative (sales, blood pressure, price, income, . . .)
- In the classification problem, y is categorical and takes values in a finite, unordered set (churner/not churner, dead/alive, signal/background, {cat, dog, mouse, . . . }, . . .)
- We have a sample of data $(x_1, y_1), \ldots, (x_n, y_n)$ which are observations (examples, instances) of these measurements

Objectives

On the basis of the available 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

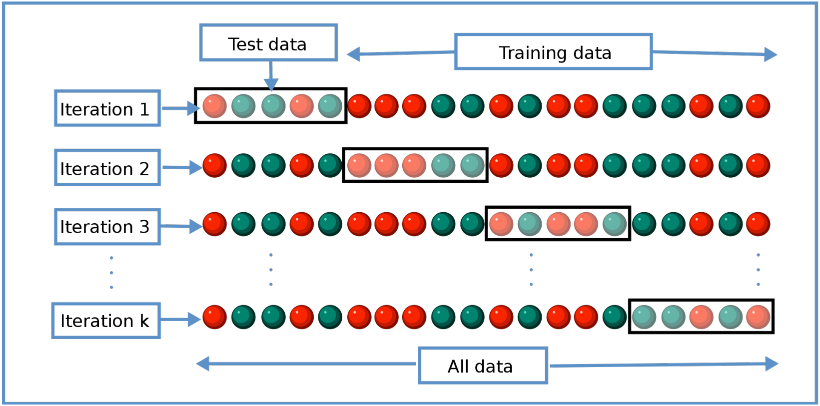
Examples – supervised learning

- Predict whether a customer of a telecommunication company will abandon the company to go to the competitors, on the basis of traffic usage, demographics and services subscriptions;
- Identify the numbers in a handwritten zip code, from a digitised image
- Estimate the probability that an insurance customer will buy another product of the same company, on the basis of client demographics and claims history
- (from Higgs Boson Machine Learning Challenge, the ATLAS experiment at CERN) The dataset comprises features derived from simulated particle collisions, with the goal of distinguishing signal events (containing evidence of the Higgs boson) from background noise.

Bias-variance trade off

	References
AS2012	Chapter 3
ISL2023	Chapter 2.2, 2.3

In figures



Source: wikipedia

Non-parametric models: regression trees

References

AS2012 Chapter 4.8 (regression), 5, 5.7

ISL2023 Chapter 8.1

MARS

	References
AS2012	Chapter 4.4.5
HTF2011	Chapter 8.9

Classification trees, combination of classifiers

	References
AS2012	Chapter 5.9
ISL2023	Chapter 8.2

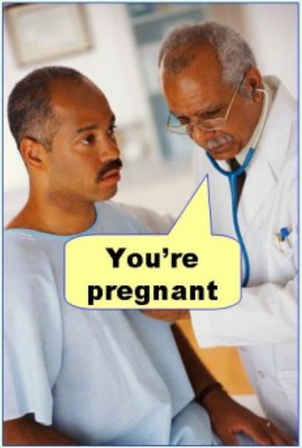
Misclassification tables

- *given* a threshold separating between two classes
- Misclassification table or confusion matrix

Prediction	Actual response		
	$-$	$+$	total
$-$	n_{11}	n_{12}	$n_{1\cdot}$
$+$	n_{21}	n_{22}	$n_{2\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Graphically

Type I error
(false positive)



Type II error
(false negative)

Misclassification tables

- *given* a threshold separating between two classes
- Misclassification table or confusion matrix

Prediction	Actual response		
	$-$	$+$	total
$-$	n_{11}	n_{12}	$n_{1\cdot}$
$+$	n_{21}	n_{22}	$n_{2\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

- Table of probability errors

Prediction	Actual response	
	$-$	$+$
$-$	$1-\alpha$	β
$+$	α	$1-\beta$
total	1	1

Probability error

Prediction	Actual response	
	$-$	$+$
$-$	$1-\alpha$	β
$+$	α	$1-\beta$
total	1	1

Probability error

Prediction	Actual response	
	$-$	$+$
$-$	$1-\alpha$	β
$+$	α	$1-\beta$
total	1	1

$$\text{sensitivity} = 1-\beta, \qquad 1-\hat{\beta} = \frac{n_{22}}{n_{12}+n_{22}}$$

Probability error

Prediction	Actual response	
	$-$	$+$
$-$	$1-\alpha$	β
$+$	α	$1-\beta$
total	1	1

$$\text{sensitivity} = 1-\beta, \qquad 1-\hat{\beta} = \frac{n_{22}}{n_{12}+n_{22}}$$

$$\text{specificity} = 1-\alpha, \qquad 1-\hat{\alpha} = \frac{n_{11}}{n_{11}+n_{21}}$$

Note: all these concepts refers to a *fixed* threshold between the two classes

Model averaging

Classification trees can be simple, but often produce noisy (bushy) is weak (stunted) classifiers

- Bagging *(Breiman, 1996)*: Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote
- Boosting *(Freund & Shapire, 1996)*: Fit many large or small trees to *reweighted* versions of the training data. Classify by weighted majority vote
- Random Forests *(Breiman 1999)*: Fancier version of bagging.

In general *Boosting* $\succ$ *Random Forests* $\succ$ *Bagging* $\succ$ *single tree*.