Data Mining

Anno Accademico 2022/23
CLAMSES - Università Milano-Bicocca

Aldo Solari
aldo.solari@unimib.it

- Insegnamento **DATA MINING M** (6 CFU)
- Insegnamento DATA SCIENCE M (12 CFU)
  - Modulo **DATA MINING** (6 CFU)
  - Modulo STATISTICAL LEARNING (6 CFU)

Che differenza c'è tra Data Science, Data Mining e Statistical Learning?

Pagine

MOODLE: https://elearning.unimib.it/course/view.php?id=44915

- Syllabus
- Avvisi
- Forum
- Spazio per la consegna
- Pubblicazione voti

WEB: https://aldosolari.github.io/DM/

- Calendario
- Materiale: slide, codice R, esercizi
- Libri di testo
- Informazioni per l'esame

Esame

La modalità di verifica consiste nell'analisi di un dataset e in una prova orale.

L'analisi dei dati ha come obiettivo la verifica delle abilità di modellizzazione dei dati a fini previsivi.

La prova orale riguarderà sia l'esposizione dell'analisi dei dati sia la verifica dello studio degli argomenti trattati a lezione.

Analisi dei dati

Per l'analisi dei dati, oltre alle previsioni, bisognerà produrre una relazione contenete la descrizione dell'analisi e il codice utilizzato.

Bisogna consegnare l'analisi dei dati sulla pagina moodle del corso almeno una settimana prima dell'appello d'esame.

E' possibile consegnare l'analisi dei dati una volta sola per A.A.

Il dataset verrà reso disponibile a fine Dicembre. Verrà data la possibilità di svolgere l'analisi dei dati in gruppi di (al massimo) tre persone (formati in maniera casuale).

Libri di testo

Azzalini, Scarpa (2012). Data analysis and data mining: an introduction. Oxford University Press. [AS].

Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning. Springer. [HTF].

Approfondimenti

Azzalini, Scarpa (2004). Analisi dei dati e data mining, Springer-Verlag Italia.

Lewis, Kane, Arnold (2019). A Computational Approach to Statistical Learning. Chapman And Hall/Crc. [LKA].

Kuhn, Johnson (2019). Feature Engineering and Selection. `http://www.feat.engineering/`. Chapman and Hall/CRC. [KJ].

Kuhn, Silge (2022). Tidy Modeling with R. `https://www.tmwr.org/`. O'Reilly Media, Inc. [KS].

Data Mining, Data Science, Statistical Learning, Machine Learning

Friedman, Jerome (1997)
Data Mining and Statistics: What's the connection?
http://www.stats.org.uk/Friedman1997.pdf

Larry Wasserman (2012)
https://normaldeviate.wordpress.com/2012/06/12/
statistics-versus-machine-learning-5-2/

Donoho, David (2017).
50 years of data science.
Journal of Computational and Graphical Statistics 26: 745-766

Robert Tibshirani (2015?)

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Trevor Hastie (2015)

Some Definitions

Machine Learning constructs algorithms that can learn from data.

Statistical Learning is a branch of applied statistics that emerged in response to machine learning, emphasizing statistical models and assessment of uncertainty.

Data Science is the extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer science, engineering, ...

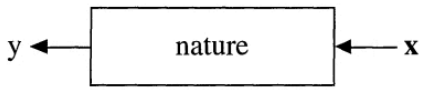
All of these are very similar — with different emphases.

Applied Statistics?

Leo Breiman (2001).
Statistical Modeling: The Two Cultures.
Statistical Science, 16:199-231

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

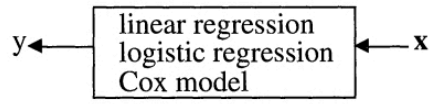
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

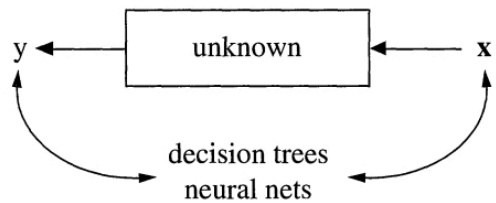
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



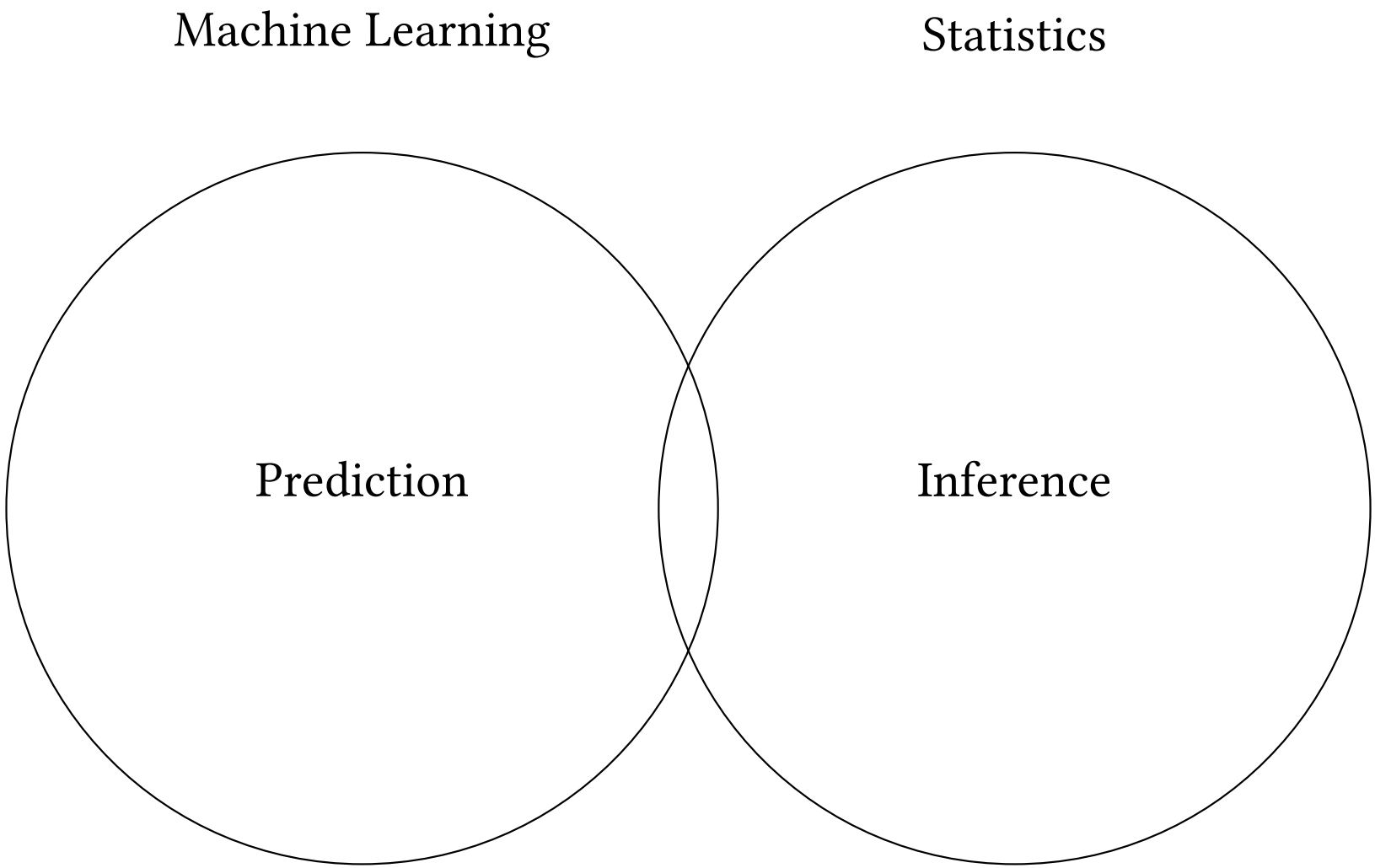
Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.


Machine Learning
Statistics
Prediction
Inference

Machine Learning	Statistics
prediction	inference
focus	
optimization/algorithmic	modelling
culture	
decision trees	linear regression
k-nearest-neighbors	discriminant analysis
neural networks	logistic regression
support vector machines	ridge regression
random forests	
adaboost	lasso
...	...
methods	

Spiegare o prevedere?

Supponiamo che il “vero” modello sia il seguente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Consideriamo ora il seguente modello “sbagliato” (sotto-specificato)

$$Y \approx \gamma_0 + \gamma_1 X_1 + \epsilon$$

Per ottenere una buona spiegazione dobbiamo stimare i coefficienti del “vero” modello, tuttavia a volte un modello ”sbagliato” può prevedere meglio Y

Galit Shmueli (2010)
To explain or to predict?
Statistical Science, 25: 289-310

```
simulation <- function(x1,x2){
n <- length(x1)
y <- 2 + x1 + x2 + rnorm(n)
fit_correct <- lm(y ~ x1 + x2)
fit_wrong <- lm(y ~ x1)
y_new <- 2 + x1 + x2 + rnorm(n)
MSE_correct <- mean( (predict(fit_correct) - y_new)^2 )
MSE_wrong <- mean( (predict(fit_wrong) - y_new)^2 )
return(c(MSE_correct, MSE_wrong))}
B = 1000
set.seed(123)
n <- 10
x1 <- rnorm(n)
x2 <- rnorm(n, x1, 0.01)
res = replicate(B, simulation(x1,x2))
row.names(res) <- c("MSE_correct", "MSE_wrong")
rowMeans(res)
MSE_correct   MSE_wrong
   1.310599    1.219693
```

Dalla statistica classica alla statistica moderna

- Statistica Classica
  - Analisi multivariata
    I libri di Anderson (1958) e di Mardia, Kent & Bibby (1979)
  - Modelli statistici
    L'articolo di Nelder & Wedderburn (1972) che introduce i GLM
- Statistica Computer-Age
  - Data Mining
  - Machine Learning
- Statistica Moderna
  - Statistical Learning
    Il libro di Hastie, Tibshirani & Friedman (2001)
  - Data Science

Carl Friedrich Gauss (1777 - 1855) è uno scienziato dei dati?

Problema astronomico

Prevedere in anticipo la posizione dell'asteroide Ceres in data 31 dicembre 1801 sulla base dei dati forniti dall'astronomo italiano Giuseppe Piazzi

Problema statistico

Determinare $\boldsymbol{\beta}$ tale che minimizzi $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

La soluzione di Gauss: metodo dei minimi quadrati
$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

Problema computazionale

Risolvere (a mano!) il sistema di equazioni $\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$

La soluzione di Gauss: algoritmo di ottimizzazione (metodo di eliminazione di Gauss)