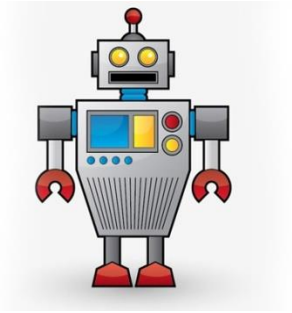
Machine learning – ML. Introducere

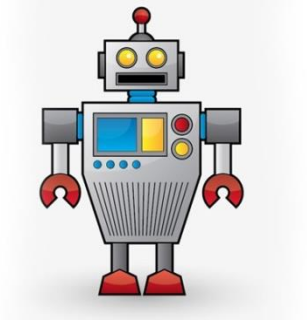
Ruxandra Stoean
rstoean@inf.ucv.ro
http://inf.ucv.ro/~rstoean

Masini inteligente

- Masini care pot fi programate sa actioneze ca oamenii



- Masina care pot fi instruite sa invete ca oamenii



Definitie

- Conform Wikipedia:
  - “ML este o disciplina stiintifica care exploreaza constructia si studiul algoritmilor care invata din date [1].
  - Astfel de algoritmi opereaza prin constructia unui model care se bazeaza pe date de intrare si il foloseste pentru a face mai degraba predictii sau decizii decat a urmari instructiuni programate explicit [2].”

[1] Ron Kovahi; Foster Provost (1998). "Glossary of terms". Machine Learning 30: 271–274.
[2] C. M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.

Definitie

- Conform cursului de Machine Learning sustinut de Prof. Andrew Ng la Universitatea Stanford [3]:

  - "Disciplina de studiu care le da calculatoarelor abilitatea de a invata fara a fi programati in mod explicit."- Arthur Samuel (1959)

  - "Un program de calculator se spune ca invata din experienta E vis-à-vis de o anumita clasa de sarcini T si o masura de performanta P daca performanta sa cu privire la sarcinile din T, masurate prin P, se imbunateste cu experienta E." – Tom Michel (1999)

[3] http://www.holehouse.org/mlclass/index.html

Aplicatii ML

- Masini care se conduc singure
  - https://www.youtube.com/watch?v=lL16AQItG1g
- Elicopterul autonom de la Stanford
  - https://www.youtube.com/watch?v=0JL04JJjocc
- Recunoasterea scrisului de mana
  - https://www.youtube.com/watch?v=xGVZsQxUv5w
- Minerit in baze de date: Google, Amazon, diagnoza medicala
  - https://www.youtube.com/watch?v=mgiUoSKrgbI

Bibliografie

- Cursul de Machine Learning sustinut de Prof. Andrew Ng la Universitatea Stanford: http://www.holehouse.org/mlclass/index.html

- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer, 2009

- Dianne Cook, Deborah F. Swayne, Graphics for Data Analysis. Interactive and Dynamic With R and Ggobi, Springer, 2007

- Andrew Webb, Keith Copsey, Statistical Pattern Recognition, Wiley, 2011

Bibliografie

- Nathalie Japkowicz, Mohak Shah, Evaluating Learning Algorithms, Cambridge, 2011

- Florin Gorunescu, Data Mining: Concepts, Models And Techniques, Springer, 2011

- Catalin Stoean, Ruxandra Stoean, Support Vector Machines and Evolutionary Algorithms for Classification: Single or Together, Springer, 2014.

Despre curs si examen

- Curs: notiuni teoretice, modele, explicatii, discutii
- Laborator: implementari modele in limbajul R[1], [2]
  - W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R, 2014
- Nota finala: nota examen + puncte laborator, proiecte

[1] http://www.r-project.org/
[2] http://www.rstudio.com/

Tipuri de invatare computationala

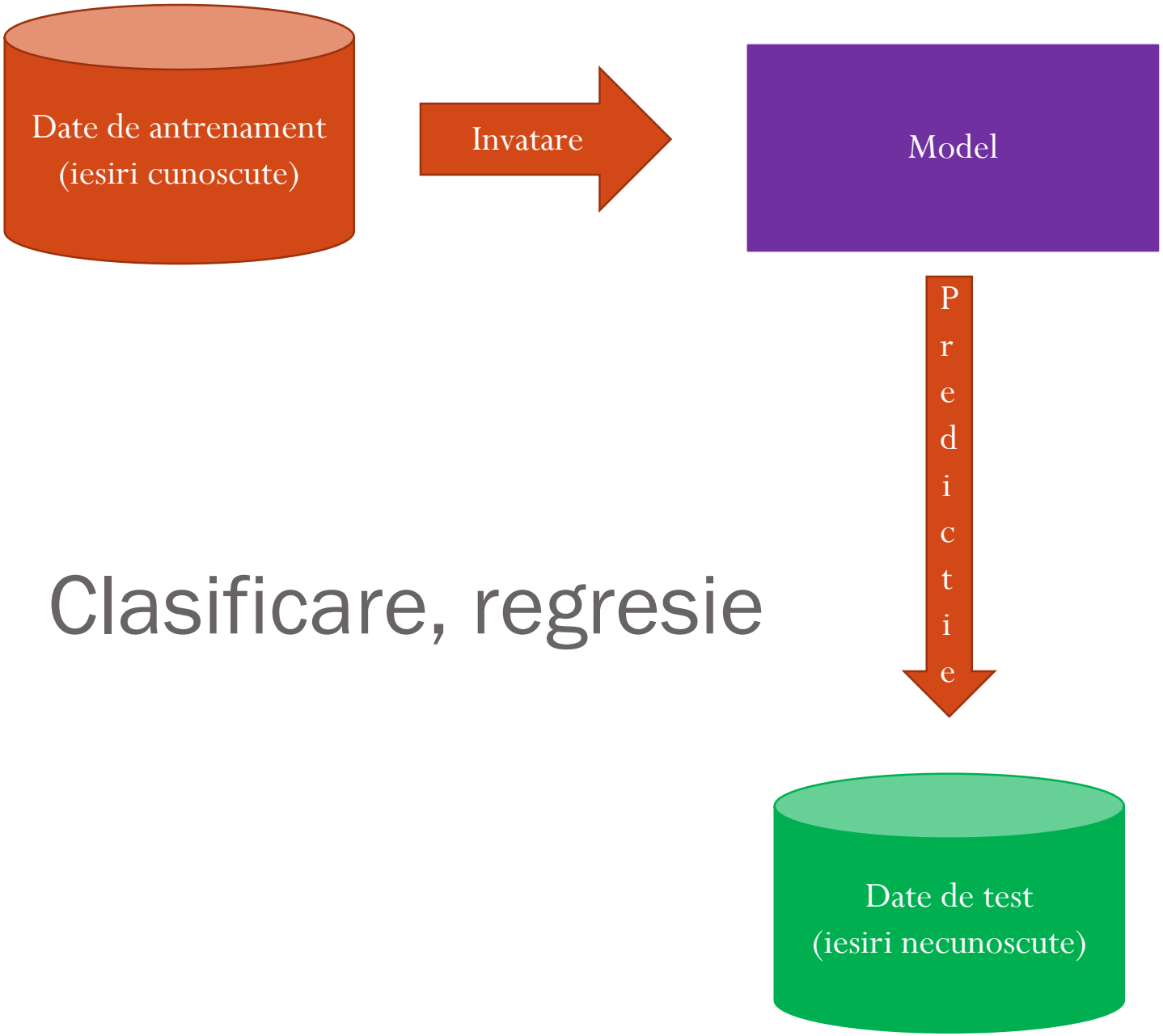
- Invatare supervizata:
  - Date de intrare cu iesiri puse la dispozitie
  - De invatat modul de asociere intrare-iesire
  - Predictie asupra iesirii unor date noi

- Invatare nesupervizata:
  - Date de intrare fara iesiri
  - De invatat/descoperit structura, caracteristici date

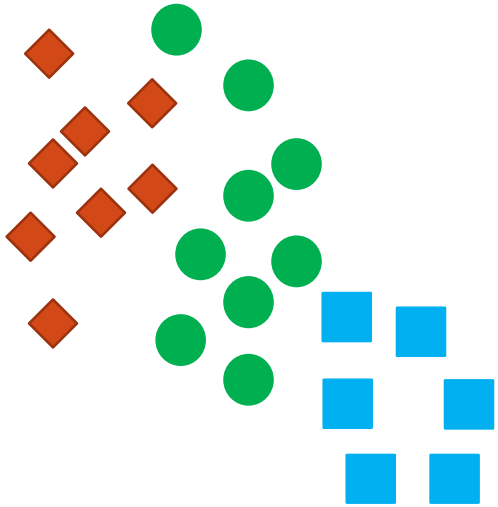
Tipuri de invatare computationala

- Invatare supervizata:
  - Clasificare
    - Iesiri grupate in doua sau mai multe clase calitative [1]
    - Predictie asupra clasei unor noi date
  - Regresie
    - Iesiri cantitative [1]
    - Predictie asupra valorii de iesire pentru intrari noi
- Invatare nesupervizata:
  - Clustering
    - Impartirea datelor in grupuri
- Colateral:
  - Selectia variabilelor si reducerea dimensionalitatii
  - Evaluarea si selectia modelelor de invatare

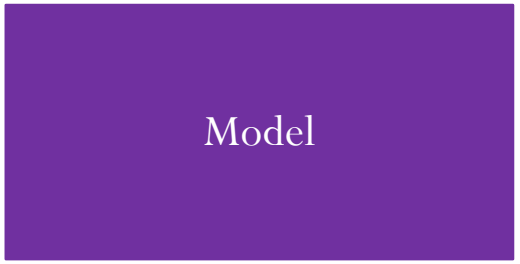
[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer, 2009

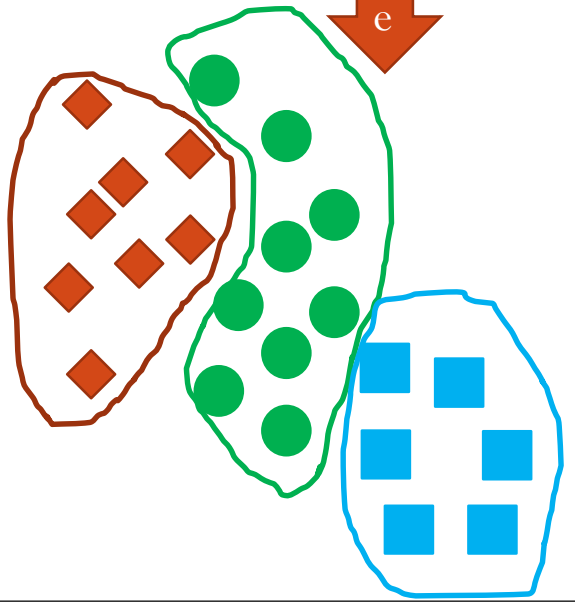


Clasificare, regresie



Date de antrenament





Clustering

Exemple

- Clasificare
  - Etichetare e-mail-uri ca spam/non-spam
  - Diagnosticare tumora pacient ca beningna/maligna
- Regresie
  - Estimare pret apartament pe baza dotarilor
  - Estimare pret masina pe baza caracteristicilor
- Clustering
  - Recomandari produse similare Amazon, Elefant etc.
  - Gruparea stirilor pe site-urile specializate

Exemplu 1 – Clasificare Iris

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.1	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*
4.8	3.0	1.4	0.1	*I. setosa*
4.3	3.0	1.1	0.1	*I. setosa*
5.8	4.0	1.2	0.2	*I. setosa*
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
5.7	2.8	4.5	1.3	*I. versicolor*
6.3	3.3	4.7	1.6	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.6	2.9	4.6	1.3	*I. versicolor*
5.2	2.7	3.9	1.4	*I. versicolor*
5.0	2.0	3.5	1.0	*I. versicolor*
5.9	3.0	4.2	1.5	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.1	2.9	4.7	1.4	*I. versicolor*
5.6	2.9	3.6	1.3	*I. versicolor*
6.7	3.1	4.4	1.4	*I. versicolor*
5.6	3.0	4.1	1.3	*I. versicolor*
5.5	2.5	4.0	1.3	*I. versicolor*
5.5	2.6	4.4	1.2	*I. versicolor*
6.1	3.0	4.6	1.4	*I. versicolor*
5.8	2.6	4.0	1.2	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
5.6	2.7	4.2	1.3	*I. versicolor*
5.7	3.0	4.2	1.2	*I. versicolor*
5.7	2.9	4.2	1.3	*I. versicolor*
6.2	2.9	4.3	1.3	*I. versicolor*
5.1	2.5	3.0	1.1	*I. versicolor*
5.7	2.8	4.1	1.3	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*

- Recunoasterea tipului unei plante noi dupa caracteristici

Iris setosa

Iris versicolor

Iris virginica

Exemplu 2 – Regresie in sport

- Pentru a forma echipe cat mai competitive.

Exemplu 3 - Clustere de galaxii

- https://www.youtube.com/watch?v=rENyyRwxpHo&list=PLQQa5REN2h_W-W4lPSt_IUx-B_JxcOZi4

Abordari invatare supervizata

- Modele liniare
- k-Nearest Neighbors (kNN)
- Masini cu suport vectorial
- Retele neuronale
- Arbori de decizie

- Metode ansamblu (Ensemble methods)
  - Bagging
  - Boosting
  - Random Forests

Abordari invatare supervizata

- Selectia si evaluarea modelelor obtinute:
  - Verificarea potrivirii cu datele / Diagnoza modelului
  - Masuri de performanta
  - Estimarea erorii
  - Comparatie intre mai multe modele. Testarea semnificatiei statistice
  - Selectia problemelor de test (benchmarking)

Abordari invatare nesupervizata

- Clustering partitional sau bazat pe centroizi
- Clustering ierarhic
- Clustering bazat pe distributie
- Clustering bazat pe densitate

- Selectie si evaluare modele obtinute:
  - Validitatea clusterelor obtinute
  - Teste statistice
  - Alegerea numarului de clustere

Selectia modelului (Model selection)

- Din mai multe modele posibile, orice paradigma computationala bazata pe invatare va alege cel mai bun pentru problema considerata.

- Atentia la echilibrul dintre deplasare si dispersie (bias-variance)
  - Generalizare - underfitting (bias ridicat)
  - Specializare - overfitting (dispersie ridicata)

Imbunatatirea performantei

- Setarea parametrilor (Parameter tuning)
  - Manual
  - Automat

- Tratarea atributelor
  - Selectia trasaturilor (Feature selection)
    - Reducerea numarului de atribute la cele importante
  - Extragerea trasaturilor (Feature extraction)
    - Combinarea atributelor originale intr-o multime mai mica de noi caracteristici

Laborator – Introducere in R [1]

- Construiti vectorii x si y care sa contina numerele 7.4, 9.3, 5, 12 si -7.3, 2, 9, -4.9, respective.
- Calculati vectorul v care sa aiba elemente obtinute dupa formula 2x+3y+1.
- Calculati pentru v lungimea, minimul, maximul, suma, produsul, media si deviatia standard ale elementelor sale.
- Sortati vectorii x si y.
- Generati un vector de tip secvential de la -10 la 10 care sa aiba elementele din 0.5 in 0.5.
- Generati un vector cu marci de masini.
- Fie n = 10, comparati secventele 1:n-1 si 1:(n-1).

[1] W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R, 2014

Exercitii (2)

- Generati un vector logic prin compararea elementelor din x cu 8.
- Folosind un vector index, selectati din y intr-un vector z numai elementele pozitive.
- Aflati elementul de pe pozitia 3 din vectorul v.
- Folosind un vector index, selectati din x intr-un vector z primele 3 elemente.
- Folosind un vector index, construiti din x un vector z excluzand primele 3 elemente ale sale.
- Inlocuiti elementele negative din y cu 0.
- Generati un vector de preturi pentru 4 tipuri de haine, atasati un vector cu hainele corespunzatoare si apoi scrieti o interogare prin care sa se calculeze pretul total pe care trebuie sa il plateasca un client pentru doua haine alese.

Exercitii (3)

- Definiti un vector de numere intregi x si schimbati-l intr-un vector caracter y care sa contina stringul asociat fiecaruia (fiecare numar intre ghilimele). Transformati-l apoi la loc in intregi dispusi intr-un vector z.
- Construiti un vector de 3 numere intregi. Apoi adaugati numarul 3 pe pozitia a 4-a. Scurtati-l ulterior la primele 2 pozitii.
- Avand un vector de tip caracter cu obiectele “unu”, “doi” si “trei”, creati un vector factor din acesta si afisati vectorul nou si nivelurile sale.

Exercitii - continuare

- La un concurs de informatica participa trei licee notate prin “unu”, “doi” si “trei”. De la fiecare liceu participa mai multi elevi care obtin note dupa urmatorii vectori:

elevi <- c(“unu”, “unu”, “trei”, “doi”, “trei”, “unu”, “doi”, “doi”, “trei”, “unu”)

note <- c(7, 3, 9, 10, 9, 8, 5, 2, 7, 9)

Transformati vectorul elevi intr-unul factor si calculati media notelor elevilor de la fiecare liceu.

Exercitii (4)

- Creati o matrice 2 x 3. Extrageti prima coloana a sa. Extrageti a doua linie. Extrageti elementul de pe pozitia (2, 3). Inlocuiti primele doua elemente de pe linia a 2-a cu 0.
- Creati un vector de lungime 3 si o matrice 3 x 2 si uniti-le pe coloana. Idem cu un vector de lungime 2 si aceeasi matrice cu unire pe linie.
- Uniti pe coloana doua matrice de 2 x 3 si 2 x 4.
- Uniti pe linie doua matrice 2 x 2 si 3 x 2.

Exercitii (5)

- Construiti doua liste cu urmatoarele componente: numele si prenumele cate unui student, anul nasterii si un vector cu doua componente - media notelor de studiu si nota de la licenta.
  - Adaugati ulterior la fiecare inca o componenta care sa specifice firma la care s-a angajat dupa alsolvire.
  - Concatenati apoi cele doua liste.
  - Calculati media finala (intre note studii si nota licenta) pentru fiecare dintre cei doi studenti.

Exercitii (6)

- Creati un fisier .data cu urmatorul cap de tabel

(marca, motor, combustibil, an, km, pret)

privind datele mai multor masini de la un parc auto.

- Creati apoi un data frame care sa importe in R datele din fisier.

- Gasiti apoi pretul minim si maxim si caracteristicile masinilor corespunzatoare.