MACHINE LEARNING

Liviu Ciortuz
Department of CS, University of Iaşi, România

What is Machine Learning?

- ML studies algorithms that $\underbrace{\text{improve with}}_{\text{learn from}}$ experience.

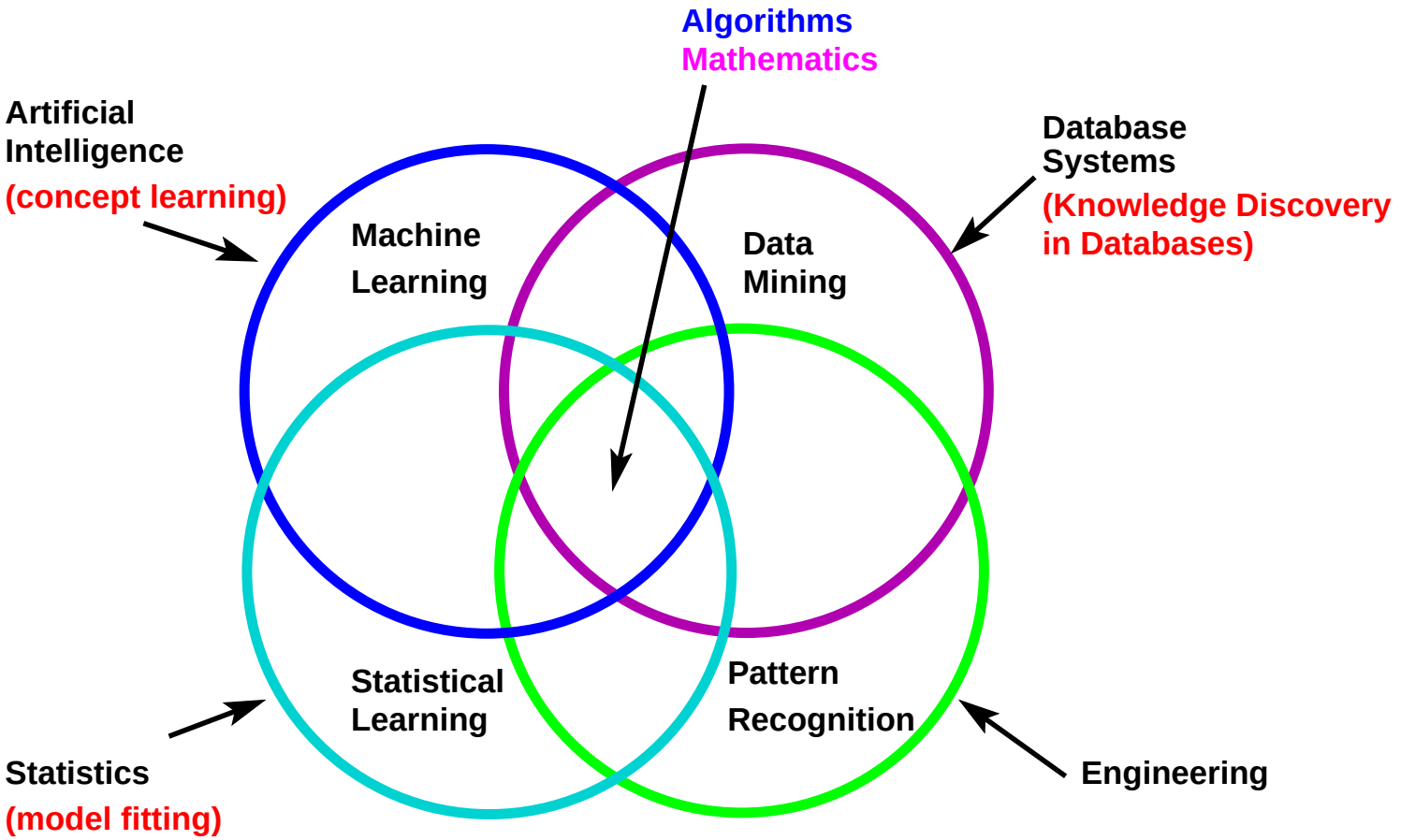
  Tom Mitchell's Definition of the [*general*] *learning problem*:
  "A computer program is said to *learn* from experience E with respect to some class of *tasks* T and *performance measure* P, if its performance on tasks in T, as measured by P, improves with experience E."

- Examples of [specific] learning problems (see next slide)
- [Liviu Ciortuz:] ML is data-driven programming
- [Liviu Ciortuz:] ML gathers a number of well-defined sub-domains/disciplines, each one of them aiming to solve in its own way the above-formulated [*general*] *learning problem*.

What is Machine Learning good for?

- natural language (text & speech) processing
- genetic sequence analysis
- robotics
- customer (financial risc) evaluation
- terrorist threat detection
- compiler optimisation
- semantic web
- computer security
- software engineering
- computer vision (image processing)
- etc.

A multi-domain view



The Machine Learning Undergraduate (2017 fall) Course: Plan

0. Probabilities Revision (Ch.Manning & H.Schütze, ch.2)

---

1. Introduction to Machine Learning (T.Mitchell, ch.1)

---

2. Decision Trees (T.Mitchell, ch.3)
3. Bayesian Learning (T.Mitchell, ch.6)
4. Instance-based Learning (T.Mitchell, ch.8)
5. Clustering Algorithms (Ch.Manning & H.Schütze, ch.14)

Bibliography

0. **"Exerciţii de învăţare automată"**
   L. Ciortuz, A. Munteanu E. Bădărău.
   Editura Universităţii "Alexandru Ioan Cuza", Iaşi, Romania, 2017

---

1. **"Machine Learning"**
   Tom Mitchell. McGraw-Hill, 1997

---

2. **"The Elements of Statistical Learning"**
   Trevor Hastie, Robert Tibshirani, Jerome Friedman. Springer, 2nd ed. 2009

---

3. **"Machine Learning – A Probabilistic Perspective"**
   Kevin Murphy, MIT Press, 2012
4. **"Pattern Recognition and Machine Learning"**
   Christopher Bishop. Springer, 2006
5. **"Foundations of Statistical Natural Language Processing"**
   Christopher Manning, Hinrich Schütze. MIT Press, 2002

Other suggested readings:
More on the theoretical side (I)

1. **"Pattern Recognition" (2nd ed.)**
   R. Duda, P. Hart, D. Stork. John Wiley & Sons Inc., 2001
2. **"Bayesian Reasoning and Machine Learning"**
   David Barber, 2012
3. **"Pattern Recognition", (Fourth Edition)**
   Sergios Theodoridis, Konstantinos Koutroumbas. Academic Press, 2008
4. **"Machine Learning. A Bayesian and Optimization Perspective",**
   Sergios Theodoridis. Elsevier, 2015

---

5. **"Apprentissage artifficiel" (2^e ed.)**
   Antoine Cornuéjols. Eyrolles, 2010

Other suggested readings: More on the theoretical side (II)

1. **"Data mining with decision trees" (2nd ed.)**
   Lior Rokach, Oded Maimon. World Scientific, 2015
2. **"Clustering"**
   Rui wu, Donald C. Wunsch II; IEEE Press, 2009

---

3. **"The EM Algorithm and Extensions" (2nd ed.)**
   Geoffrey J. McLachlan, Thriyambakam Krishnan. John Wiley & Sons, 2008
4. **"A Tutorial on Support Vector Machines for Pattern Recognition"**
   Christopher Burges, 1998
5. **"Support Vector Machines and other kernel-based learning methods"**
   Nello Cristianini, John Shawe-Taylor. Cambridge University Press, 2000.

---

6. **"Apprentissage statistique. Réseaux de neurones, cartes topologiques, machines à vecteurs supports" (3^e ed.)**
   G. Dreyfus, J.-M. Martinez, M. Samuelides, M.B. Gordon, F. Badran, S. Thiria. Eyrolles, 2007

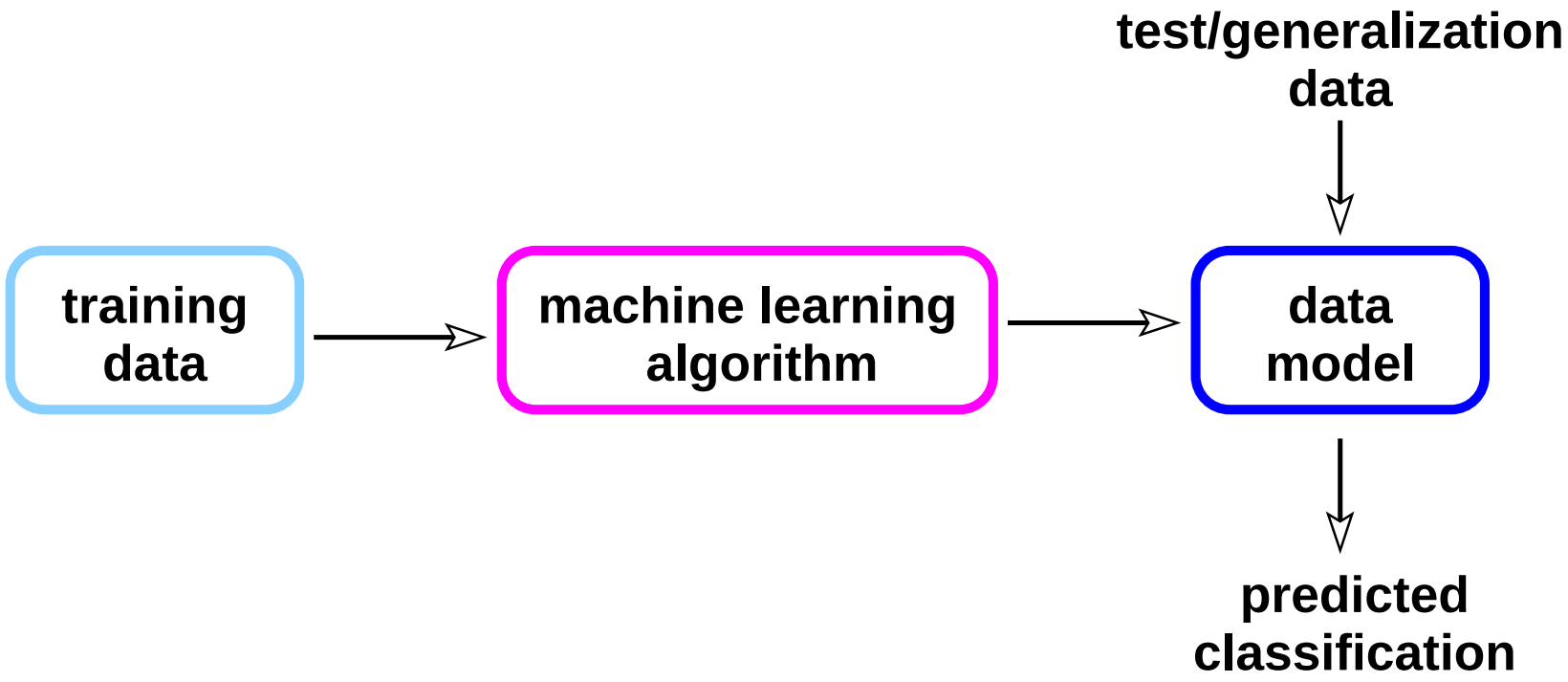
Other suggested readings:
More on the practical side

1. **"Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Ian Witten, Eibe Frank (3rd ed.). Morgan Kaufmann Publishers, 2011**
2. **"An Introduction to Statistical Learning"**
   Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer, 2013
3. **"Applied Predictive Modeling"**
   Max Kuhn, Kjell Johnson; Springer, 2013
4. **"An introduction to Pattern Recognition: A Matlab approach",**
   Sergios Theodoridis, Konstantinos Koutroumbas. Academic Press, 2010
5. **"Machine Learning with R", Brett Lantz. PACT Publishing, 2013**
6. **"Data Mining with R – Learning with Case Studies"**
   Luís Torgo. CRC Press, 2011

---

7. **"Mining of Massive Datasets"**
   Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman; 2013

A general schema for machine learning methods



Basic ML Terminology

1. **instance x, instance set X**
   concept $c \subseteq X$, or $c : X \to \{0, 1\}$
   example (labeled instance): $\langle x, c(x) \rangle$; positive examples, neg. examples
2. **hypotheses $h : X \to \{0, 1\}$**
   hypotheses representation language
   hypotheses set H
   hypotheses consistent with the concept c: $h(x) = c(x), \forall$ example $\langle x, c(x) \rangle$
   version space
3. **learning = train + test**
   supervised learning (classification), unsupervised learning (clustering)
4. ***error***$_h = |\{x \in X, h(x) \neq c(x)\}|$
   training error, test error
   accuracy, precision, recall
5. **validation set, development set**
   n-fold cross-validation, leave-one-out cross-validation
   overfitting

The Inductive Learning Assumption

Any hypothesis found to conveniently approximate the target function over a sufficiently large set of training examples

will also conveniently approximate the target function over other unobserved examples.

Inductive Bias

Consider

- **a concept learning algorithm L**
- **the instances X, and the target concept c**
- **the training examples $D_c = \{\langle x, c(x)\rangle\}$.**
- **Let $L(x_i, D_c)$ denote the classification assigned to the instance x_i by L after training on data D_c.**

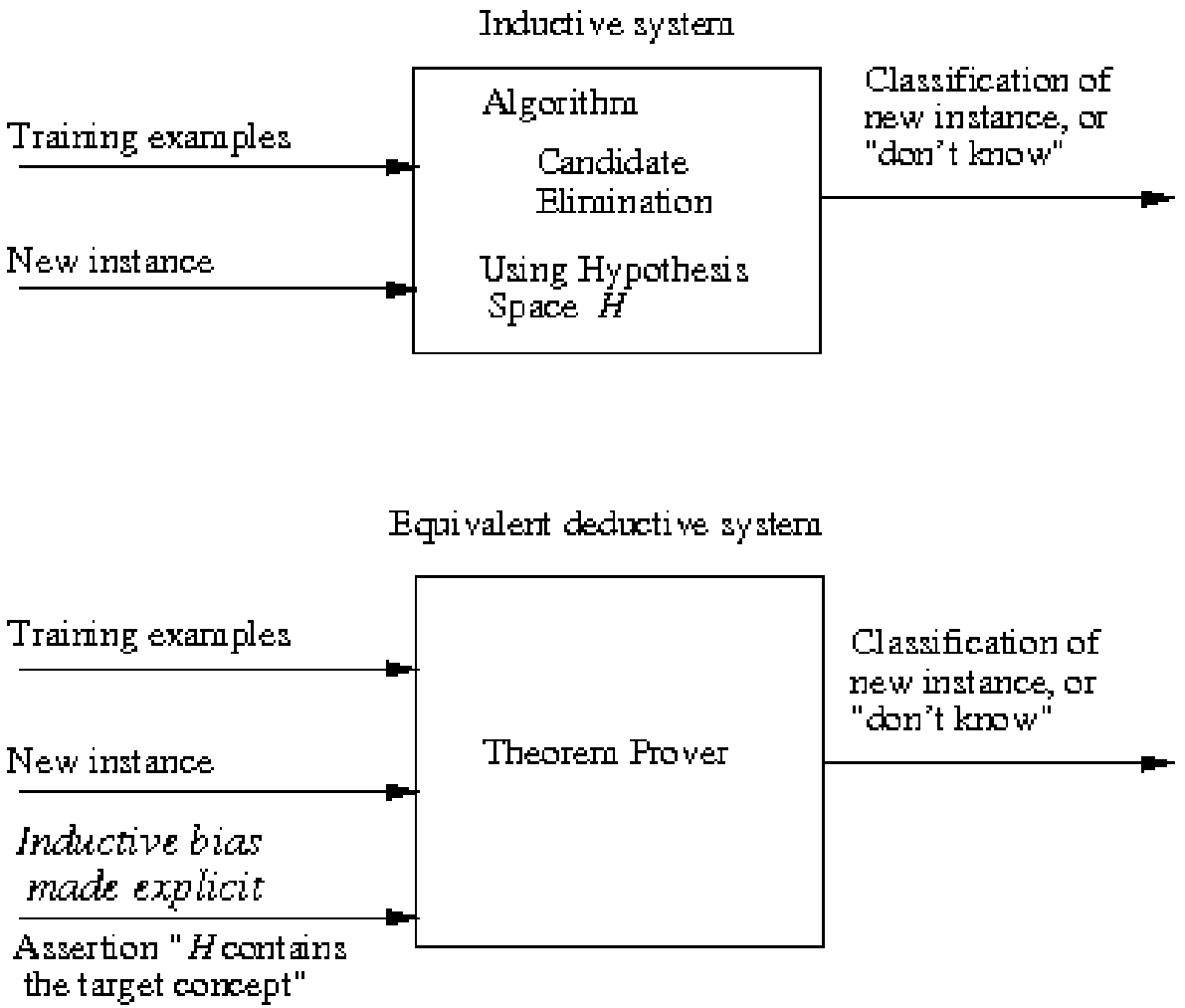
Definition:

The inductive bias of L is any minimal set of assertions B such that

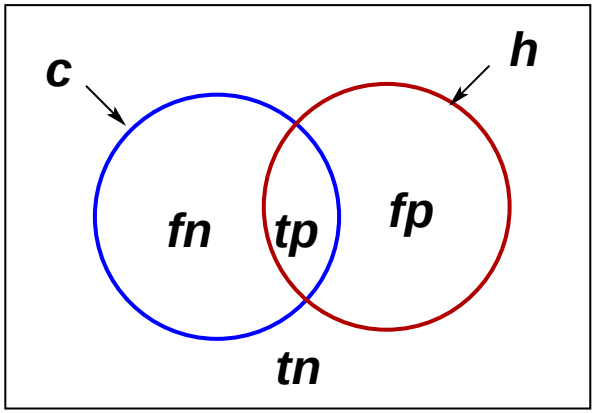
$$(\forall x_i \in X)[(B \vee D_c \vee x_i) \vdash L(x_i, D_c)]$$

for any target concept c and corresponding training examples D_c. ($A \vdash B$ means A logically entails B)

Inductive systems can be modelled by equivalent deductive systems



Evaluation measures in Machine Learning



tp – true positives
fp – false positives
tn – true negatives
fn – false negatives

accuracy: $Acc = \frac{tp + tn}{tp + tn + fp + fn}$

precision: $P = \frac{tp}{tp + fp}$

recall (or: sensitivity): $R = \frac{tp}{tp + fn}$

F-measure: $F = \frac{2\ P \times R}{P+R}$

specificity: $Sp = \frac{tn}{tn + fp}$

follout: $= \frac{fp}{tn + fp}$

Mathew's Correlation Coefficient:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp) \times (tn + fn) \times (tp + fn) \times (tn + fp)}}$$

Lazy learning vs. eager learning algorithms

Eager: generalize before seeing query

- ID3, Backpropagation, Naive Bayes, Radial basis function networks, ...
- Must create global approximation

Lazy: wait for query before generalizing

- k-Nearest Neighbor, Locally weighted regression, Case based reasoning
- Can create many local approximations

Does it matter?

If they use the same hypothesis space H, lazy learners can represent more complex functions.
E.g., a lazy Backpropagation algorithm can learn a NN which is different for each query point, compared to the eager version of Backpropagation.

Who is Liviu Ciortuz?

- **Diploma (maths and CS) from UAIC, Iaşi, Romania, 1985**
  PhD in CS from Université de Lille, France, 1996
- **programmer:**
  Bacău, Romania (1985-1987)
- **full-time researcher:**
  Germany (DFKI, Saarbrücken, 1997-2001),
  UK (Univ. of York and Univ. of Aberystwyth, 2001-2003),
  France (INRIA, Rennes, 2012-2013)
- **assistant, lecturer and then associate professor:**
  Univ. of Iasi, Romania (1990-1997, 2003-2012, 2013-today)

ADDENDA

"...colleagues at the Computer Science department at Saarland University have a strong conviction, that nothing is as practical as a good theory."

Reinhard Wilhelm,
quoted by Cristian Calude,
in *The Human Face of Computing*,
Imperial College Press, 2016

"Mathematics translates concepts into formalisms and applies those formalisms to derive insights that are usually NOT amenable to a LESS formal analysis."

Jürgen Jost,
Mathematical Concepts,
Springer, 2015

"Mathematics is a journey that must be shared, and by sharing our own journey with others, we, together, can change the world."

"Through the power of mathematics, we can explore the uncertain, the counterintuitive, the invisible; we can reveal order and beauty, and at times transform theories into practical objects, things or solutions that you can feel, touch or use."

Cedric Villani,
winner of the Fields prize, 2010

cf. http://www.bbc.com/future/sponsored/story/20170216-inside-the-mind-of-a-mathematician, **15.03.2017**

ADMINISTRATIVIA

Teaching assistants for the ML undergraduate course 2017 (fall semester)

- Lect. dr. Anca Ignat (. . . Image processing)
- Silviu Nedelciuc (MSc)
- Bogdan Cazacu (MSc)
- Oriana Oniciuc (MSc student)
- Sebastian Ciobanu (MSc student)
- Lucian Nevoe (MSc student)

Related courses

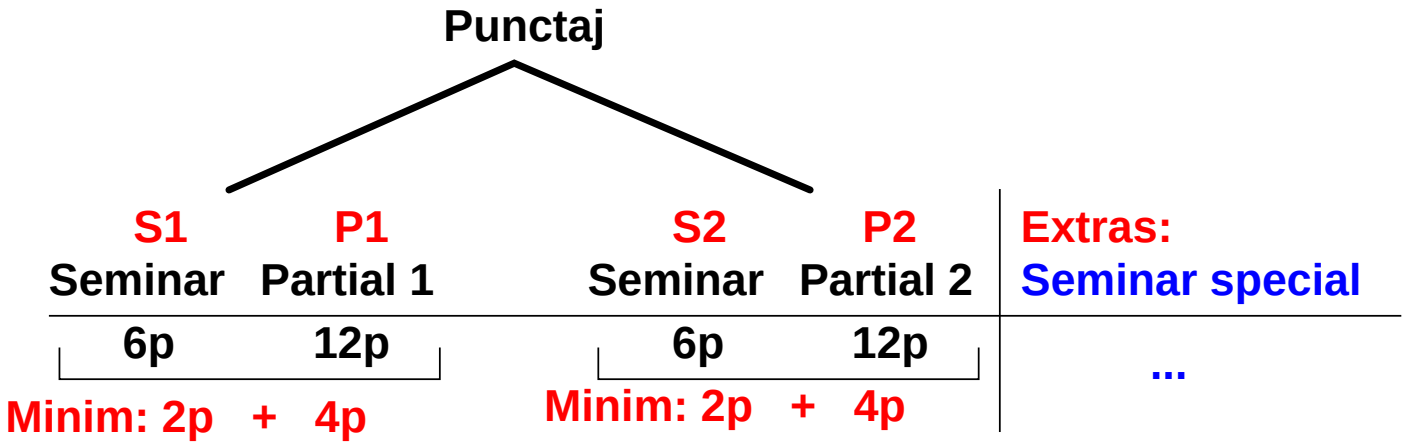
- Genetic Algorithms
- Artificial Neural Networks
- Probabilistic programming

- Special Chapters of Machine Learning
- Data Mining
- Nature-inspired computing methods
- Big Data Analytics
- Image Processing
- Exploratory Data Analysis
- Special Chapters of Artificial Neural Networks

- Bioinformatics

Grading standards for the ML undergraduate course 2017 (fall semester)

Obiectiv: invatare pe tot parcursul semestrului!



Prezenta la seminar: obligatorie!
Penalizare: 0.1p pentru fiecare absenta de la a doua incolo

Nota = (4 + S1 + P1 + S2 + P2) / 4
Pentru promovare: S1 + P1 + S2 + P2 >= 14

REGULI generale pentru cursul de Învăţare automată

Regulile de organizare a cursului de Învăţare Automată (engl., Machine Learning, ML), 2017-2018, sem. I, sunt specificate în *fişa disciplinei*
http://profs.info.uaic.ro/∼ciortuz/ML.2017f.fisa-disciplinei.RO.pdf

- **Bibliografie minimală: vezi slide #5**
- **Planificarea materiei, pentru fiecare săptămâna (curs + seminar):**
http://profs.info.uaic.ro/∼ciortuz/ML.what-you-should-know.pdf
- **Prezenţa la curs: recomandată!**
- **Săptămânal se va ţine un seminar suplimentar, destinat pentru acei studenţi care sunt foarte interesaţi de acest domeniu. (Vezi secţiunile "Advanced issues" şi "Implementation exercises" din documentul**
http://profs.info.uaic.ro/ ciortuz/ML.what-you-should-know.pdf.)
Ziua şi ora la care se va ţine acest "seminar suplimentar" vor fi anunţate în curând.

REGULI generale pentru cursul de Învăţare automată (cont.)

- **Slide-uri de imprimat** (în această ordine şi, de preferat, COLOR):

http://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf

https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.ProbStat.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.EstimP.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.Regression.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.DT.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.Bayes.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.IBL.pdf
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.Cluster.pdf

(Atenţie: acest set de slide-uri va fi actualizat pe parcursul semestrului!)

- De imprimat (ALB-NEGRU):

http://profs.info.uaic.ro/~ciortuz/SLIDES/ml0.pdf
http://profs.info.uaic.ro/~ciortuz/SLIDES/ml3.pdf
http://profs.info.uaic.ro/~ciortuz/SLIDES/ml6.pdf
http://profs.info.uaic.ro/~ciortuz/SLIDES/ml8.pdf
http://profs.info.uaic.ro/~ciortuz/SLIDES/cluster.pdf

- De imprimat opţional (ALB-NEGRU):

Companion-ul practic pentru culegerea ,,Exerciţii de învăţare automată“:
https://profs.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/ML.ex-book.Companion.pdf

REGULI generale pentru cursul de Învăţare automată (cont.)

Observaţie (1)

Pentru seminarii, nu se admit mutări ale studenţilor de la o grupă la alta, decât în cadrul grupelor care au acelaşi asistent / profesor responsabil de seminar.

Observaţie (2)

La fiecare curs şi seminar, studenţii vor veni cu cartea de exerciţii şi probleme (de L. Ciortuz et al) şi cu o fasciculă conţinând slide-urile imprimate.

Observaţie (3)

Profesorul responsabil pentru acest curs, Liviu Ciortuz NU va răspunde la email-uri care pun întrebări pentru care raspunsul a fost deja dat

– fie în aceste slide-uri,
– fie la curs,
– fie în documentul
http://profs.info.uaic.ro/ ciortuz/ML.what-you-should-know.pdf.

REGULI generale pentru cursul de Învăţare automată (cont.)

- **Vă recomand să accesaţi săptămânal:**

http://profs.info.uaic.ro/~ciortuz/ML.what-you-should-know.pdf

http://profs.info.uaic.ro/~ciortuz/ML.ex-book/ML.ex-book.overview.pdf,
acest document oferă o sinteză (un conspect) asupra materiei;

- **O carte foarte bună (în limba engleză), care conţine practic toate noţiunile din matematica de liceu de care aveţi nevoie la acest curs:**
"Mathematics for the International Student" (2012)

- **Pentru buna gestionare a timpului individual de pregătire pentru seminar şi pentru examene, vă recomand călduros să citiţi:**

http://profs.info.uaic.ro/~ciortuz/SLIDES/time_management.SLIDES.pdf

Guidelines for the first ML seminary

Pentru seminariile din prima săptămână,

– veţi recapitula noţiunile din slide-urile 2-5 din
http://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf
(vezi bibliografia indicata in slide-ul #0)

– veţi citi / studia (în prealabil) problemele rezolvate din documentul
http://profs.info.uaic.ro/~ciortuz/ML.ex-book/sem1.pdf
şi veţi rezolva problemele propuse acolo.

Recomandarea profesorului responsabil de curs (L. Ciortuz) este ca la acest seminar să fie ascultaţi la tablă primii (3-5) studenţi la catalog, de la fiecare grupă.