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 (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 a tasks in T, as measured by P, improves with experience E."

- **Examples of [specific] learning problems (see next slide)**

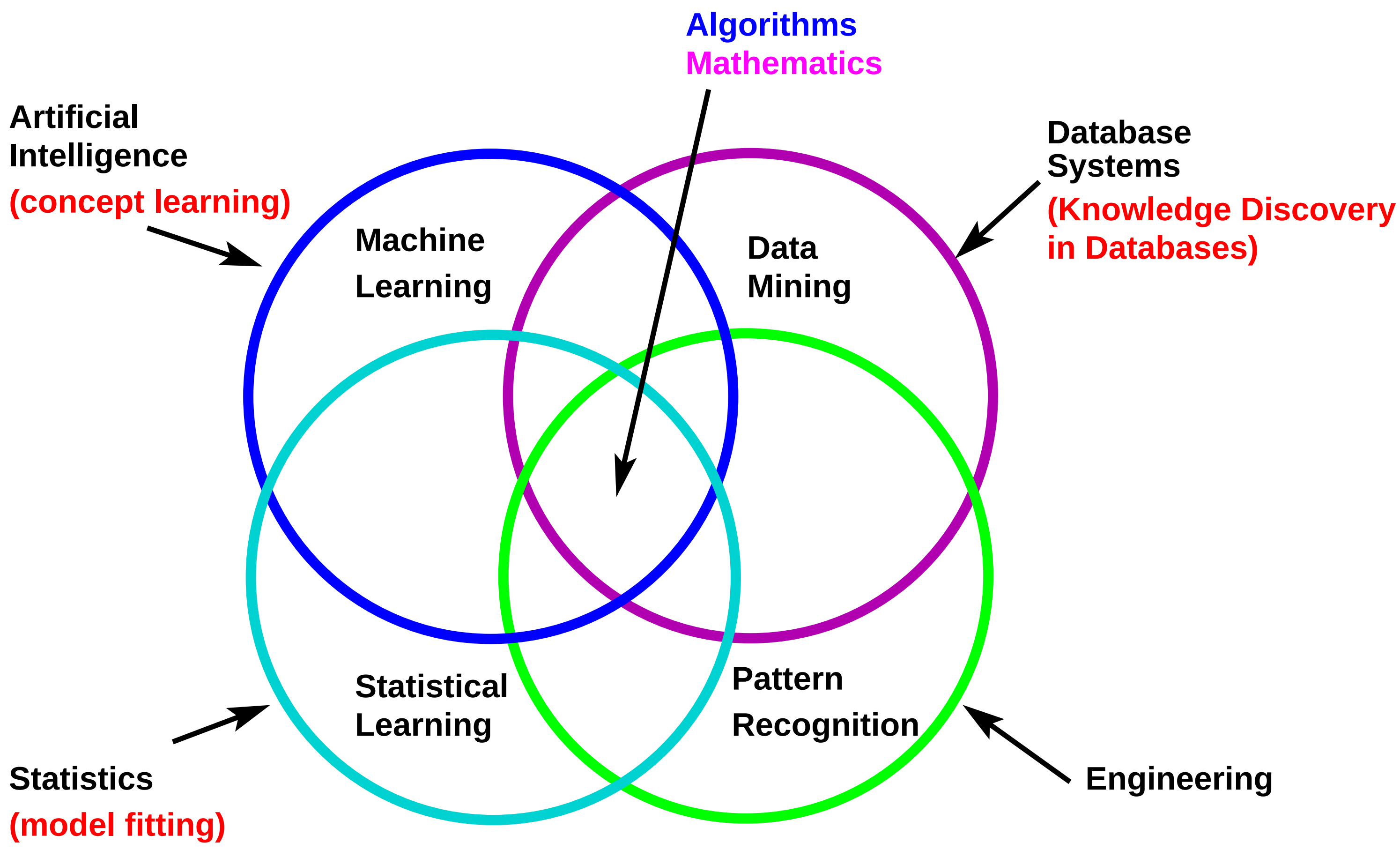
◦ **[Liviu Ciortuz:] ML is data-driven programming**

◦ **[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 BSc Course: Plan

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

---

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

---

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)
6. The EM algorithmic schemata (T.Mitchell, ch.6.12)

Bibliography

1. **“Machine Learning”**
   Tom Mitchell; McGraw-Hill, 1997

2. **“Foundations of Statistical Natural Language Processing”**
   Christopher Manning, Hinrich Schütze; MIT Press, 2002

---

3. **“Exerciţii de învăţare automată”**
   L. Ciortuz, A. Munteanu E. Bădărău,
   Editura Universităţii “Alexandru Ioan Cuza”, Iaşi, 2015 (in print)

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

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

---

2. "Pattern Recognition", (Fourth Edition)
   Sergios Theodoridis, Konstantinos Koutroumbas, Academic Press, 2008
3. "Machine Learning. A Bayesian and Optimization Perspective",
   Sergios Theodoridis, Elsevier, 2015
4. "Machine Learning – A Probabilistic Perspective"
   Kevin Murphy, MIT Press, 2012

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

1. "Pattern Recognition" (2nd ed.)
   Duda, Hart, Stork; John Wiley & Sons Inc., 2001
2. "Pattern Recognition and Machine Learning"
   Christopher Bishop; Springer, 2006
3. "'Apprentissage artifficiel' (2^e éd.)
   Antoine Cornuéjols, Eyrolles, 2010
4. "Bayesian Reasoning and Machine Learning"
   David Barber, 2012
5. "Data mining with decision trees" (2nd ed.)
   Lior Rokach, Oded Maimon, World Scientific, 2015
6. "A probabilistic theory of pattern recognition"
   Luc Devroye, László Györfi, Gábor Lugosi, Springer, 1996
7. "Clustering"
   Rui wu, Donald C. Wunsch II; IEEE Press, 2009

---

8. "Support Vector Machines and other kernel-based learning methods"
   Nello Cristianini, John Shawe-Taylor, Cambridge University Press, 2000.

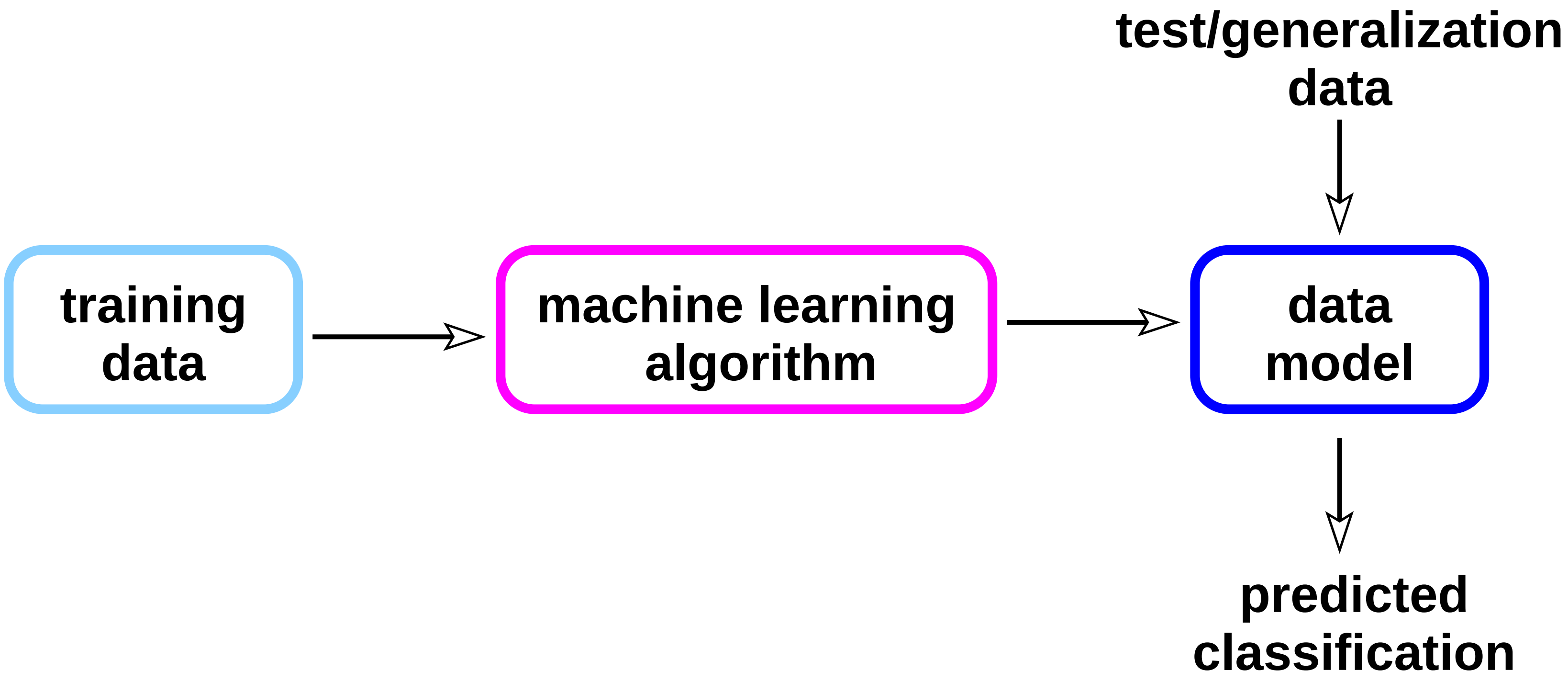
Other suggested readings: More on the practical side

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, 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. **"Top Ten Algorithms in Data Mining"
   Xindong Wu, Vipin Kumar; CRC Press, 2009**
6. **"Machine Learning with R", Brett Lantz, PACT Publishing, 2013**
7. **"Data Mining with R – Learning with Case Studies"
   Luís Torgo; CRC Press, 2011**

---

8. **"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 $c : X \rightarrow \{0, 1\}$
   example (labeled instance): $\langle x, c(x) \rangle$; positive examples, neg. examples
2. **hypotheses $h : X \rightarrow \{0, 1\}$**
   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$_h$* $= \mid \{x \in X, h(x) \neq c(x)\} \mid$**
   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

- **concept learning algorithm** L
- **instances** X**, target concept** c
- **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**.**

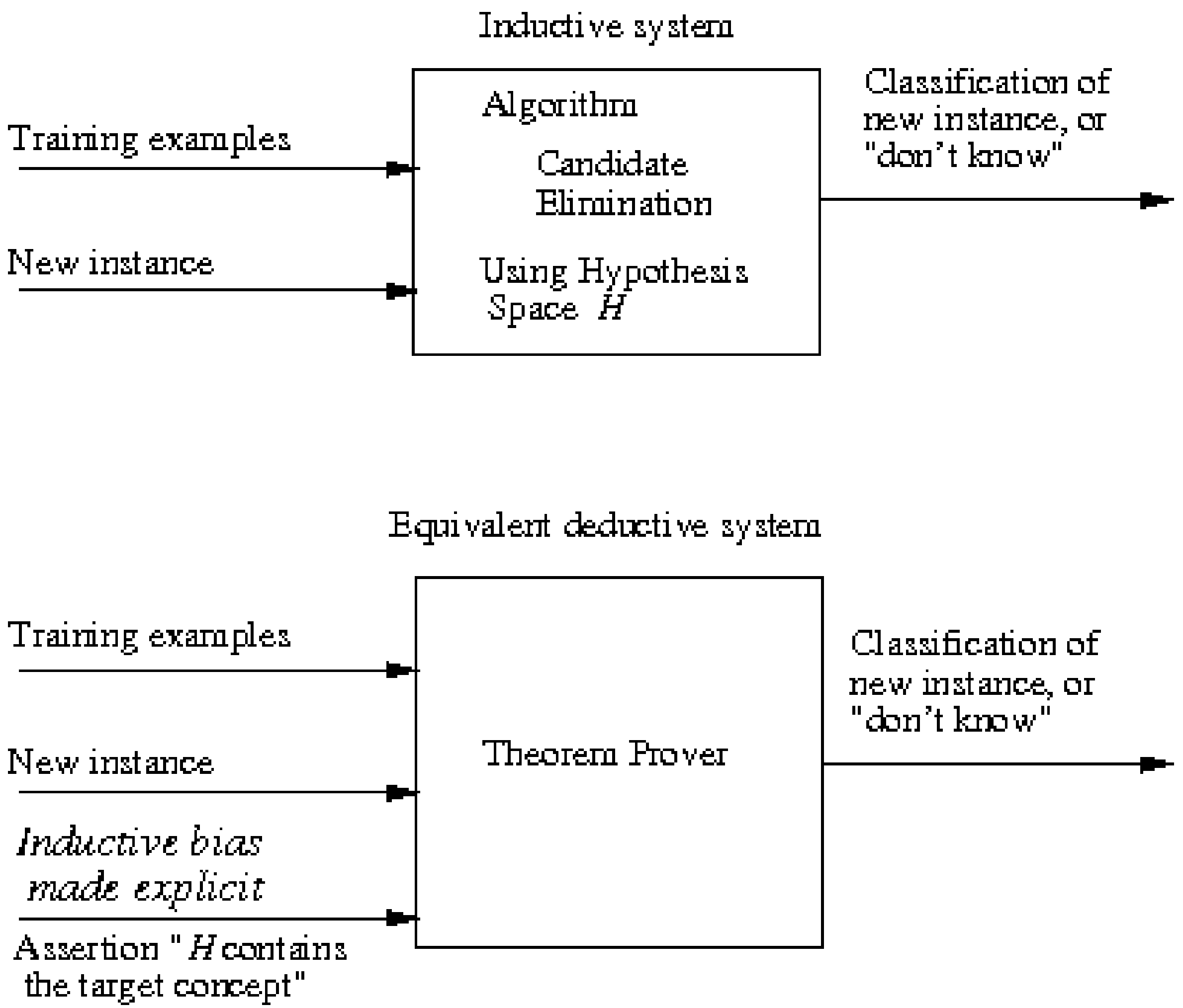
Definition:

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

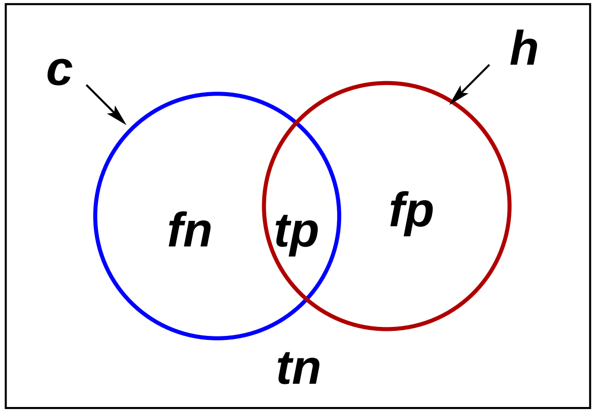
$$(\forall x_i \in X)[(B \vee D_c \vee x_i) \vdash L(x_i, D_c)]$$

for any target concept c **and corresponding training examples** D_c

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.

ADMINISTRATIVIA

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 associated professor:**
  Univ. of Iasi, Romania (1990-1997, 2003-2012, 2013-today)

Teaching assistants for the ML undergraduate course 2016 (fall semester)

- Lect. dr. Anca Ignat (. . . Image processing)
- Conf. dr. Mihaela Breabăn (. . . Big data)
- Tiberius Dumitriu (“Gh. Asachi” Tehnical Univ. of Iaşi)
- Elena Bădărău
- Oriana Oniciuc
- Ştefan Panţiru (?)

Related courses

- Genetic Algorithms
- Artifical Neural Networks
- Probabilistic programming

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

◦ Bioinformatics

General RULES for the ML course

Regulile de organizare a cursului de Învăţare Automată (engl., Machine Learning, ML), 2016-2017, sem. I, sunt specificate în documentul
http://profs.info.uaic.ro/∼ciortuz/ML.txt

- **Bibliografie minimală: vezi slide #5**

- **Planificarea materiei, pentru fiecare săptămâna (curs + seminar):**
http://profs.info.uaic.ro/∼ciortuz/ML.what-you-should-know.2016f.pdf
varianta incluzând exerciţii de implementare:
http://profs.info.uaic.ro/∼ciortuz/ML.what-you-should-know.+implementations.2016f.pdf

- **Prezenţa la curs: recomandată! Prezenţa la seminar: obligatorie!**
Pentru fiecare absenţă la seminar, începând de la a doua absenţă încolo, se aplică o penalizare/depunctare de 0.1 puncte din S1, respectiv din S2. (Vezi formula de notare.)
Regulile se aplică inclusiv studenţilor reînmatriculaţi şi cursanţilor.

- **Săptămânal se va ţine un seminar suplimentar, destinat pentru cei mai buni studenţi sau 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.+implementations.2016f.pdf.)**
Ziua şi ora la care se va ţine acest "seminar suplimentar" vor fi anunţate în curând.

General RULES for the ML course (cont'd)

Sistemul de notare

Nota = (4 + S1 + P1 + S2 + P2) / 4,
unde

> **S1 = punctajul la seminar pe prima jumătate de semestru (0-6 puncte)**
> **S2 = punctajul la seminar pe a doua jumătate de semestru (0-6 puncte)**
> **P1 = punctajul la primul examen parţial (0-12 puncte)**
> **P2 = punctajul la al doilea examen parţial (0-12 puncte)**

Punctajele S1 si S2 se obţin (fiecare) ca medie aritmetică a două punctaje, pentru
– răspunsuri "la tablă"
– test scris (anunţat în prealabil)

Condiţii de promovare:

S1 ≥ 2; S2 ≥ 2; P1 ≥ 4, P2 ≥ 4, nota ≥ 4.5
În consecinţă, punctajul minimal de îndeplinit din suma S1+P1+S2+P2 este 14.

Atenţie:

S1 < 2 (sau S2 < 2) implică imediat nepromovarea acestui curs în anul universitar 2016–2017!

General RULES for the ML course (cont'd)

- **Slide-uri (de imprimat în această ordine):**

http://profs.info.uaic.ro/~ciortuz/SLIDES/ml0.pdf
http://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.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

---

http://profs.info.uaic.ro/~ciortuz/SLIDES/ML.ex-book.SLIDES.pdf
(Atenţie: acest set de slide-uri va fi actualizat pe parcursul semestrului!)

- **Vă recomand să imprimaţi şi**

http://profs.info.uaic.ro/~ciortuz/ML.what-you-should-know.+implementations.2016f.pdf
http://profs.info.uaic.ro/~ciortuz/ML.ex-book/ML.ex-book.overview.pdf,
acest document oferă o sinteză (un conspect) asupra materiei;

http://profs.info.uaic.ro/~ciortuz/ML.ex-book/ML.ex-book.synopsis.pdf,
în acest document se indică sursele exerciţiilor şi se marchează (cu ■) care sunt exerciţiile pentru care avem deja slide-uri făcute;

---

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

General RULES for the ML course (cont'd)

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.ex-book.what-you-should-know.2016f.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.

Guidelines for the second ML seminary

Pentru seminariile din a doua săptămână,

– veţi recapitula noţiunile din slide-urile 6-15 din
http://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf
(vezi bibliografia indicata in slide-ul #0)

– veţi citi / studia (în prealabil) cât mai multe dintre problemele rezolvate din documentul
http://profs.info.uaic.ro/~ciortuz/ML.ex-book/sem2.pdf
şi veţi rezolva problemele propuse acolo.

ADDENDA

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

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

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