ML course, 2016 fall
What you should know:

Week 1, 2 and 3.$\frac{1}{2}$: Basic issues in Probabilities and Information theory

Read: Chapter 2 from the *Foundations of Statistical Natural Language Processing* book by Christopher Manning and Hinrich Schütze, MIT Press, 2002.[1]

Week 1: Random events

(slides 3-6 from https://profs.info.uaic.ro/∼ciortuz/SLIDES/foundations.pdf)

Concepts/definitions:

- sample space, random event, event space
- probability function
- conditional probabilities
- independent random events (2 forms); conditionally independent random events (2 forms)

Theoretical results/formulas:

- elementary probability formula: $\frac{\text{\# favorable cases}}{\text{\# all possible cases}}$
- the "multiplication" rule; the "chain" rule
- "total probability" formula (2 forms)
- Bayes formula (2 forms)

Exercises illustrating the above concepts/definitions and theoretical results/formulas, in particular: proofs for certain properties derived from the *definition of the probability function* for instance: $P(\emptyset) = 0$, $P(\bar{A}) = 1 - P(A)$, $A \subseteq B \Rightarrow P(A) \leq P(B)$

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 1-5 [6-7], 8, 39-42 [43-45]

[1]For a more concise / formal introductory text, see *Probability Theory Review for Machine Learning*, Samuel Ieong, November 6, 2006 (https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf) and/or *Review of Probability Theory*, Arian Maleki, Tom Do, Stanford University.

Week 2: Random variables, and (several) usual probabilistic distributions

(slides 7-9, 13-16, 36-37 [10-12, 17-22, 38-44]
from https://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf)

Concepts/definitions:

- random variables;
random variables obtained through function composition
- discrete random variables;
probability mass function (pmf)
examples: Bernoulli, binomial, geometric, Poisson distributions
- cumulative function distribution
- continuous random variables;
probability density function (pdf)
examples: Gaussian, exponential, Gamma, Laplace distributions
- expectation (mean), variance, standard variation; covariance. (**See definitions!**)
- multi-valued random functions;
joint, marginal, conditional distributions
- independence of random variables;
conditional independence of random variables

Advanced issues:

- vector of random variables;
covariance matrix for a vector of random variables;
pozitive [semi-]definite matrices,
negative [semi-]definite matrices
- the likelihood function (see *Estimating Probabilities*, additional chapter to the *Machine Learning* book by TomMitchell, 2016)

Theoretical results/formulas:

- for any discrete variable X:
$\sum_x p(x) = 1$, where p is the pmf of X

  for any continuous variable X:
$\int p(x)\, dx = 1$, where p is the pdf of X
- $E[X+Y] = E[X] + E[Y]$
$E[aX+b] = aE[X] + b$
$E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i E[X_i]$.

  $Var[aX] = a^2\, Var[X]$
$Var[X] = E[X^2] - (E[X])^2$
$Cov(X,Y) = E[XY] - E[X]E[Y]$
- X, Y independent variables $\Rightarrow$
$Var[X+Y] = Var[X] + Var[Y]$
- X, Y independent variables $\Rightarrow$
$Cov(X,Y) = 0$, i.e. $E[XY] = E[X]E[Y]$

Advanced issues:

- For any vector of random variables, the covariance matrix is symmetric and positive semi-definite.

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 25

Exercises illustrating the above concepts/definitions and theoretical results/formulas, concentrating especially on:

- identifying in a given problem's text the underlying probabilistic distribution: either a basic one (e.g., Bernoulli, binomial, categorial, multinomial etc.), or one derived [by function composition or] by summation of identically distributed random variables
- computing probabilities
- computing means / expected values of random variables
- verifying the [conditional] independence of two or more random variables

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 9-16 [17-22], 46-55 [57-63], 64 (additional: ch. *Bayesian classification*, ex. 3)

Implementation exercises for advanced issues:

1. CMU, 2009 fall, Geoff Gordon, HW3, pr. 3
Implement *Linear Regression* and apply it to the task of predicting the level of PSA (Prostate Specific Agent) in prostate tissue, using a set of 8 variables (medical test results).[2]

[2] A somehow simpler exercise, CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 4, uses linear regression on the compute the quantity of insulin to be injected into a patient based on his/her blood sugar level.

Week 3.$\frac{1}{2}$: Elements of Information Theory

(slides 28-31 [32-33] from https://profs.info.uaic.ro/∼ciortuz/SLIDES/foundations.pdf)

Concepts/definitions:

- entropy;
specific conditional entropy;
average conditional entropy;
joint entropy;
information gain (mutual information)

Advanced issues:

- relative entropy;
- cross-entropy

Theoretical results/formulas:

- $0 \leq H(X) \leq H(\underbrace{1/n, 1/n, \ldots, 1/n}_{n \text{ times}}) = \log_2 n$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
(generalisation: the chain rule, $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$)
- $H(X,Y) = H(X) + H(Y)$ iff X and Y are indep.
- $IG(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $IG(X;Y) \geq 0$
- $IG(X;Y) = 0$ iff X and Y are independent

Exercises illustrating the above concepts/definitions and theoretical results/formulas, concentrating especially on:

– computing different types of entropies (see **Ciortuz et al.'s exercise book:** ch. *Foundations*, ex. 31, 32, 35, 65, 71 [37, 70]);
(additional: ch. *Bayesian classification*, ex. 3)

– proof of some basic properties (see **Ciortuz et al.'s exercise book:** ch. *Foundations*, ex. 29, 30, [33, 34,] 36, [38], 66-69), including the functional analysis of the entropy of the Bernoulli distribution, as a base for drawing its plot.

Week $3.\frac{2}{2}$, 4 and 5: Decision Trees

Read: Chapter 3 from Tom Mitchell's *Machine Learning* book.

Important Note:
See the Overview (rom.: "Privire de ansamblu") document for Ciortuz et al.'s exercise book, chapter *Decision Trees*. It is in fact a "road map" for what we will be doing here. (This *note* applies also to all chapters.)

Week $3.\frac{2}{2}$, 4:
decision trees and the ID3 algorithm: applications;
properties of decision trees;
analysis of the ID3 algorithm (as an algorithm *per se*):
Ciortuz et al.'s exercise book, ch. *Decision trees*, ex. 1-4, 8, 21a, 28-36, 47

• extensions to the ID3 algorithm:
– handling of attributes with many values:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 13
– handling of attributes with costs:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 14
– using other impurity neasures as local optimality criterion in ID3:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 15

Week 5:

• extensions to the ID3 algorithm:
– handling of continuous attributes:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 9-11, 39-41
– decision surfaces, decision boundaries:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 9, 40, and ch. *Instance-based learning*, ex. 11b

• analysis: ID3 as a Machine Learning algorithm;
– *inductive bias* for ID3:
[LC: a hierachical structure of the model/knowledge, compatibility/consistency with the data, and]
compactness of the resulting decision tree;
– error analysis/computation: training error, validation error, n-fold cross-validation, CVLOO
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 5-7, 21d, 37-38
– ID3 as eager learner:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 16
- ID3 and [non-]robustness to noises, and *overfitting*:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 9, 21bc, 40
– *pruning* strategies for decision trees:
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 19-20, 45-46

• other issues (optional):
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 12, 17-18, 42-44

Important Note:
Some of the exercises listed above (for weeks 4 and 5) would be done in class (i.e., at seminaries) in an easier/nicer way if students would priorily do at home the exercise 31, which askes for the **implementation** of the **information gain** (and also entropy, conditional specific entropy and conditional average entropy), starting form the counts (i.e., data partitions) associated to the leaf nodes of a **decision stump**. This implementation could be later extanded to an implementation of ID3 algorithm (the basic form); see ex. 58.

Implementation exercises:

0. CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2
Given a Matlab/Octave implementation for ID3, work on synthetic, noisy data, and study the relationship between model complexity, training set size, train and test accuracy;

1. CMU, 2012 spring, Roni Rosenfeld, HW3
Complete a given C (incomplete) implementation for ID3; work on a simple example (Play Tennis from TM's ML book) and on a real dataset (Agaricus-Lepiota Mushrooms); perform *reduced-error (top-down vs. bottom-up) pruning* to cope with *overfitting*.

CMU, 2011 spring, T. Mitchell, A. Singh, HW1, pr. 3
Similar to the above one, except that *pruning* a node is conditioned on getting at least an ϵ increase in accuracy; dataset: mushrooms.[3]

◦ CMU, 2008 spring, T. Mitchell, HW1, pr. 2
Asked for doing an ID3 implementation, including *reduced-error prunning* and *rule-based prunning*; work on a real dataset: German Credit Approval.
See the interesting Note at the end of the 'german-description.txt' file!

◦ CMU, 2009 spring, Tom Mitchell, HW1
Do an ID3 implementation, including *rule post-pruning*; work on a real dataset: predicting the votes in the US House of Representatives.

2. CMU, 2011 fall, T. Mitchell, A. Singh, HW1, pr. 2
Working with continuous attributes, complete a given a Matlab/Octave implementation for ID3, perform *reduced-error pruning*; implement another splitting criterion: the *weighted misclassification rate*; work on a real dataset: Breast Cancer.

[3] A similar exercise — CMU, 2011 spring, Roni Rosenfeld, HW3 — uses a chess dataset.

Week 6 and 7: Bayesian Classifiers

Read: Chapter 6 from Tom Mitchell's *Machine Learning* book (except subsections 6.11 and 6.12.2).

Week 6:

Bayes' theorem:
Ciortuz et al.'s exercise book, ch. *Foundations*, ex. 6-7, 43-45;

classes of machine learning hypotheses: MAP hypotheses vs. ML hypotheses:
Ciortuz et al.'s exercise book, ch. *Bayesian classification*, ex. 1-3, 12-13, 21, 32

application of Naive Bayes and Joint Bayes algorithms:

Ciortuz et al.'s exercise book, ch. *Bayesian classification*, ex. 4-7, 22-26

Week 7:
computation of the [training] error rate of Naive Bayes:
Ciortuz et al.'s exercise book, ch. *Bayesian classification*, ex. 8-10, 27-29;

some properties of Naive Bayes and Joint Bayes algorithms:
Ciortuz et al.'s exercise book, ch. *Bayesian classification*, ex. 11, 33;

comparisons with other classifiers:
Ciortuz et al.'s exercise book, ch. *Bayesian classification*, ex. 30-31.

Implementation exercises:

0. CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 4
Implement the Naive Bayes classification algorithm,
and perform CVLOO on a toy ("weather prediction") dataset;
do *feature selection* based on CVLOO.

1. Stanford, 2012 spring, Andrew Ng, pr. 6
Implement the Naive Bayes (train and test) algorithm;
use it as a spam filter on a subset of the Ling-Spam dataset.

2. CMU, 2011 spring, Tom Mitchell, HW2, pr. 3
Implement the Naive Bayes classification algorithm and perform n-ary classification on the *20 Newsgroups* datset;
for the $P(X_i|Y)$ parameters, do MAP estimation (instead of MLE) using as prior the Dirichlet distribution;[4]
identify the *key words* (for classification) using *conditional entropy*; analyse its effectiveness relative to the *information gain*.

Week 8: midterm EXAM

[4] An earlier exercise, CMU, 2009 spring, T. Mitchell, HW3, pr. 2, centered on Naive Bayes and the MAP estimation with Dirichlet prior, but instead of asking the student to perform n-ary classification (on the 20 newsgroups dataset), it limited itself to binary classification on a much simpler dataset: hokey vs. baseball newsgroups. A similar exercise — CMU, 2014f, W. Cohen and Z. Bar-Joseph, HW2, pr. 6 — uses a simple measure for feature (i.e. keyword) selection, and classifies texts from The Economist and The Onion.

Errata la
Exerciţii de învăţare automată, versiunea septembrie 2017

1. cap. *Fundamente*, ex. 2, pag. 23, rândul 12 de jos:
   orice $p \in (0, 1) \longrightarrow p = 1/2$

2. cap. *Fundamente*, ex. 33, pag. 66, rândul 6 de sus:
   d. Similar, $\longrightarrow$ Similar,

3. cap. *Fundamente*, ex. 34, pag. 69, rândul 8 de jos:
   $-\sum_x p(x) \log \frac{p(x)}{q(x)} \longrightarrow -\sum_x p(x) \log \frac{q(x)}{p(x)}$

4. cap. *Fundamente*, ex. 34, pag. 69, rândul 6 de jos:
   $KL(p||q) \geq 0 \longrightarrow KL(p||q) = 0$

5. cap. *Fundamente*, ex. 70, pag. 87, rândul 3 de sus:
   $CH(P_{true}, P_A)$, $CH(P_{true}, P_A) \longrightarrow CH(P_{true}, P_A)$, $CH(P_{true}, P_B)$

6. cap. *Arbori de decizie*, ex. 11, pag. 155, la al treilea compas de decizie, eticheta de pe ramura din dreapta:
   T $\longrightarrow$ N

7. cap. *Arbori de decizie*, ex. 11, pag. 155, în primul desen (mai precis, arborele de decize cu două niveluri de test), la al treila nod de decizie:
   $- \longrightarrow +$
   (Alexandra Minghel, studentă FII, anul III)

8. cap. *Arbori de decizie*, ex. 11, pag. 156, nota de subsol 57:
   322.5 $\longrightarrow$ 232.5 (la toate cele 4 apariţii)

9. cap. *Arbori de decizie*, ex. 11, pag. 157, rândul 1 de sus:
   Confrom $\longrightarrow$ Conform

10. cap. *Arbori de decizie*, ex. 22, pag. 181, rândul 15 de jos:
    cu sunt $\longrightarrow$ cum sunt

11. cap. *Arbori de decizie*, ex. 22, pag. 181, rândul 1 de jos:
    deciât $\longrightarrow$ decât

12. cap. *Arbori de decizie*, ex. 22, pag. 182, rândul 2 de sus:
    diminueaza probabilitatea alocată instanţelor incorect clasificate $\longrightarrow$ diminueaza probabilitatea alocată instanţelor corect clasificate
    (Lucian Nevoe, student FII, master)

13. cap. *Arbori de decizie*, ex. 22, pag. 184, rândul 4 de jos:
    $$\frac{\partial}{\partial \alpha_m} \longrightarrow \frac{\partial}{\partial \alpha_t}$$

14. cap. *Arbori de decizie*, ex. 23, pag. 185, în tabelul din dreapta jos:
    eticheta (adică y_i) pentru x_7 trebuie să fie -1 (în loc de $+1$)

15. cap. *Arbori de decizie*, ex. 23, pag. 186, rândul 5 de jos, exceptând notele de subsol:
    perchi $\longrightarrow$ perechi

16. cap. *Arbori de decizie*, ex. 23, pag. 187, în primul tabel, linia din mijloc, ultima coloană:
    $$\frac{2}{9} + \frac{2}{9} = \frac{2}{3} \longrightarrow \frac{4}{9} + \frac{2}{9} = \frac{2}{3}$$

17. cap. *Arbori de decizie*, ex. 23, pag. 187, în al doilea tabel, linia din mijloc, penultima şi ultima coloană:
$\frac{1}{9} + \frac{2}{9} = \frac{2}{9} \longrightarrow \frac{2}{9} + \frac{1}{9} = \frac{1}{3}$
şi respectiv
$\frac{1}{3} \longrightarrow \frac{2}{9}$

18. cap. *Arbori de decizie*, ex. 23, pag. 188, rândul 14 de jos:
$h1 \longrightarrow h_1$

19. cap. *Arbori de decizie*, ex. 23, pag. 189, rândul 1 de jos, exceptând formulele:
distrbuţii $\longrightarrow$ distribuţii

20. cap. *Arbori de decizie*, ex. 23, pag. 189, rândul 2 de sus (din *Observaţia* 2):
$X_2 \geq 7/2 \longrightarrow X_2 < 7/2$

21. cap. *Arbori de decizie*, ex. 23, pag. 190, rândul 1 de jos:
$X_1 \geq 5/2 \longrightarrow X_1 < 5/2$

22. cap. *Arbori de decizie*, ex. 2, pag. 231, rândul 24 de jos:
,,...uşă¡‘ $\longrightarrow$,,...uşă! “

23. cap. *Arbori de decizie*, ex. 2, pag. 231, rândul 14 de jos:
,,...iniţial¿‘ $\longrightarrow$,,...iniţial? “

24. cap. *Învăţare bazată pe memorare*, ex. 5, pag. 294, rândul 9 de sus:
sunt prezentate $\longrightarrow$ este prezentată

25. cap. *Învăţare bazată pe memorare*, ex. 9, pag. 303, rândul 14 de sus:
aplicaţille practice care $\longrightarrow$ aplicaţiile practice în care

26. cap. *Clusterizare*, ex. 11, pag. 348, rândul 6 de jos:
instanţle $\longrightarrow$ instanţele

27. cap. *Clusterizare*, ex. 11, pag. 362, rândurile 2 şi 4 de jos:
destribuţii $\longrightarrow$ distribuţii

28. cap. *Clusterizare*, ex. 11, pag. 362, rândul 3 de jos:
fiindca funţia $\longrightarrow$ fiindcă funcţia

29. cap. *Clusterizare*, ex. 30, pag. 390, rândul 19 de jos:
sunt situate $\longrightarrow$ sunt situaţi