ML course, 2020 fall
What you should know:

Week 1:

PART I (course): A brief introduction to Machine Learning

(slides 0-9, 19-24 from https://profs.info.uaic.ro/~ciortuz/SLIDES/ml0.pdf)

PART II (seminary): Revision: Basic issues in Probabilities[1]

(slides 3-16 from https://profs.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf)

Read: Chapter 2 (section 2.1) from the *Foundations of Statistical Natural Language Processing* book by Christopher Manning and Hinrich Schütze, MIT Press, 2002.[2]

PART II.1: Random events

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 (2020) ch. *Foundations*, ex. 1-5 [6-7] 8, 79-82 [83-84] 85

[1] Professors / teaching assistants who are in charge with the seminaries may decide to proceed through this revision on more than one week or, alternatively, to address these probabilities issues "by need", i.e. when required by the machine learning algorithms that students will learn / apply in class.

[2] 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.

Part II.2: Random variables [and a few basic probabilistic distributions]

Concepts/definitions:

- random variables;
  random variables obtained through function composition
- discrete random variables;
  probability mass function (p.m.f.)
  examples: Bernoulli, categorical, binomial [multinomial, geometric, Poisson] 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

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
- $E[X + Y] = E[X] + E[Y]$
  $E[aX] = aE[X]$

  Corollary: the *liniarity* of expectation:
  $E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i E[X_i]$.

  $Var[aX] = a^2\, Var[X]$
  $Var[X] = E[X^2] - (E[X])^2$

  $Cov(X, Y) = E[XY] - E[X]E[Y]$

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

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

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

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 9.ab, [9.c-11] 12-16, 21-23 [24-25] 86-88 [89] 90-92

Week 2.$\frac{1}{2}$: Introduction to Information Theory

Read: Chapter 2 (section 2.2) from the *Foundations of Statistical Natural Language Processing* book by Christopher Manning and Hinrich Schütze, MIT Press, 2002.
(slides 28-31 [32-33] from https://profs.info.uaic.ro/∼ciortuz/SLIDES/foundations.pdf)

Concepts/definitions:

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

Theoretical results/formulas:

- $0 \leq H(X) \leq H(\underbrace{1/n, 1/n, \ldots, 1/n}_{n \text{ times}}) = \log_2 n$
- $IG(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $IG(X;Y) \geq 0$
- $IG(X;Y) = 0$ iff X and Y are independent
  - $IG(X;X) = H(X)$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
  (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.

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

- computing different types of entropies:
  Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 51-53 [54] 56-57, 126 [127] 128;
- proof of some basic properties:
  Ciortuz et al.'s exercise book: ch. *Foundations*, ex. [50] 55, 58, 129-132, 136.

Weeks 2.$\frac{2}{2}$, 3 and 4: Decision Trees — illustrating basic issues in ML

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

Important Note:
See (i.e., do *not* skip!) the Overview (rom.: "Sumar") section for the *Decision Trees* chapter in Ciortuz et al.'s exercise book. It is in fact a "road map" for what we will be doing here. (This *note* applies also to all chapters.)

Week 2.$\frac{2}{2}$:

Decision trees and the **ID3 algorithm:**
applications;
analysis of the ID3 algorithm (as an algorithm *per se*);
properties of ID3 trees:
Ciortuz et al.'s exercise book, ch. *Decision trees*, ex. 1-9, 21.a, 31-42, 52 [53]

- **decision trees:**
seen as data structures: ex. 1, 7.b, 31
and as logic programs: ex. 2.e, 38.bc
- **ID3 algorithm:**
simple applications: ex. 2-3, 5, 36-37, 39-40
- **analysis of ID3 as an algorithm *per se*:**
recursive, divide-et-impera
greedy: ex. 4, 21.a, 38
search algorithm: ex. 3, 37
- **properties of ID3 trees:** ex. 2-4, 7-9, 21.a, 32, 37-38 [41] 52
- **implementation exercises:** ex. 34 [53]

Important Note:
Some of the exercises listed above would be done in class (i.e., at seminaries) in an easier / nicer way if students would priorily do at home the exercise 34, which asks for the **implementation** of the **information gain** (and also entropy, specific conditional entropy and average conditional entropy), starting form the counts (more precisely, from the data partitions) associated to the leaf nodes of a **decision stump**.[3] Alternatively, the exercise 35 advises the student on how to conveniently use a **pocket calculator** in order to calculate the above mentioned entropies and the information gain.

[3] This implementation could be later extanded to an implementation of ID3 algorithm (the basic form); see ex. 53.

Weeks 3-4:
extensions of the ID3 algorithm;
analysis of ID3 as a Machine Learning algorithm;
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 10-16 [17-18] 19 [20] 21, 43-46, [48-49] 50 [51]

• **extensions of the ID3 algorithm**
– handling of continuous attributes: ex. 10-12, 43-46
– decision surfaces, decision boundaries: ex. 10, 43, and ch. *Instance-based learning*, ex. 11.b

◦ **other extensions to the ID3 algorithm**
– handling of attributes with many values: ex. 13
– handling of attributes with costs: ex. 14
– using other impurity neasures as local optimality criterion in ID3: ex. 15
– reducing the greedy behaviour of the ID3 algorithm: ex. 17-18 [48-49]

• **analysis: ID3 as a Machine Learning algorithm**
– *inductive bias* for ID3:
[LC: a hierarchical structure of the model, compatibility/consistency with the data, and]
compactness of the resulting decision tree;
– error analysis/computation: training error, validation error, n-fold cross-validation, CVLOO:
ex. 6-8, 10, 21.d, 41-42, 44, 46.d
– ID3 as "eager" learner: ex. 16
– ID3 and [non-]robustness to noises, and *overfitting*: ex. 10, 21.bc, 45
– *pruning* strategies for decision trees: ex. 19 [20] [49] 50 [51]

Weeks 5-6: Bayesian Classifiers

Read:
Chapter 6 from T. Mitchell's *Machine Learning* book (except subsections 6.11 and 6.12.2);
(slides #3-5, 11-12, 14 in https://profs.info.uaic.ro/∼ciortuz/SLIDES/ml6.pdf)

Week 5: The Naive Bayes and Joint Bayes classifiers

- Bayes' theorem:
  Ciortuz et al.'s exercise book, ch. *Foundations*, ex. 6,7, 83-84;
- conditionally independent random [events,] variables:
  Ciortuz et al.'s exercise book, ch. *Foundations*, ex. 15-16 [88-89] 90-92;
- ◦ classes of machine learning hypotheses: MAP hypotheses vs. ML hypotheses:
  Ciortuz et al.'s exercise book, ch. *Bayesian classification*,[4] ex. 1-4, 24-25, 37;
- pseudo-code: ML book, page 177, and slide #14 in https://profs.info.uaic.ro/∼ciortuz/SLIDES/ml6.pdf
- **applications of Naive Bayes and Joint Bayes algorithms:** ex. 5-9, 26-30.

Week 6:

- **computation of the [training] error rate of Naive Bayes:** ex. 10-11 [12] 31-34;
- *sample complexity* of Naive Bayes and Joint Bayes: ex. 13;
- the nature of the *decision boundary* determined by Naive Bayes (and the relationship to logistic regression): ex. 14;
- ◦ comparisons with other classifiers: ex. 35-36;
- ◦ revision: ex. 38.

[4] This chapter is equally the source for all exercises listed below.

Week 7: Instance-Based Learning

Read: Chapter 8 from Tom Mitchell's *Machine Learning* book.

application of the k-NN algorithm:
Ciortuz et al.'s exercise book, ch. *Instance-based learning*, ex. 1-7, 12.ab, 15-23, 24.a, 26;

comparisons with the ID3 algorithm: ex. 11, 12.c, 13, 24.b;

Shepard's algorithm: ex. 8.

Week 8: midterm

Week 9: [Introduction to] The AdaBoost Algorithm

- pseudocode + intro. to theoretical foundations: ex. 22 [23];
  applications: ex. 24, 25.a, 54-58;
- revision: ex. 30, 66

Weeks 10-14: Clustering

Weeks 10-12: Hierarchical and Partitional Clustering

Read: Chapter 14 from Manning and Schütze' *Foundations of Statistical Natural Language Processing* book.

Week 10: Hierarchical Clustering:

ex. 1-6, 26-33 [34-36]

Weeks 11-12: Partitional Clustering: **The k-Means Algorithm**

See section 2.2 in the *Overview* of the *Clustering* chapter in Ciortuz et al.'s exercise book;
application: ex. 7-11, 15.a, 17.a, 21.a, 22.a, 37-38;
properties (convergence, optimality and other issues): ex. 12-13, 39-43;
k-Means for image compression: ex. 44;
using another distance metric (than the Euclidian one): ex. 45;
k-Means++: ex. 46;
a "kernelized" version of k-Means: ex. 47;
comparison with the hierarchical clustering algorithms: ex. 14, 48;
implementation: ex. 50.

Weeks 13-14: Model-based Clustering
Using the **EM algorithm** to solve **GMMs (Gaussian Mixture Models)**.

• **EM for GMM**, the **uni-variate case:**

pseudo-code: slide #30 in https://profs.info.uaic.ro/∼ciortuz/SLIDES/cluster.pdf
Note: it can be seen as a probabilistic version of the K-means algorithm expressed as in slide #67 in https://profs.info.uaic.ro/∼ciortuz/ML.ex-book/SLIDES/ML.ex-book.SLIDES.Cluster.pdf

Read: Tom Mitchell, *Machine Learning*, sections 6.12.1 and 6.12.3;
see section 3 in the *overview* of the *Clustering* chapter in Ciortuz et al.'s exercise book;

– Learning the means [of the Gaussians that build up the GMM]:
Ciortuz et al.'s exercise book, ch. *Clustering*, ex. 15, 16, 17.b, 51;

– Learning also [the] other parameters of the GMM:
Ciortuz et al.'s exercise book, ch. *Clustering*, ex. 17.c, 18, 25.a, 52-55.

• **EM for GMM**, the **multi-variate** case, when the **covariance matrices** (Σ_k) are **diagonal** (i.e., the variables are mutually independent)

Ciortuz et al.'s exercise book, ch. *Clustering*, ex. 21.b, 22.b, 56-57, 63.

• **The EM algorithmic schema:**

Read: Tom Mitchell, *Machine Learning*, section 6.12.2;
Slide #30 in https://profs.info.uaic.ro/∼ciortuz/SLIDES/cluster.pdf
Ciortuz et al.'s exercise book, ch. *Clustering*, ex. 19, 25.b.

Prerequisites:

– The **Gaussian distribution:** See *Advanced issues...*

– The **likelihood function:** See *Advanced issues...*

– **Mixtures of probabilistic distributions:**
Ciortuz et al.'s exercise book, ch. *Foundations*, ex. 103, 26, 107 [36, 108].

Weeks 15-16: [final] EXAM