ML course, 2019 fall
What you should know:

Week 1, 2: Basic issues in Probabilities

Read: Chapter 2 (section 2.1) from the *Foundations of Statistical Natural Language Processing* book by Christopher Manning and Hinrich Schütze, MIT Press, 2002.[1]

Week 1:

PART I: A brief introduction to Machine Learning

(slides 0-9, 20-24 from https://profs.info.uaic.ro/∼ciortuz/SLIDES/ml0.pdf)

PART II: 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 (2019) ch. *Foundations*, ex. 1-5 [6-7], 8, 61-64 [65-66], 67

Advanced issues:

- Taking *A self evaluation test for the ML course*, CMU, 2014 fall, W. Cohen: http://www.cs.cmu.edu/∼wcohen/10-601/self-assessment/Intro_ML_Self_Evaluation.pdf

◦ Similar tests:
http://www.cs.cmu.edu/∼ninamf/courses/601sp15/hw/homework1.pdf (CMU, 2015 spring, N. Balcan)
http://curtis.ml.cmu.edu/w/courses/images/8/88/Homework1.pdf (CMU, 2016 spring, W. Cohen, N. Balcan)
http://www.cs.cmu.edu/∼mgormley/courses/10601b-f16/files/hw1_questions.pdf (CMU, 2016 fall, N. Balcan, M. Gormley)

[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 a few basic probabilistic distributions]

(slides 7-16 from https://profs.info.uaic.ro/$\sim$ciortuz/SLIDES/foundations.pdf)

Concepts/definitions:

- random variables;
  random variables obtained through function composition
- discrete random variables;
  probability mass function (p.m.f.)
  examples: Bernoulli, binomial [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. 9-15, 19, 20-21, [22] 68-75, 81, 82-83

Advanced issues (I):

Concepts/definitions:

- cumulative function distribution
- continuous random variables;
  probability density function (p.d.f.)
  examples: Gaussian, exponential, [Gamma, Beta, Laplace] distributions

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 25-28, 76-78

Theoretical result:

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

Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 18

Advanced issues (II):

- the *likelihood function* (see *Estimating Probabilities*, additional chapter to the *Machine Learning* book by Tom Mitchell, 2016)
- the *Bernoulli distribution*: MLE and MAP estimation of the parameter

Week $3.\frac{1}{2}$: 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. 35-36, 39, 92;
- proof of some basic properties:
Ciortuz et al.'s exercise book: ch. *Foundations*, ex. [33] 34, 40, 93-94 [95], 96, 98.

Advanced issues:

- cross-entropy
Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 41-42, 97;
- relative entropy (Kullback-Leibler divergence)
Ciortuz et al.'s exercise book: ch. *Foundations*, ex. 38.

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

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

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, 22.a, 29-40, 52 [53]

- **decision trees:**
seen as data structures: ex. 1, 7.b, 29
and as logic programs: ex. 2.e, 36.bc
- **ID3 algorithm:**
simple applications: ex. 2-3, 5, 34-35, 37-38
- **analysis of ID3 as an algorithm *per se*:**
recursive, divide-et-impera
greedy: ex. 4, 22.a, 36
search algorithm: ex. 3, 35
- **properties of ID3 trees:** ex. 2-4, 7-9, 22.a, 30, 35-36 [39] 52
- **implementation exercises:** ex. 32 [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 32, 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**.[2] Alternatively, the exercise 33 advises the student on how to conveniently use a **pocket calculator** in order to calculate the above mentioned entropies and the information gain.

Implementation exercises:

CMU, 2012 spring, Roni Rosenfeld, HW3
- Complete a given C (incomplete) implementation for ID3.
- Work firstly on a simple example (Play Tennis from Tom Mitchell's *Machine Learning* book) and secondly on a real dataset (Agaricus-Lepiota Mushrooms).
- ○ Perform *reduced-error (top-down vs. bottom-up) pruning* to cope with *overfitting*.[3]

  Note: CMU, 2011 spring, Roni Rosenfeld, HW3 – a similar problem to the above one — uses a *chess* dataset generated by Alen Shapiro (see *Structured Induction in Expert Systems*, 1983, 1987).

[2] This implementation could be later extanded to an implementation of ID3 algorithm (the basic form); see ex. 53.
[3] CMU, 2011 spring, T. Mitchell, A. Singh, HW1, pr. 3 is similar to the above problem, except that *pruning* a node is conditioned on getting at least an ε increase in accuracy. (Dataset: mushrooms.)

Week 4:
extensions of the ID3 algorithm;
analysis of ID3 as a Machine Learning algorithm;
Ciortuz et al.'s ex. book, ch. *Decision trees*, ex. 10-12, 14-17 [18-19] 20 [21] 22, 42-45, [48-49] 50 [51]

• extensions of the ID3 algorithm
− handling of continuous attributes: ex. 10-12, 42-45
− decision surfaces, decision boundaries: ex. 10, 42, and ch. *Instance-based learning*, ex. 11.b

◦ other extensions to the ID3 algorithm
− handling of attributes with many values: ex. 14
− handling of attributes with costs: ex. 15
− using other impurity neasures as local optimality criterion in ID3: ex. 16
− reducing the greedy behaviour of the ID3 algorithm: ex. 18-19 [48-49]
− other splitting criteria (optional): ex. 13, 46-47

• 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, 22.d, 39-40, 43, 44.d
− ID3 as "eager" learner: ex. 17
− ID3 and [non-]robustness to noises, and *overfitting*: ex. 10, 22.bc, 43
− *pruning* strategies for decision trees: ex. 20 [21] [49] 50 [51]

Implementation exercises:

CMU, 2011 fall, T. Mitchell, A. Singh, HW1, pr. 2

- Working with continuous attributes on a real *dataset*: Breast Cancer.
- Complete a given a Matlab/Octave implementation for ID3.
- Perform *reduced-error pruning*.
- Implement another splitting criterion: the *weighted misclassification rate*.

Week 5: The AdaBoost Algorithm

Weeks 6-7: 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 8: midterm

Week 9: Instance-Based Learning

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

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 13:

The **likelihood function**, and the Maximum Likelihood method for estimating parameters of probabilistic distributions (abbrev., **MLE**)

Read: *Estimating Probabilities*, additional chapter to the *Machine Learning* book by Tom Mitchell, 2016.

Week 14: Model-based Clustering
Using the **EM algorithm** to solve **GMMs (Gaussian Mixture Models), the uni-variate case**.

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;

Weeks 15-16: [final] EXAM