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. 1, 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
- $E[X + Y] = E[X] + E[Y]$
  $E[aX + b] = aE[X] + b$
  $Var[aX] = a^2\, Var[X]$
  $Var[X] = E[X^2] - (E[X])^2$
  $Cov(X, Y) = E[XY] - E[X]E[Y]$
- 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. 1, 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. 1, ex. 9-16 [17-22], 46-55 [57-63], 64

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.

2. CMU, 2014 fall, William Cohen, Ziv Bar-Jospeh, HW3, pr. 1
Implement *Logistic Regression* and apply it to the task of hand-written character recognition.[3]

   ◦ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW3, pr. 1.4-6
   Implement *Multinomial Logistic Regression* and apply it to the *ORL Faces* dataset [while pr. 1.1-3,7 compares (M)LR with K-NN, Gaussian Naive Bayes and Gaussian Joint Bayes on this dataset].

3. CMU, 2010 fall, Ziv Bar-Jospeh, HW2, pr. 4
Study the regularization effect for logistic regression, using L2 and (especially) L1 norm, especially for *feature selection*. Work on the *communities and crime* dataset.

[3] A similar exercise, CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 4.1-2, applies logistic regression (LR) on the *Breast Cancer* dataset [while pr. 4.3-4 compares LR with the Rosemblatt *perceptron* on this dataset].

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:

- $0 \leq H(X) \leq H(\underbrace{1/n, 1/n, \ldots, 1/n}_{n \text{ times}}) = \log_2 n$
- $IG(X;Y) \geq 0$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ (generalisation: the chain rule)
- $IG(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $H(X,Y) = H(X) + H(Y)$ iff X and Y are indep.
- $IG(X;Y) = 0$ 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. 1, ex. 31, 32, 35, 65 [37, 70]);
- proof of some basic properties (see **Ciortuz et al.'s exercise book:** ch. 1, ex. 29, 33, [33, 34,] 36, [38,], 66-69, 71), 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 3. 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 their properties; application of ID3 algorithm, properties of ID3:
Ciortuz et al.'s exercise book, ch. 3, ex. 1-8, 28-38

Errata la

Exerciţii de învăţare automată, versiunea septembrie 2017

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

2. cap. 1, pr. 33, pag. 66, rândul 6 de sus:
   d. Similar, $\longrightarrow$ Similar,

3. cap. 1, pr. 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. 1, pr. 34, pag. 69, rândul 6 de jos:
   $KL(p||q) \geq 0 \longrightarrow KL(p||q) = 0$

5. cap. 1, pr. 70, pag. 87, rândul 3 de sus:
   $CH(P_{true}, P_A)$, $CH(P_{true}, P_A) \longrightarrow CH(P_{true}, P_A)$, $CH(P_{true}, P_B)$