Department of Mathematics

Técnicas Matemáticas para Big Data

Data: 15 of December of 2024
Duration: 2h30
Obs: Every question should be justified. Any code presented should be written in the exam sheet in Python or pseudo-code.

Exam (version A)

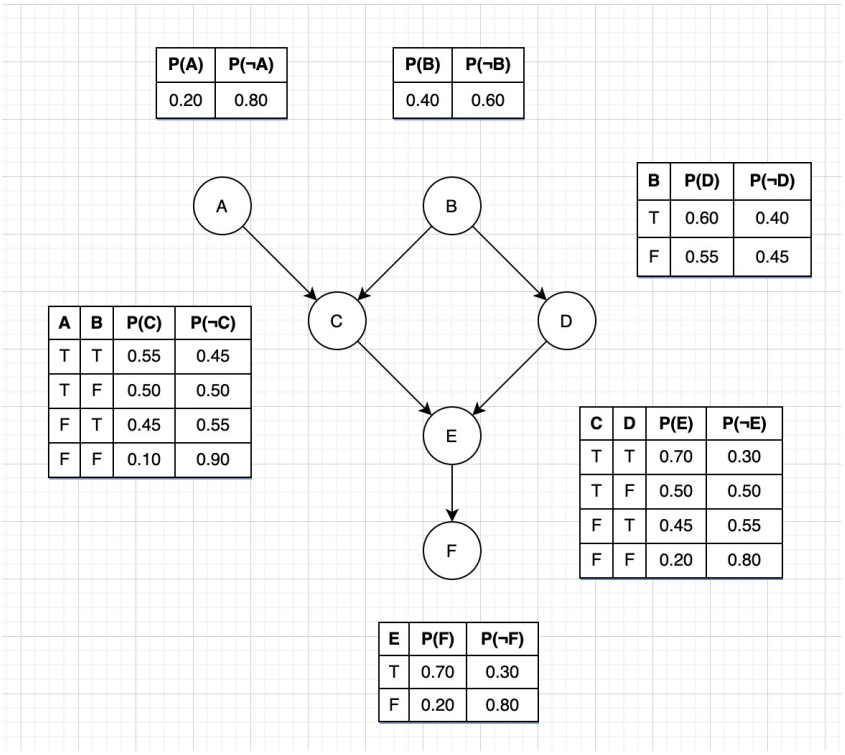
(3.0) **1.** Answer the following questions:

(a) Explain how relevant is a family of hash functions to a Bloom filter algorithm.

(b) State the CAP theorem and its consequences on storing/accessing geographic data.

(c) Identify the different perspectives of Big Data, in particular, explain in detail the perspective $HighVelocity$.

(4.5) **2.** Consider the following Bayesian network:



P(A)	P(¬A)
0.20	0.80

P(B)	P(¬B)
0.40	0.60

B	P(D)	P(¬D)
T	0.60	0.40
F	0.55	0.45

A	B	P(C)	P(¬C)
T	T	0.55	0.45
T	F	0.50	0.50
F	T	0.45	0.55
F	F	0.10	0.90

C	D	P(E)	P(¬E)
T	T	0.70	0.30
T	F	0.50	0.50
F	T	0.45	0.55
F	F	0.20	0.80

E	P(F)	P(¬F)
T	0.70	0.30
F	0.20	0.80

(a) Determine the probability of an event X with probability given by $P(A = true, B = false, C = true, D = false, F = true)$. Notice that X depends indirectly on E.;

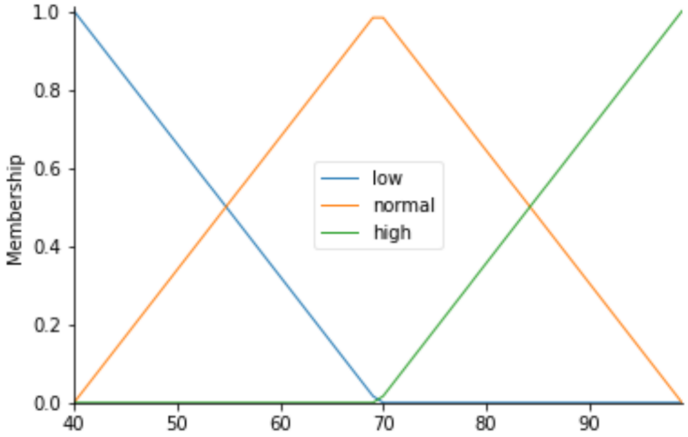
(b) Expalin if the above network can be seen as a Hidden Markov Model.

(1.0) **3.** For a dataset with 3 million records, consider a Machine Learning method based on weights with an update method

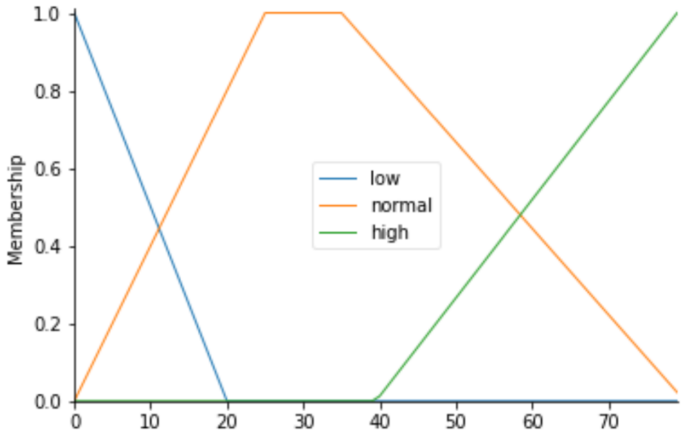
$$w_{k+1} = w_k + h\ (3x - w_k)^2\ f(k, x),$$

for some function $1 < f(k, x) < 2$, $k \in \{1, 2, \dots\}$ and $x \in [0, 5]$. How can you compute the value of the constant $h > 0$ to ensure that the method converges? Write the associated differential equation for the Euler method and explain how can be used to improve the original method.

(4.5) **4.** Consider the Fuzzy sets, related to the operation of a heavy machine, with variables *Consumption* and *Performance* with graphs:



Comsumption



Performance

The graphs end in the values 100 and 80, respectively.

(a) Prove that $S(a, b) = (a + b - 2ab)/(1 - ab)$ is a S-norm.

(b) Let $CO2\ emission$ be a consequence Fuzzy set with a graph similar to *Consumption* but multiplied by the factor 2.0. Determine the region of $CO2\ emission$ for the rule:

IF x is $Consumption['normal']$ OR y is $Performance['high']$ THEN z is $CO2\ emission['high']$

when $x = 60$, $y = 70$ and using the function S;

(c) Which is the approximate value of $CO2\ emission$, by using the "min of maximum". Justify.

(4.0) **5.** Consider the following stream $S = [-1.2, -1.2, 16.3, 4.4, 16.3, -1.2, -1.2, 6.8, 2.6]$, where the value 2.6 corresponds to the current time instant, the value 6.8 corresponds to the previous time instant, and so forth.

(a) With the values of S, write a streaming algorithm to determine the outliers by a Z-score $[\mu - \alpha\sigma, \mu + \alpha\sigma]$, with constant $\alpha = 3.7$ instead of the constant $\alpha = 3.5$.

(b) With all the values of S, determine if there are lower outliers below the barrier $Q_{0.25} - \alpha(Q_{0.75} - Q_{0.25})$ with $\alpha = 1.9$ and where Q_p is the quantile of order $p \in [0, 1]$.

(3.0) **6.** Consider the following questions.

(a) In which conditions, you may choose to apply Self Organizing Maps (SOM) instead of t-SNE.

(b) Explain the different benefits between a Principal Component Analysis approach or a t-SNE approach for reducing data into two dimensions.

(c) What happen to the volume of a sphere when the dimension highly increases? What are the expect consequences to the results obtained by some machine learning algorithms?