Data Engineering: Exercises.

1. Discuss the area of Data Engineering in comparison to Data Mining.
2. Explain the concept of 'the graph of a system'.
3. Define two observables to model the equation(s) of motion for a particle (mass point) governed by Newton's second law.

   - Define and illustrate the equivalence classes and quotient sets induced in the phase-space.
4. Derive the equation of the probability $Pr(\cdot)$ of the (crisp) event $A \subset X$, by using the definition of the expectation operator $E[\cdot]$.

   - Illustrate or describe why/how probability theory and measure theory are related.
5. i. State the definition of the expectation operator, i.e the expectation of a real-valued function f defined on X, with respect to $g(x)$.

   ii. From i. derive the mean value of a random variable $\mathbf{x}$.

   iii. From i. derive the probability of an event A where A is a crisp subset, in X, described by $\zeta_A \colon X \to \{0, 1\}$.

   iv. Generalise the result from iii. for the probability of the fuzzy event $A = \{(x, \mu_A(x)) \colon x \in X\}$ and provide an interpretation for the equation that determines the probability of a fuzzy event.

   v. Let the fuzzy event A be described by the following membership function :

$$\mu_A(x) = \begin{cases} e^{-(4-x)} & \text{if } 0 \leq x \leq 4, \\ e^{-\left(\frac{x-4}{2}\right)} & \text{if } x > 4. \end{cases}$$

   Calculate (with step-by-step solution) a numerical value for the probability $Pr(A)$ if the probability density function on X is defined by

$$p(x) = \begin{cases} 2 \cdot e^{-2x} & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

6. Explain the difference between 'descriptive' and 'inferential' statistics.
7. Consider three data points $\mathbf{m}_1 = (-4, 0)$, $\mathbf{m}_2 = (1, 3)$, $\mathbf{m}_3 = (3, 6)$, find the least-squares parameter estimates $\hat{\boldsymbol{\theta}}$ for the model $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{E}$ and determine the vector of residuals, $\mathbf{E}$.
8. Describe the principle idea of maximum likelihood estimation and define the 'likelihood function'.
9. Explain the EM-algorithm (it is not necessary to state all equations).
10. State the definitions and discuss the differences of

    - the partition spaces M_{hc}, and M_{fc},
    - the objective functions J_{hc} and J_{fc}.

    where the indices hc and fc refer to hard-c-means and fuzzy-c-means clustering respectively.
11. Outline the steps required in preparation and execution of the fuzzy-c-means algorithm.
12. Define and compare the following fuzzy model structures

    - Linguistic fuzzy model,
    - Takagi-Sugeno fuzzy model.
13. For the affine-linear Takagi-Sugeno model we are given

$$\boldsymbol{\theta}_1 = [0.6, 2]^T \quad \text{and} \quad \boldsymbol{\theta}_2 = [0.2, 9]^T$$
$$\mu_{A_1}(x) = 0.3 \quad \text{and} \quad \mu_{A_2}(x) = 0.8\ .$$

    Define (state) the rule-based system and calculate its output value $y = f(x)$ for $x = 13$.

14. For the 'product-inference engine' with singleton input data and 'centre-average' defuzzification, calculate the output value of the system. Three triangular output fuzzy sets have their peaks (for which $\mu(y) = 1$), at $y_0^{(1)} = -2$, $y_0^{(2)} = 0$, and $y_0^{(3)} = 2$. There are two input variables for which the following degree of membership are known:

$$\mu_{A_{11}}(x) = 0.5, \qquad \mu_{A_{12}}(x) = 1.0, \qquad \mu_{A_{21}}(x) = 0.10\ ,$$
$$\mu_{A_{22}}(x) = 0.5, \qquad \mu_{A_{31}}(x) = 0.3, \qquad \mu_{A_{32}}(x) = 0.95\ .$$

15. Prove that the following De Morgan law is true for fuzzy sets : $(A \cup B)^c = A^c \cap B^c$, where A and B are fuzzy sets defined by their membership functions μ_A and μ_B and ^{c} denotes the complement. Use the "Zadeh definitions" for union, intersection and complement. State results and calculations in terms of membership functions.

16. Let $\mu_A(x) = 0.3$ and $\mu_B(y) = 0.7$, calculate $\mu_R(x, y)$, the degree of membership of fuzzy relation are using the Zadeh implication.

17. Fuzzy sets are commonly used to describe fuzzy concepts. For example, let X be a reasonable age interval of human beings : $[0, 100]$. Assume that the concept of "young" is represented by a fuzzy set Y whose membership function is :

$$\mu_Y(x) = \begin{cases} 1 & \text{if } x \leq 25, \\ (40 - x)/15 & \text{if } 25 < x < 40, \\ 0 & \text{if } x \geq 40. \end{cases}$$

Let the concept "old" be represented by a fuzzy set O whose membership function is :

$$\mu_O(x) = \begin{cases} 0 & \text{if } x \leq 50, \\ (x - 50)/15 & \text{if } 50 < x < 65, \\ 1 & \text{if } x \geq 65. \end{cases}$$

i. State the set-theoretic operation and determine the equation for the membership function representing the concept "neither young nor old" (="middle aged").
ii. Show that "old" implies "not young".
iii. Determine the equations and sketch graphically **all** membership functions involved, including those of complement sets and the sets asked for in the two previous items.

18. Consider the two fuzzy relations $R_1(x, y)$ and $R_2(y, z)$ as defined below. Determine the max-min-composition on $X \times Z$ for $x = x_1, z = z_1$ and $y = y_i$, where $i = 1, \ldots, 5$.

$$R_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccccc} y_1 & y_2 & y_3 & y_4 & y_5 \end{array} \\ \begin{pmatrix} 0.1 & 0.2 & 0.0 & 1.0 & 0.7 \\ 0.3 & 0.5 & 0.0 & 0.2 & 1.0 \\ 0.8 & 0.0 & 1.0 & 0.4 & 0.3 \end{pmatrix} \end{array} \qquad R_2 = \begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{pmatrix} 0.9 & 0.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.8 & 0.0 \\ 0.8 & 0.0 & 0.7 & 1.0 \\ 0.4 & 0.2 & 0.3 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.8 \end{pmatrix} \end{array}$$

19. Let $f(x) = x^2 - 3$, and fuzzy set

$$A = \{(-2, 0.1), (-1, 0.4), (0, 0.8), (1, 0.9), (2, 0.3)\}\ .$$

Use the extension principle to obtain fuzzy set B from A.

20. Explain the principle idea of the 'extension principle' and illustrate it using a drawing and equations (for the simple case of $y = f(x)$ where f is not injective and the argument is fuzzy set $A \subset X$ while $B \subset Y$.

21. Suppose f is a function mapping ordered pairs from

$$U_1 = \{-1, 0, 1\} \quad \text{and} \quad U_2 = \{-2, 2\} \qquad \text{to} \qquad V = \{-2, -1, 2, 3\}\ ,$$

and

$$f(x_1, x_2) = x_1^2 + x_2\ .$$

Let A_1 and A_2 be fuzzy sets defined on U_1 and U_2 respectively, such that

$$A_1 = \{(-1, 0.5), (0, 0.1), (1, 0.9)\} \qquad \text{and} \qquad A_2 = \{(-2, 0.4), (2, 1)\} \ .$$

Use the extension principle to derive $f(A_1, A_2)$, the membership function of the fuzzy set B. (Illustrate the calculation process with a table.)

22. Let $U = \{\text{Reading}, \text{Manchester}, \text{Stockport}\}$ and $V = \{\text{London}, \text{Manchester}\}$. Using a fuzzy relation to define the relational concept "very far", between the two sets of cities, we obtain the following (fuzzy) relational matrix :

	London	MCr
Reading	0.3	0.9
MCr	1	0
Stockport	0.95	0.1

Use the following notation: $y \in V, x \in U, z \in W$!

i. State the formal definition of the projection of fuzzy relation Q in $U_1, \times \cdots \times U_n$ on $U_{i1} \times \cdots \times U_{ik}$ defined by the membership function $\mu_{Q_P}(u_{i1}, \ldots, u_{ik})$.

ii. From i., and from fuzzy relation "very far", determine the projections Q_1 and Q_2 on U and V.

iii. State the formal definition of the extension of fuzzy relation Q_P in $U_{i1} \times \cdots \times U_{ik}$ to $U_1 \times \cdots \times U_n$ defined by the membership function $\mu_{Q_{PE}}(u_1, \cdots, u_n)$.

iv. From iii. determine the extensions of Q_{1E} and Q_{2E} of Q_1 and Q_2 in ii. to $U \times V$.

v. From ii. and iv., taking the projection of a fuzzy relation and then extending it, what can be observed regarding the obtained fuzzy relation?

vi. Let $W = \{\text{Sheffield}, \text{Oxford}\}$ and let $P(U, V)$ denote the fuzzy relation "very far", defined above. The fuzzy relation "very near" in $V \times W$, denoted $Q(V, W)$, is defined by the relational matrix :

	Sheffield	Oxford
London	0.95	0.1
MCr	0.1	0.9

   i. Write out P and Q as finite sets.

   ii. State the definitions of the max-min composition and max-product composition of P and Q and compute both compositions. Note that $U \times W$ contains six elements: (Reading,Sheffield), (Reading,Oxford), (MCr,Sheffield), (MCr,Oxford), (Stockport,Sheffield), (Stockport,Oxford). State final $P \circ Q$ matrices and step-by-step solutions for $\mu_{P \circ Q}$(Reading,Sheffield) and $\mu_{P \circ Q}$(Reading,Oxford) using the max-min composition.

23. Let the fuzzy relation R on $X \times X$ be defined as

$$R = \big\{[(x_1, x_1), 0.2], [(x_1, x_2), 1], [(x_2, x_1), 0.1], [(x_2, x_2), 0.6]\big\}$$

Show that the max-min composition $R \circ R$ is transitive with respect to set-inclusion ($A \subseteq B$ if and only if $\mu_A(x) \leq \mu_B(x)$).

24. Let U_1, U_2 be two reference sets (universe of discourse) defined as

$$U_1 = \{a, b, c, d, e, f\} \qquad U_2 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

and fuzzy relation R is defined by

$$R = \big\{[(a, 1), 0.1], [(b, 1), 0.4], [(b, 2), 0.3], [(b, 5), 0.2], [(d, 1), 0.9], [(d, 2), 0.7], [(e, 1), 0.8], [(e, 2), 0.7], [(e, 5), 0.1]\big\}$$

Find the two projections R_1 and R_2 onto the subspaces U_1 and U_2.

25. Let the fuzzy relation "almost equal" be defined as follows:

$$R_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0.8 & 0.0 & 0.0 \\ 0.8 & 1 & 0.8 & 0.0 \\ 0.0 & 0.8 & 1.0 & 0.8 \\ 0.0 & 0.0 & 0.8 & 1 \end{pmatrix} \end{array}$$

What is the membership function of the fuzzy set $B =$"rather small integer" if it is interpreted as the composition $A \circ R$?

26. Given fuzzy sets $\mu_{A'}(x)$, $\mu_B(y)$, and fuzzy relation $\mu_R(x, y)$, consider the modus ponens inference rule:

Implication:	IF **X** is A, THEN **Y** is B.
Premise:	**X** is A'.
Conclusion:	**Y** is B'.

Describe

i. the basic idea of the compositional rule of inference.

ii. in three steps how to obtain the membership function $\mu_{B'}(y)$ using the compositional rule of inference.

27. Consider an evidential reasoning system considering two bodies of evidence, represented by the two spaces $R = \{r_1, r_2\}$, and $S = \{s_1, s_2\}$ with given probabilities $p(r_1) = 0.8$ and $p(s_1) = 0.4$, respectively. The problem is described by two variables **X** and **Y** that take their possible values from the spaces R, S and $T = [0, 100]$, respectively. The set of rules that associate evidential elements to hypotheses are:

IF **X** is r_1, THEN **Y** in $A_1 = [0\ , 30]$.
IF **X** is r_2, THEN **Y** in $A_2 = [0\ , 90]$.

IF **X** is s_1, THEN **Y** in $B_1 = [0\ , 45]$.
IF **X** is s_2, THEN **Y** in $B_2 = [38, 80]$.

(a) Determine the probability assignments m_R and m_S on T.

(b) Calculate the probability interval for $D = [0, 35]$, considering only the evidence provided on R.

(c) Calculate the probability interval for $D = [0, 35]$, considering only the evidence provided on S.

(d) Determine the combined probability assignment $m_R \oplus m_S$ for the two independent evidential sources.

(e) Determine the probability interval, that is, the belief and plausibility measure regarding the hypothesis $D = [0, 35] \subset T$.

(f) Compare the results from c), and d), with e).

28. Describe the structure and operation of a fuzzy logic control system, explaining the functions of the fuzzifier, defuzzifier, fuzzy rule-base and fuzzy inference. Discuss the alternative mathematical possibilities for implementing the fuzzy inference and defuzzification procedures.

29. In fuzzy modeling of dynamic systems the unknown system $y = f(\mathbf{x})$ generates data for $y(t)$ and $\mathbf{x}(t)$ measured at $t = k, k-1, \ldots$ then the aim is to construct a model that is a reasonable approximation for $y = f(\mathbf{x})$, where $f(\cdot)$ is unknown. Let $\mathbf{x} = [x_1, \ldots, x_r]^T$ and $y \in \mathbb{R}$. The model is represented as a collection of IF-THEN fuzzy rules.

a) Describe the Mamdani or linguistic fuzzy model:

i. Model structure: general form and 'conjunctive form'.

ii. Simplified inference scheme (in four steps): 'fuzzification' to obtain β_i, $1 \le i \le n_R$; output fuzzy sets B'_i; 'rule-aggregation'; 'defuzzification' using the 'centre of gravity' (COG) method.

(Denote with N_q the number of discretised values y_q in Y, with p the number of fuzzy sets per rule. Use the min-operator for implication.)

b) Describe the Takagi-Sugeno fuzzy model.

i. Rule structure: general form and 'affine linear form'.

ii. Inference mechanism: extending the COG defuzzification for the 'singleton model'.

c) Determine the numerical output of y for the following model :

$$R_1\colon \text{IF } x_1 \text{ is } A_{11} \text{ AND } x_2 \text{ is } A_{12} \text{ THEN } y = x_1 + x_2\ ,$$
$$R_2\colon \text{IF } x_1 \text{ is } A_{21} \text{ THEN } y = 3x_1\ ,$$
$$R_3\colon \text{IF } x_2 \text{ is } A_{32} \text{ THEN } y = 2x_2\ .$$

Let $x_1 = 12$, $x_2 = 5$, with the structure is as in b) and

$$\mu_{A_{11}}(x_1) = \begin{cases} 1 & x_1 < 0 \\ -0.0625x_1 + 1 & 0 \leq x_1 \leq 16 \\ 0 & x_1 > 16\ . \end{cases}$$

$$\mu_{A_{12}}(x_2) = \begin{cases} 1 & x_2 < 0 \\ -0.125x_2 + 1 & 0 \leq x_2 \leq 8 \\ 0 & x_2 > 8\ . \end{cases}$$

$$\mu_{A_{21}}(x_1) = \begin{cases} \min(0, 1 - |\frac{20-x_1}{10}|) & 10 \leq x_1 \leq 20 \\ 0 & \text{otherwise}\ . \end{cases}$$

$$\mu_{A_{32}}(x_2) = \begin{cases} 0 & x_2 \leq 2 \\ 1 & x_2 \geq 20 \quad \text{(straight line in between).} \end{cases}$$

30. Consider the affine-linear Takagi-Sugeno fuzzy model for two clusters, $c = 2$. Given the set of data $\{(1,1),(2,1),(3,3)\}$, the fuzzy c-means clustering algorithm returned the partition matrix $\mathbf{U}$ with $u_{11} = 1$, $u_{22} = 0$, $u_{13} = 1$.. .

    i. State and complete $\mathbf{U}$ in matrix form. Comment on the partition matrix.

    ii. State and calculate the least squares solution for consequence parameter vector $\boldsymbol{\theta}_i = [\mathbf{a}_i^T, b_i]^T$. Provide all matrices and steps involved. State first general solutions, notations and only then apply equations to given numerical values. Provide step-by-step solutions for matrix operations.

    iii. From the solutions $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, calculated before, state $\mathbf{a}_1$ and b_1 and comment on the rule-base R_i.

    iv. Let $x_1 = 1.5$ and $A_1 = \exp\left(-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2\right)$ where $c = 2$ and $\sigma = 1$. State the inference equation for the Takagi-Sugeno model in general and calculate the output for the given values.

31. Describe the procedure for using fuzzy logic to construct a control law from a set of linguistic rules relating the error and change in error to change of control input. Explain how a deterministic control input is generated from the membership function of a fuzzy set, and show that the resulting control law can be regarded as a non-linear PI controller.

32. The iterative fuzzy c-means clustering algorithm is used to identify the premise portion of a IF-THEN rule-based fuzzy system with two rules. The following data vectors are given :

$$\mathbf{m}_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \qquad \mathbf{m}_2 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \mathbf{m}_3 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

    The partition matrix for $l = 1$ is

$$\mathbf{U} = \begin{bmatrix} ? & 0.9197 & ? \\ 0.3271 & ? & 0.7746 \end{bmatrix}$$

    Let w be 2 and the initial cluster clusters, randomly chosen be

$$\mathbf{c}_1^{(l=0)} = \begin{bmatrix} 1.89 \\ 3.76 \end{bmatrix} \qquad \mathbf{c}_2^{(0)} = \begin{bmatrix} 2.47 \\ 4.76 \end{bmatrix}$$

    i. Describe the effect/role of the weighting factor.

    ii. Calculate the missing elements in the partition matrix $\mathbf{U}$. Note: Provide a step-by-step solution only for u_{11}!

    iii. Calculate the cluster centres $\mathbf{c}_i^{(l)}$ for $l = 1$.