Exam: Principles of Machine Learning / Beginselen van Machine Learning

This document contains part of an earlier exam. (Not all questions have been included.)
Dit is een gedeelte van een eerder gehouden examen. (Niet alle vragen zijn behouden.)

This exam is bilingual. Guidelines and questions are identical in both languages.
Dit examen is tweetalig. De richtlijnen en vragen zijn identiek in beide talen.

Richtlijnen — belangrijk!

1. Dit examen is gesloten boek. Je mag enkel het materiaal gebruiken dat op het examen voorzien wordt, en een eenvoudige wetenschappelijke rekenmachine.
2. Lees elke vraag aandachtig. Antwoord alleen op de vraag, geef geen bijkomende informatie die niet gevraagd is (irrelevante informatie geven kan een lagere score opleveren).
3. Beantwoord elke vraag op een **duidelijke, gestructureerde** manier. Antwoord **bondig**, **precies** en **ter zake**. Schrijf geen volzinnen als bv. een oplijsting van termen even duidelijk is.
4. Alle antwoorden moeten op deze pagina's gegeven worden. De voorziene ruimte volstaat. Waar een antwoordkader gegeven wordt, moet het antwoord binnen dat kader opgeschreven worden. Waar geen kader voorzien is, mag alle lege ruimte op de bladzijde gebruikt worden.
5. Waar een maximale lengte (aantal woorden) opgegeven wordt, moet je je daaraan houden. Te lange antwoorden kunnen tot een lagere score leiden. Tekeningen en wiskundige formules tellen niet mee als woorden.
6. Het examen bestaat uit 3 delen. Deel 1 hangt aan deze pagina vast. Delen 2 en 3 zijn apart samengeniet. Geef alle delen tegelijk af.
7. Het examen duurt 4 uur.

Succes!
Profs. Blockeel, Davis, De Raedt

Guidelines — important!

1. This is an **closed book** *exam. You can only use the materials provided to you at the exam, and a basic scientific calculator.*
2. Read each question carefully. Just answer the question, do not provide information that is not asked (if you do, it may lower your score).
3. Answer each question in a **clear, structured** *way. Be* **concise**, **precise** *and* **to the point**. *It is not always necessary to write full sentences, bulleted lists may suffice.*
4. All questions should be answered on these pages only. Sufficient space is provided. When answer boxes are provided, write the requested answer in the box. When no box is provided, you can use all empty space on the page.
5. When a maximum length (number of words) is mentioned, do not ignore it. Too lengthy answers may lower you score. Drawings and mathematical formulas do not count as words.
6. The exam consists of 3 parts. Part 1 is connected to this page. Parts 2 and 3 are stapled together separately. Hand in everything at the same time.
7. You have 4 hours to complete the exam.

Good luck!
Profs. Blockeel, Davis, De Raedt

---

(1) Prof. Blockeel : 4 vragen, 7.5 punten • 4 questions, 7.5 points

Q 1 (1p) Leg kort uit (2-3 zinnen) wat de rol is van "pooling"-lagen in convolutionele neurale netwerken. Soms spreekt men van "max-pooling" lagen; wat denk je dat daarmee bedoeld wordt? • *Explain briefly (2-3 sentences) the role of pooling layers in convolutional neural networks. When people talk about a "max-pooling layer", what do you think they mean?*

① Reduce the computation cost

② Generalization

Q 2 (2p) Beschouw de volgende dataset. • *Consider the following dataset.*

age	**astigmatism**	**TPR**	**lens**
young	no	normal	soft
young	yes	normal	hard
young	no	reduced	none
young	no	normal	soft
young	yes	normal	hard
pre-presbyopic	no	normal	soft
pre-presbyopic	no	reduced	none
pre-presbyopic	yes	normal	none
presbyopic	no	reduced	none
presbyopic	yes	reduced	none
presbyopic	no	normal	soft
presbyopic	yes	reduced	none

Stel dat een beslissingsboom geleerd wordt met de bedoeling de waarde van **lens** te voorspellen uit de andere variabelen. Leg in het grote kader uit hoe een leersysteem als C4.5 beslist welke test in de wortel van de boom komt. Toon de concrete berekeningen (gebruik informatiewinst als heuristiek). Beantwoord ook de vragen in onderstaande tabel. • *Assume a decision tree is learned to predict the value of* **lens** *from other values. Explain (in the large box below) how a tree learner like C4.5 will determine the test in the top node of the tree. Include concrete calculations (use* information gain *as a heuristic). Then answer the following questions.*

Klasse-entropie in de hele dataset • *Class entropy in the entire dataset*:	
Informatiewinst van de geselecteerde test • *Information gain of the chosen test*:	
Stel dat de uiteindelijk geleerde boom maar 1 niveau bevat (d.w.z. onder de wortel zijn er enkel bladeren), welke klasse zal dan voorspeld worden voor het tupel (presbyopic, yes, normal)? • *Assuming the final tree consists of only one level (so there are only leaves below the top node), what class will be predicted for the unlabeled instance (presbyopic, yes, normal)?*	

Berekeningen (wordt vervolgd op volgende bladzijde) • *Calculations (continued on next page)*:

Q 3 (2p) Als je de minst algemene generalisatie onder θ-subsumptie van de volgende clausules berekent, met behulp van het algoritme dat we gezien hebben, hoeveel literals zal het niet-gereduceerde resultaat dan bevatten? • *When computing the least general generalization under θ-subsumption of the following two clauses, using the algorithm we have seen, how many literals will occur in the (non-reduced) clause that is obtained?*

friends(ann,bob) ← likes(ann, swimming), likes(ann,dining), likes(bob,jogging), likes(bob,dining).

friends(ann,carla) ← likes(ann, swimming), likes(ann,dining), likes(carla,swimming), likes(carla, jogging).

Aantal literals in de niet-gereduceerde clausule: • *Number of literals in the non-reduced clause*: [ ]

Na reductie van de clausule wordt een van de volgende clausules bekomen. Dewelke? • *After reducing that clause, one of the following clauses is obtained; which one is it?*

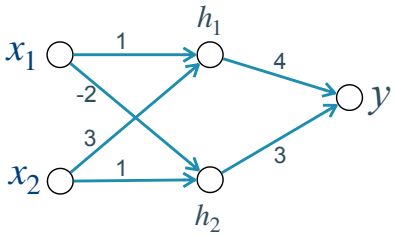
(1) friends(ann,X) ← likes(ann,dining), likes(ann,swimming), likes(X,jogging)
(2) friends(ann,X) ← likes(ann,dining), likes(ann,swimming), likes(X,swimming)
(3) friends(ann,X) ← likes(ann,dining), likes(ann,swimming), likes(X,Y), likes(ann,Y)
(4) friends(ann,X) ← likes(ann,dining), likes(ann,swimming), likes(X,jogging), likes(X,Y), likes(ann,Y)
(5) friends(ann,X) ← likes(ann,dining), likes(ann,swimming), likes(X,jogging), likes(X ,Y)

De gereduceerde clausule is clausule nummer • *The reduced clause is clause number* [ ]

Bondige argumentatie • *Brief argumentation* :

Q 4 (2.5p) Beschouw het neurale network en de mini-dataset hieronder. De activatiefuncties zijn ReLU voor de verborgen knopen, en de identiteitsfunctie voor de uitvoerknoop. Stel dat we dit netwerk trainen met behulp van gradient descent, op basis van kwadratisch verlies op de voorziene mini-dataset. Bereken de gradiënt voor de getoonde netwerkconfiguratie, gebruik makend van backpropagation. Schrijf de gradiënt op in de tabel hieronder. (De parameters zijn per laag opgesomd, en van boven naar beneden zoals ze in de figuur voorkomen; voor de duidelijkheid zijn hun waarden zoals ze in de figuur staan in de tabel getoond.) Toon in het grote kader hoe deze gradiënt berekend wordt door het backpropagation-algoritme. (Het kader gaat verder op de volgende pagina.) • *Consider the neural network and the mini-dataset shown below. The hidden nodes use a RELU activation function, the output node uses the identity function as activation function. Assume we train the network using gradient descent, and using quadratic loss on the provided mini-dataset. Compute the gradient of the loss for the shown network configuration, using backpropagation. Write the gradient in the answer table below. (The parameters are listed per layer, from top to bottom as they appear in the picture; for clarity, their values as shown in the picture are repeated in the table.) Show in the big answer box how this gradient is computed by the backpropagation algorithm. (The box extends onto the next page.)*

Network:



Data:

x_1	x_2	y
1	**1**	**5**
2	**0**	**3**

Antwoord • Answer:

Parameters	1	−2	3	1	4	3
Gradient						

Uitleg • Explanation:

Vervolg • Continuation :

*

(2) Prof. Davis: 3 vragen, 4.5 punten • 3 questions, 4.5 points

Q 5 (1.5 pts) NEDERLANDS *Je gebruikt voor een binair classificatieprobleem met* $y \in \{0, 1\}$ *het volgende model:*

$$F(x) = \frac{1}{1 + e^{w_0 + \sum_{j=1}^{d} w_j \times x_j}}$$

met x *een feature vector. Je gebruikt de volgende lossfunctie:*

$$\ell^{log}(y, F(x)) = (1 - y) \ln(1 - F(x)) + \ y \ln(F(x))$$

Je hebt een dataset met n *voorbeelden. Het doel is om een accurate maar toch spaarse gewichtsvector te leren voor dit model. Formuleer een optimalisatieprobleem waarmee dit doel bereikt wordt. Schrijf je antwoord in het onderstaande kader.*

ENGLISH *For a binary classification problem with* $y \in \{0, 1\}$*, you select the following model:*

$$F(x) = \frac{1}{1 + e^{w_0 + \sum_{j=1}^{d} w_j \times x_j}}$$

where x *is a feature vector. You pick the loss function:*

$$\ell^{log}(y, F(x)) = (1 - y) \ln(1 - F(x)) + \ y \ln(F(x))$$

Given a data set with n *examples, your goal is to learn an accurate yet sparse weight vector for this model. Formulate an optimization problem that achieves this goal. Write your answer in the box below.*

Table 1: Codewoorden voor vraag 9. Code words for Question 9.

Class	Code Word
0	011110
1	011100
2	010110
3	000100
4	001010
5	001001
6	110010
7	110001
8	111111
9	101101
10	100011
11	100001

Q 6 (1 pts) NEDERLANDS *Irrelevante features veroorzaken grote problemen bij kNN. Bij beslissingsbomen zijn irrelevante features een minder groot probleem. Leg in minder dan 50 woorden uit waarom dit zo is.*

ENGLISH *Irrelevant features cause huge problems for kNN. However, irrelevant features are less problematic for decision tree learners. Please explain why this is the case in less than 50 words.*

Q 7

(2 pts) NEDERLANDS *Voor een multiclass probleem met 12 klasses train je een Error-Correcting Output Codes ensemble waarbij je gebruikmaakt van de codewoorden voor in Tabel 1. De classifiers in het ensemble geven de volgende voorspellingen voor een testvoorbeeld x*:*

1. *positief*
2. *positief*
3. *positief*
4. *negatief*
5. *positief*
6. *negatief*

Welke klasse zal worden voorspeld voor x? Waarom wordt deze voorspelling gemaakt?*

ENGLISH *For a multiclass problem with 12 classes, you train an Error-Correcting Output Codes (ECOC) ensemble using the code words given in Table 1 for each class. For test example x^*, the classifiers in the ensemble make the following predictions:*

1. *positive*
2. *positive*
3. *positive*
4. *negative*
5. *positive*
6. *negative*

What class will the model predict for x^* and why will it make this prediction?

Voorspelling / Prediction:

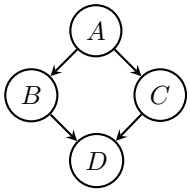
Beschrijf kort waarom deze voorspelling wordt gemaakt:
Briefly explain why this prediction is made:

*

(3) Prof. De Raedt : 5 vragen, 6 punten • 5 questions, 6 points

Q 8 (1pt) Beschouw het volgende Bayesiaans netwerk en veronderstel dat de waarschijnlijkheidsverdeling aan de volgende verzameling van onafhankelijkheden voldoet. • *Consider the following Bayesian network and assume that the probability distribution P is satisfying the following set of independencies:*

$$\{B \perp\!\!\!\perp C|A; \quad A \perp\!\!\!\perp D|B, C\}$$



1. Is de grafe een D-map voor de verdeling ? • *Is the graph a D-map for the distribution ?*

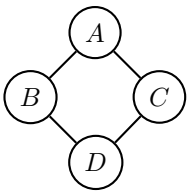
   Omcirkel JA / NEE en geef de reden.• *Circle YES / NO and justify.*

2. Is de grafe een I-map voor de verdeling ? • *Is the graph a I-map for the distribution ?*

   Omcirkel JA / NEE en geef de reden.• *Circle YES / NO and justify.*

Beschouw nu het volgende Markov model en veronderstel dat de waarschijnlijkheidsverdeling aan de volgende verzameling van onafhankelijkheden voldoet. • *Consider now the following Markov model and assume that the probability distribution P is satisfying the following set of independencies:*

$$\{B \perp\!\!\!\perp C|A, D; \quad A \perp\!\!\!\perp D|B, C\}$$



3 Is de grafe een D-map voor de verdeling ? • *Is the graph a D-map for the distribution ?*

   Omcirkel JA / NEE en geef de reden.• *Circle YES / NO and justify.*

4 Is de grafe een I-map voor de verdeling ? • *Is the graph a I-map for the distribution ?*

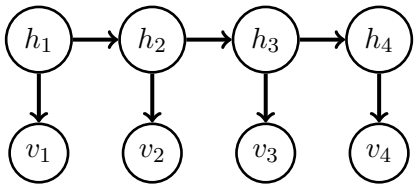
   Omcirkel JA / NEE en geef de reden.• *Circle YES / NO and justify.*

Q 9 (1 pt) Welke eigenschappen maken de Beta-verdeling zo nuttig en zo populair bij het Bayesiaanse leren? Leg elke eigenschap uit in maximaal 2 zinnen. •

What are the properties that make the Beta-distribution so useful and popular for Bayesian learning? Explain each property using at most 2 sentences.

Verklaring • Explanation:

Q 10 (1,5 pt) Beschouw het volgende HMM en neem aan dat je de marginale verdeling $P(H_2)$ wil berekenen met het message passing algoritme. • Consider the following HMM and assume you want to compute the marginal of $P(H_2)$ with message passing.



1. Teken de factor grafe die je hiervoor nodig hebt. • *Draw the factor graph that is needed to do this.*

2. Duid op de grafe aan welke boodschappen je nodig hebt om $P(H_2)$ te berekenen en ook de volgorde waarin die moeten berekend worden. Gebruik een nummer om de volgorde en een pijltje om de richting aan te geven. Er wordt niet gevraagd om de boodschappen zelf te definiëren. • *Mark the messages needed to compute $P(H_2)$ on the graph and also the order in which they have to be computed. Use numbers to indicate the order and arrows to indicate the direction. There is no need to define the messages themselves.*

3. Geef aan hoe je $P(H_2)$ kan berekenen a.d.h.v. de finale boodschappen. • *State how $P(H_2)$ to compute in terms of the final messages.*

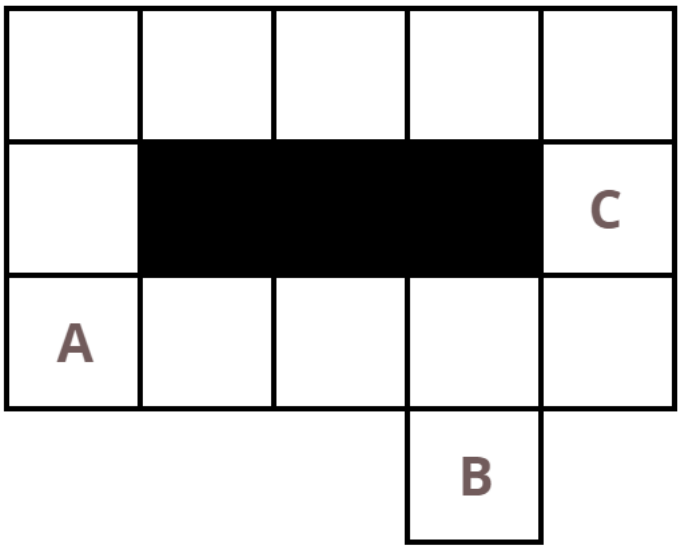
Q 11 (1,5 pt) Beschouw de grid hieronder. De agent start in positie A en neemt de volgende sequenties van acties in twee verschillende episodes. • *Consider the grid below. An agent starts in position A and takes the following sequence of actions in two different epochs:*

1. $\uparrow \uparrow \rightarrow \rightarrow \rightarrow \rightarrow \downarrow$
2. $\rightarrow \rightarrow \rightarrow \downarrow$

De acties zijn deterministisch, er is een discountfactor γ en de beloningen voor elke toestand zijn als volgt (met $r > 0$) • *The outcomes are deterministic, there is a discountfactor γ and the rewards for each state are as follows (with $r > 0$)*

$$R(s) = \begin{cases} -r & \text{if } s = B \\ r & \text{if } s = C \\ 0 & \text{otherwise} \end{cases}$$

1. Neem aan dat we Q-learning gebruiken en dat alle waarden geïnitialiseerd werden als 0. Duid op de figuur alle waarden aan die aangepast werden en duidt ook hun nieuwe waarde aan. Neem aan dat de updates achterwaarts gebeuren na elke episode. • *Assume we are performing Q-learning, and that all values are initialised as 0. Mark on the figure all Q-values that changed as well as their value, assuming that the updates are done backwards after every epoch.*



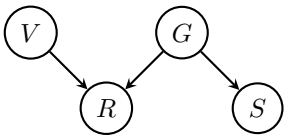
2. Is de resulterende policy optimaal ? Omcirkel JA of NEE en leg uit. • *Is the resulting policy optimal ? Circle YES or NO and explain.*

Q 12 (1pt) Beschouw het volgende Bayesiaanse netwerk met Booleaanse toevalsvariabelen en bijhorende conditionele waarschijnlijkheidstabellen (CPTs). • *Consider the following Bayesian network with boolean random variables and accompanying conditional probability tables (CPTs).*

$$p(V=0)=\delta; \quad p(G=0)=\alpha$$

$p(S=0\|G)$	G
γ	0
β	1

$p(R=0\|V,G)$	V	G
0.4	0	0
0.7	0	1
0.8	1	0
0.9	1	1

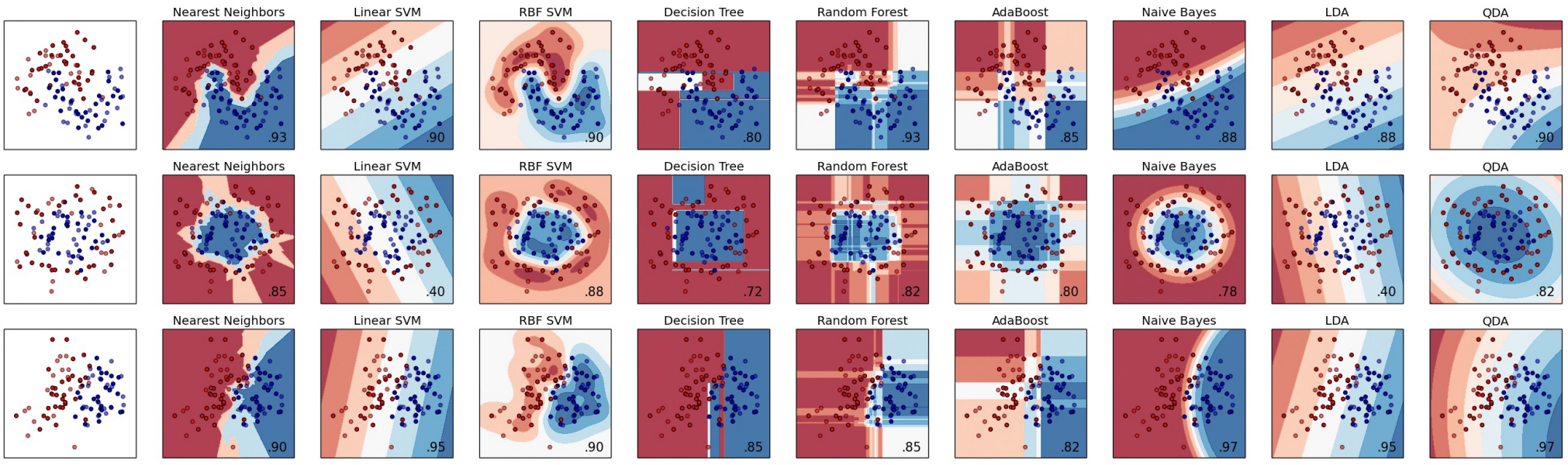


We schatten de parameters γ, β in de CPT voor S door gebruik te maken van het EM algoritme. Geef aan hoe de waardes van γ, β veranderen na een iteratie door middel van een symbolische updatevergelijking. (We zijn niet geïnteresseerd in de andere variabelen). (Hint: Denk goed na voor je rekent). • *We are estiming the parameters γ, β in the CPT for S using the EM algorithm. State how the values of γ, β change after one iteration. Use a symbolic update equation.(We are not interested in the other parameters). (Hint: Think before you compute).*

V	G	R	S
1	1	1	1
1	1	0	1
1	0	0	0
1	1	?	?

Update γ

Update β


Nearest Neighbors
Linear SVM
RBF SVM
Decision Tree
Random Forest
AdaBoost
Naive Bayes
LDA
QDA
.93
.90
.90
.80
.93
.85
.88
.88
.90
Nearest Neighbors
Linear SVM
RBF SVM
Decision Tree
Random Forest
AdaBoost
Naive Bayes
LDA
QDA
.85
.40
.88
.72
.82
.80
.78
.40
.82
Nearest Neighbors
Linear SVM
RBF SVM
Decision Tree
Random Forest
AdaBoost
Naive Bayes
LDA
QDA
.90
.95
.90
.85
.85
.82
.97
.95
.97