CS781 Mid-semester Exam (Autumn 2025)

Max marks: 30 Duration: 120 mins Roll No.

- *You are required to answer each question only in the space provided with each question.*
- *Only material written within the allotted answering space for each question will be graded.*
- *The spaces allotted for answering questions should give a rough indication of the relative lengths of correct answers to the questions.*
- *Please attach all your rough sheets.*
- *The exam is open book and notes. However, you are not allowed to search on the internet or consult others over the internet for your answers.*
- *Be brief, complete and stick to what has been asked.*
- *Unless asked for explicitly, you may cite results/proofs covered in class without reproducing them.*
- ***If you need to make any assumptions, state them clearly.***
- ***Do not copy solutions from others. Penalty for offenders: FR grade.***

1. *[10 marks]* Consider the boosted tree classifier for granting a loan to an applicant, shown below. The decision of whether a loan will be granted on not depends on whether the "scores" obtained from each of the trees add up to a value strictly greater than 0.

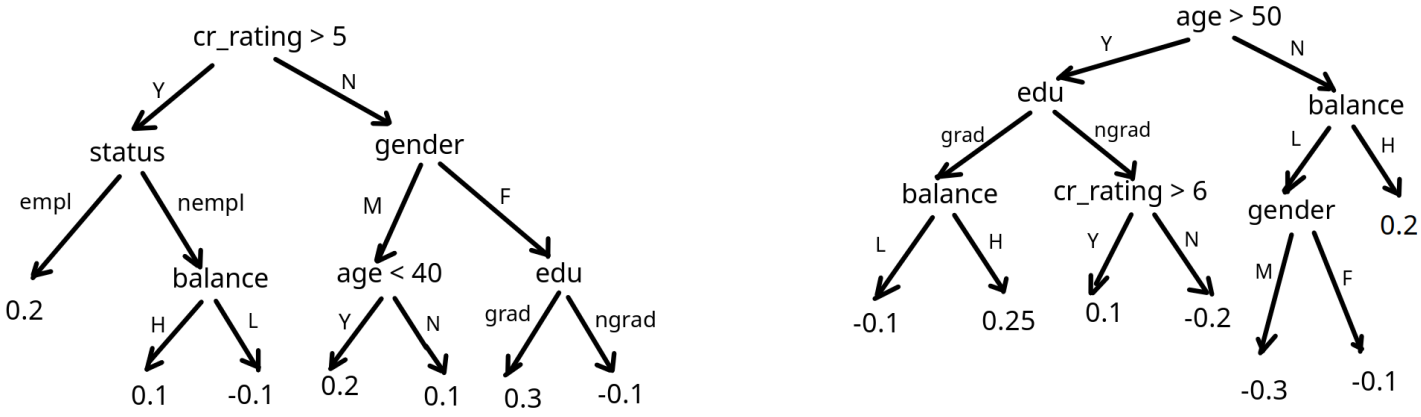

Figure 1: Neural network with ReLU activation

Consider a loan applicant with the feature vector: age = 55, gender = "F", edu = "ngrad", status = "nempl", cr_rating = "5.5", balance = "H".

(a) *[1 mark]* Determine whether the boosted tree classifier grants a loan to the applicant. State your reasons.

Score from left tree = 0.1

Score from right tree = −0.2

∴ Total score = −0.2 + 0.1 = −0.1

∴ Loan is not granted

(b) *[3 marks]* Find a subset-minimal abductive explanation for the classifier's decision.

There are many possible answers. Below is one of them.

Step 1: Is empty set an AXP?

No: consider all feature values same as for the loan applicant, except cr-rating = 6.5

This gives score from left tree = 0.1

& from right tree = 0.1

∴ Total score = 0.2; hence loan granted.

Step 2: Difference feature set: {cr_rating}

Step 3: Minimal Hitting set = {cr_rating}

Step 4: Is {cr-rating} an AXP?

No: Consider all feature values same as given, except edu = grad.

This gives left score = 0.1, right score = 0.25

∴ Total score = 0.1 + 0.25 = 0.35; loan granted.

Step 5: Difference feature set = {edu}

Step 6: Minimal HS for {cr_rating}, {edu} = {cr_rating, edu}

Step 7: Is {cr-rating, edu} an AXP?

No: Consider all feature values same as given, except age = 40

This gives left score = 0.1, right score = 0.2

∴ Total score = 0.1 + 0.2 = 0.3; loan granted.

Step 8: Difference feature set = {age}

Step 9: Minimal HS for {cr-rtg}, {edu}, {age} = {cr-rtg, edu, age}

Step 10: Is {cr-rtg, edu, age} an AXP?

Yes it is! Left score ≤ 0.2, Right score = −0.2

∴ Total score ≤ 0; loan not granted.

(c) *[3 marks]* Find a subset-minimal contrastive explanation for the classifier's decision.

Any CxP must be a hitting set for the set of all subset-minimal AxPs.

∴ Any CxP must have at least one of {cr-rtg, edu, age}

From our argument for part (b), we already see three CxPs: {cr-rtg}, {edu}, {age}

[We can flip the decision by changing the value of any one of {cr-rtg, edu, age}]

Since each of these are already subset-minimal (we can't flip the decision without changing any feature value), each is a subset-minimal CxP.

Choose {cr-rating} as a subset-minimal CxP

(d) *[3 marks]* Suppose we now add the following two domain rules

- (age > 50)$\wedge$ (balance = "H") $\rightarrow$ (cr_rating > 5)
- (gender = "F") $\wedge$ (status = "nempl") $\rightarrow$ (edu = "ngrad")

Find a domain-rule aware subset-minimal contrastive explanation for the classifier's decision. Give justification for your answer.

The CXP: {cr_rating} is already a domain-rule aware subset-minimal CXP.

WHY? If we keep all feature values same as in the given loan application, except cr_rating = 6.5 (see Step 1 in soln. of (b)), we satisfy all domain rules and can flip the decision.

Similarly, CXP: {age} is already a domain-rule aware subset-minimal CXP.

However, CXP: {edu} is not a domain-rule aware subset-minimal CXP, since if the values of features "gender" and "status" are unchanged, the second domain rule forbids changing the value of feature "edu".

2. *[10 marks]* Recall the paper on synthesizing Pareto-optimal interpretations for black-box ML models. In that paper, we used Boolean variables $m_{i,s}$ to denote whether the decision diagram rooted at the i^{th} node of the symbolic interpretation predicts the right classification for sample s. We also constructed a formula $\phi_{\mathcal{S}}$ that uniquely defined $m_{i,s}$ for each sample s, and each node i in the symbolic interpretation. This works fine when the ML model generates exactly one of two output labels (binary classification) for each input.

   Suppose we want to use a similar idea in a situation where the ML model generates one of 4 possible output labels, say $0, 1, 2, 3$, for each input. Assume that these outputs are encoded using two Boolean variables, i.e. 0 is encoded as 00, 1 and 01 and so on. Each sample s is therefore represented as $(s.input, s.out_0, s.out_1)$, where $s.out_0$ (respectively, $s.out_1$) donotes the least (respectively, most) significant bit of the output's binary encoding.

   The decision diagram (interpretation) to be synthesized in this case must also yield two Boolean outputs for each input. Accordingly, we will use two boolean variables $m_{i,s,0}$ and $m_{i,s,1}$ for each node i in the (symbolic) interpretation, and for each sample s. Specifically, we want $m_{i,s,1}m_{i,s,0}$ to be the binary encoding (MSB $= m_{i,s,1}$) of the absolute value of the difference between the ML model's output and the output of the decision diagram rooted at the i^{th} node of the symbolic interpretation, for every sample s.

   (a) *[6 marks]* Taking cue from the way $m_{i,s}$ is defined for binary classifiers, give a formula $\phi_{\mathcal{S}}$ that defines $m_{i,s,0}$ and $m_{i,s,1}$ for each i and s. You can give an inductive definition, like that given in the paper. However, your formula $\phi_{\mathcal{S}}$ must allow us to unambiguously obtain $m_{i,s,0}$ and $m_{i,s,1}$ for each i and for each sample s. You must provide justification for your steps.

Each leaf of the (symbolic) decision diagram must give two boolean values encoding the prediction of the decision diagram. Thus there must be 4 leaves, say $L_{00}, L_{01}, L_{10}, L_{11}$, where leaf $L_{\alpha\beta}$ predicts α for the MSB of the output class' encoding, and β for the LSB of same.

$\therefore$ If i is leaf $L_{\alpha\beta}$, then

$$m_{i,s,0} \leftrightarrow \begin{cases} 0 & \text{if } \beta = s.out_0 \\ 1 & \text{o/w} \end{cases}$$

if LSB of prediction matches LSB of sample's label, abs. value of diff is even, i.e., LSB of abs. diff. $= 0$

$$m_{i,s,1} \leftrightarrow \begin{cases} 1 & \text{if } |\alpha\beta - s.out_1\, s.out_0| \geq 2 \\ 0 & \text{o/w} \end{cases}$$

if the abs. diff. ≥ 2, the MSB of the diff. $= 1$ (abs. diff. can only be 0, 1, 2, or 3)

for non-leaf node i, we have formulas for $m_{i,s,0}$ and $m_{i,s,1}$ that mimic the formula for $m_{i,s}$ given in the paper.

Overall, we have ϕ_S:

$$\bigwedge_{s \in S} \bigwedge_{L_{\alpha\beta} \in L} \Big(m_{L_{\alpha\beta}, s, 0} \leftrightarrow (\beta \leftrightarrow \neg s.\text{out}_0) \;\wedge\; m_{L_{\alpha\beta}, s, 1} \leftrightarrow (|\alpha\beta - s.\text{out}_1, s.\text{out}_0| \geq 2) \Big)$$

Boolean formula with set of sat. assignments $\alpha\beta\, s.\text{out}_1, s.\text{out}_0 = \{0010, 0011, 1000, 1100, 0111, 1101\}$

$$\wedge$$

$$\bigwedge_{s \in S} \bigwedge_{1 \leq i \leq K} \bigwedge_{0 \leq t \leq 1} m_{i,s,t} \leftrightarrow \bigvee_{p \in P} \bigvee_{c \in B(p)} \Big(\lambda_{i,p} \wedge \text{func}(s,p,c) \wedge \bigwedge_{j \in \{i+1, \ldots K\} \cup L} (T_{i,c,j} \rightarrow m_{j,s,t}) \Big)$$

Evaluates to 1 iff p evaluated on s.input gives ctn outcome of p

(b) *[4 marks]* For the above black-box ML model (generating outputs $0, 1, 2, 3$), we wish to find Pareto-optimal interpretations, where the accuracy measure is defined as the sum of absolute values of differences between the ML model's output and the output of the decision diagram (interpretation). The explainability measure is assumed to be the same as that used in the paper studied in class. We will use the same $\phi_{\mathcal{E}}$ and $\phi_{\Delta_{\mathcal{E}}}$ as used in the paper. Assuming that you have a correct $\phi_{\mathcal{S}}$ from the first part of this question, indicate (a) what $\phi_{\Delta\mathcal{C}}$ you must use, and (b) what weights you must assign to which literals, such that solving the corresponding weighted MaxSAT problem generates a Pareto-optimal interpretation using the new correctness measure. Provide justification for your steps.

We want $\phi_{\Delta C}$ to be a conjunction of clauses s.t. satisfying as many of these clauses as possible maximizes the sum of abs. value of differences.

$$\therefore \phi_{\Delta C} = \bigwedge_{s \in S} m_{1,s,1} \wedge m_{1,s,0}$$

with weight 1 assigned to all $m_{1,s,0}$ and weight 2 assigned to all $m_{1,s,1}$.

This ensures that $m_{1,s,1} = \alpha$ and $m_{1,s,0} = \beta$, where $\alpha, \beta \in \{0, 1\}$ contributes $(2 \cdot \alpha + \beta)$ to the sum of abs. values of differences.

[Note that in real life, we may want to minimize the sum of absolute values of differences. In this case, we would have

$$\phi_{\Delta C} = \bigwedge_{s \in S} \neg m_{i,s,1} \wedge \neg m_{i,s,0}$$ with weight 2 for $\neg m_{i,s,1}$ and weight 1 for $\neg m_{i,s,0}$]

3. *[10 marks]* In this question, we wish to use the principles of α-β CROWN, but instead of optimizing for α and β, we are going to always use $\alpha = 0.5$ and $\beta = 100$. Consider the simple neural network shown below, where ReLUs are shown shaded. Suppose each of the inputs take values in $[-1, 3]$ (both ends inclusive).

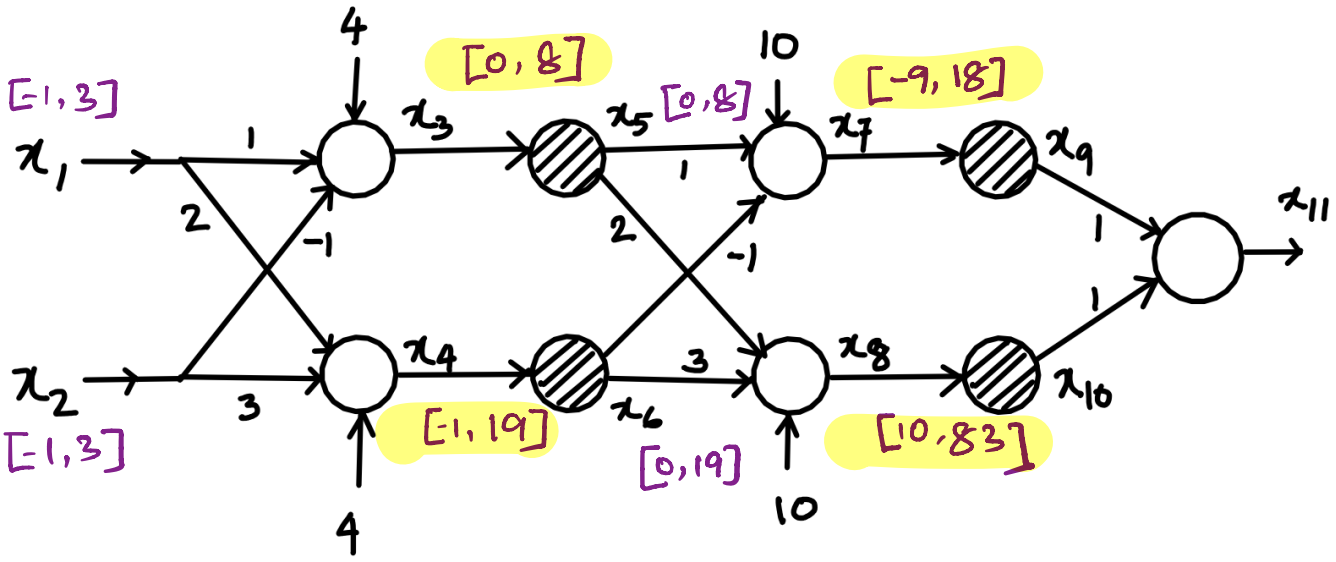

Figure 2: Neural network with ReLU activation

(a) *[2 marks]* Use interval analysis (only, nothing else) to find bounds on the inputs of each ReLU. You must clearly show the lower and upper bounds for each ReLU.

The relevant intervals are shown highlighted above:

$x_3 \in [0, 8]$

$x_4 \in [-1, 19]$

$x_7 \in [-9, 18]$

$x_8 \in [10, 83]$

(b) *[5 marks]* For each unstable ReLU, we wish to case split on when the input to the ReLU is positive and when it is non-positive. For ~~each such~~ any one such case-split, use the provided α and β values to obtain linear expressions for upper and lower bounds of the output in terms of the inputs.

Depending on which case-split you choose, there are multiple possible answers. Only one of them is shown below.

Let's propagate constraints backward from x_{11}.

$x_9 + x_{10} \leq x_{11} \leq x_9 + x_{10}$

Using $x_{10} = x_8$ and $\frac{18}{27} x_7 + \frac{9}{27} \times 18 \geq x_9 \geq \frac{1}{2} x_7$ (approx. of $x_9 = \text{ReLU}(x_7)$; $\frac{1}{2} x_7$ ← given α) we get

$\frac{1}{2} x_7 + x_8 \leq x_{11} \leq \frac{2}{3} x_7 + x_8 + 6$

Suppose we now choose the split-constraint $x_7 \geq 0$ for the unstable ReLU with x_7 as input.

This adds the constr $\{x_7 \geq 0\}$ and we must introduce a Lagrange multiplier accordingly.

When considering the upper bound. $x_{11} \leq \frac{2}{3} x_7 + x_8 + 6$, we wish to maximize.

So, we use $x_{11} \leq \frac{2}{3} x_7 + 100 \cdot x_7 + x_8 + 6$ (100 ← given β)

When considering $\frac{1}{2} x_7 + x_8 \leq x_{11}$, we wish to minimize the lower bound.

So, we use $\frac{1}{2} x_7 + x_8 - 100 x_7 \leq x_{11}$ (−100 ← given β)

This gives $-\frac{199}{2} x_7 + x_8 \leq x_{11} \leq \frac{302}{3} x_7 + x_8 + 6$

Propagating backward further, we get.

$-\frac{199}{2}(x_5 - x_6 + 10) + 2x_5 + 3x_6 + 10 \leq x_{11} \leq \frac{302}{3}(x_5 - x_6 + 10) + 2x_5 + 3x_6 + 10 + 6$

which is:

$-\frac{195}{2} x_5 + \frac{205}{2} x_6 - \frac{1970}{2} \leq x_{11} \leq \frac{308}{3} x_5 - \frac{293}{3} x_6 + \frac{3068}{3}$

[continued on next page]

approx of $x_6 = \text{ReLU}(x_4)$

Using $x_5 = x_3$ and $x_6 \le \frac{19}{20}x_4 + \frac{19}{20}$; $x_6 \ge \frac{1}{2}x_4$ (given α), we get:

$$-\frac{195}{2}x_3 + \frac{205}{8}x_4 - \frac{1970}{2} \le x_{11} \le \frac{308}{3}x_3 - \frac{293}{6}x_4 + \frac{3068}{3}$$

Suppose we now choose the split-constraint $x_4 < 0$. Using same reasoning as earlier, we get:

$$-\frac{195}{2}x_3 + \frac{205}{8}x_4 + 100x_4 - \frac{1970}{2} \le x_{11} \le \frac{308}{3}x_3 - \frac{293}{6}x_4 - 100x_4 + \frac{3068}{3}$$

(+100 x_4: given β; −100 x_4: given β)

Simplifying, we get

$$-\frac{195}{2}x_3 + \frac{1005}{8}x_4 - \frac{1970}{2} \le x_{11} \le \frac{308}{3}x_3 - \frac{893}{6}x_4 + \frac{3068}{3}$$

Propagating backward one last time, we get

$$-\frac{195}{2}(x_1 - x_2 + 4) + \frac{1005}{8}(2x_1 + 3x_2 + 4) - \frac{1970}{2} \le x_{11} \le$$

$$\frac{308}{3}(x_1 - x_2 + 4) - \frac{893}{6}(2x_1 + 3x_2 + 4) + \frac{3068}{3}$$

Simplifying, we get:

$$\frac{615}{4}x_1 + \frac{3795}{8}x_2 + \frac{225}{2} \le x_{11} \le -\frac{585}{3}x_1 - \frac{3295}{6}x_2 + \frac{2514}{3}$$

Note: The final answer precludes certain values of x_1, x_2, although they lie within $[-1, 3]$.

For example, with $x_1 = 1 = x_2$, we get

$$740.625 \le x_{11} \le 93.833 \quad \text{---- an impossibility!}$$

This is because with $x_1 = 1 = x_2$, we cannot satisfy the two split constraints we had assumed. Indeed, with $x_1 = 1 = x_2$, we get $x_3 = 4$, $x_4 = 9$, violating our split constraint $x_4 < 0$

This shows how β-CROWN propagates constraints on intermediate vars. to the inputs.

(c) *[3 marks]* Suppose we are told that the inputs are actually in a L_2-norm ball of radius 2 around $(x_1 = 1, x_2 = 1)$. Give as best upper and lower bounds (constants) on the value of the neural network's outputs as you can.

Since the question mentions an L_2-norm ball, we should try to use Holder's inequality.

Towards this end, we can calculate the linear expressions for upper and lower bounds of x_{11}, but without considering split constraints.

Reusing steps from part (b), we know:

$$\frac{1}{2}x_7 + x_8 \leq x_{11} \leq \frac{2}{3}x_7 + x_8 + 6$$

Propagating backwards:

$$\frac{1}{2}(x_5 - x_6 + 10) + (2x_5 + 3x_6 + 10) \leq x_{11} \leq \frac{2}{3}(x_5 - x_6 + 10) + (2x_5 + 3x_6 + 10) + 6$$

Simplifying:

$$\frac{5}{2}x_5 + \frac{5}{2}x_6 + 15 \leq x_{11} \leq \frac{8}{3}x_5 + \frac{7}{3}x_6 + \frac{68}{3}$$

Using $x_5 = x_3$ and $x_6 \leq \frac{19}{20}x_4 + \frac{19}{20}$; $x_6 \geqslant \frac{1}{2}x_4$ (given α), we get:

$$\frac{5}{2}x_3 + \frac{5}{4}x_4 + 15 \leq x_{11} \leq \frac{8}{3}x_3 + \frac{133}{60}x_4 + \frac{1493}{60}$$

One last propagation backwards gives:

$$\frac{5}{2}(x_1 - x_2 + 4) + \frac{5}{4}(2x_1 + 3x_2 + 4) + 15 \leq x_{11} \leq \frac{8}{3}(x_1 - x_2 + 4) + \frac{133}{60}(2x_1 + 3x_2 + 4) + \frac{1493}{60}$$

Simplifying:

$$5x_1 + \frac{5}{4}x_2 + 30 \leq x_{11} \leq \frac{213}{30}x_1 + \frac{239}{60}x_2 + \frac{2665}{60}$$

Using Holder's ineq (ref. helper slides used in class):

$$\max_{\vec{x} \in \mathbb{B}_2((1,1),2)} \left(\frac{213}{30}x_1 + \frac{239}{60}x_2 + \frac{2665}{60}\right) \leq \left(\frac{213}{30}\cdot 1 + \frac{239}{60}\cdot 1 + \frac{2665}{60}\right) + 2 \cdot \left\|\left(\frac{213}{30}, \frac{239}{60}\right)\right\|_2$$

($p = 2$, $q = 2$)

$\because \frac{1}{p} + \frac{1}{q} = \frac{1}{2} + \frac{1}{2} = 1$

$$\therefore \quad x_{11} \leq \frac{3330}{60} + 2 \times \frac{488.47}{60} \leq 71.79$$

Similarly,

$$\min_{\vec{x} \in B_2((1,1),2)} \left(5x_1 + \frac{5}{4}x_2 + 30 \right) =$$

$$- \max_{\vec{x} \in B_2((1,1),2)} \left(-5x_1 - \frac{5x_2}{4} - 30 \right)$$

$$\geq - \left(\left(-5 \times 1 - \frac{5}{4} \times 1 - 30 \right) + 2 \cdot \left\| \left(-5, \frac{-5}{4} \right) \right\|_2 \right)$$

$$\geq - \left(-\frac{145}{4} + 2 \times \frac{20.62}{4} \right)$$

$$\therefore \quad x_{11} \geq 25.94$$