CAS CS 640 Homework 1

Alec Hoyland (U83403624)

September 26, 2019

Part I
Responsible AI

Question 1.

Who codes matters

Code is written by people with thoughts on how to implement a certain algorithm or tool. What their experiences are directly influences what sorts of bugs or flaws we notice, and what our priorities are. Joy's facial recognition problems in college were caused by engineers neglecting to test their software on people with dark complexions.

How we code matters

Rigorous testing should include not just bug testing, but construct validation. Is your algorithm biased? In statistics, this can be rigorously evaluated, but if equity in your dataset isn't on the forefront of your mind, it's easy to omit construct validation testing if no one makes it a point to remember.

Why we code matters

The purpose, or objective, of code shapes how it affects the world. If code is written, someone is often footing the bill, someone with their own motivations. Code written for the EFF has a much different use-case than that written for police departments. We should keep the teleology of our software in mind.

Question 2.

2.1) I ducked for "Northpointe AI". Here are the first four relevant results:

(a) "Rise of the racist robots" (The Guardian)
(b) "Inspecting algorithms for bias" (MIT Tech Review)
(c) "Towards a code of ethics in artificial intelligence" (Future of Life)
(d) "Machine Bias" (ProPublica)

Let me begin by saying, *holy shit*, this is *not* a responsible use of AI. There are obvious benefits to using AI in the legal system, for the same reason that it's useful in medicine. Seeing a judge or doctor is a time-intensive endeavor, since there are not a lot of judges and doctors, and reviewing a case reviews synthesizing a lot of documentation. Handling large amounts of data and making a decision is a thing that artificial intelligence is great at. The problem is, neural networks just do a kind of nonlinear regression where the basis functions and the weights vary during training. This means that neural networks are the victims of their training datasets, and the twin curses of overfitting and dimensionality.

In the ProPublica report, they showed that not only were the Northpointe ROC curves shadier than previously reported, but that when compared race-by-score, that black defendants were actually less at-risk than white defendants for recidivism. In order to test for false positives or false negatives, they culled the dataset, and found that in all cases, black defendants were judged too harshly by the algorithm.

I believe this is likely due to a flawed dataset. Reality does not lie, but data are what they measure, not what you think they measure. If the legal system of Broward County included even an iota of racism, *e.g.* harsher sentencing, more arrests/criminal complaints, fewer economic and educational opportunities... this would likely show up in the dataset used to train the `COMPAS` algorithm. If racism shows up in the dataset, then it will show up in the trained classifier.

While I understand that automated classification will speed up the legal process – which is a good thing, it needs to be rigorously tested for bias before being deployed, since people's lives and freedoms are at risk here. Broward County and the `COMPAS` designers should be held responsible.

2.2) Deferral rates refer to cases where the algorithm does not make a decision, but passes the case to a human evaluator instead. At this point, there are two possible options. Either the case is evaluated, or it is tabled for later. In the case of a prisoner, this is the difference between being considered for parole, and being left in prison until the next possible parole date, which is tantamount to a "high risk" score by the `COMPAS` algorithm.

So long as the first case is true, it will result in a generally more fair legal system, provided that the people running the legal apparatus are

not racists themselves. The deferral rate essentially represents how much work cannot be offloaded to the algorithm, so an algorithm with a deferral rate of about 20% with respect to the total caseload, will require humans to do 20% of the work themselves. So, an algorithm with a low deferral rate is better for the workloads of the legal workers. Of course, one could design an algorithm that just flips a coin for each case, and has a deferral rate of 0%. That would be the most efficient, and save so much paperwork! It also has the advantage of not being racist at all. But of course, there are other issues to consider.

If the deferral rate is uneven between races or genders or what-have you, this basically results in a system where one group is being punished more for the same crimes, because they are serving longer sentences. This has economic and social ramifications, since the incarcerated are essentially disenfranchised in political, social, and economic spheres.

2.3) Canetti *et al.* discuss several different methods for making AI decision systems more equitable. In Example 3.1, they show that even using a group-blind post-processor, the unprivileged group is held to a higher standard in order to make the positive predictive values equal. This means that a non-deferring threshold post-processor has fundamental difficulties promoting totally equitable situations. The deferral rates are higher in the Canetti model vs. the original `COMPAS` algorithm, because Canetti *et al.* are taking the extra step of equalizing the accuracies by deferring more cases. This leads to an increased number of deferrals. While deferring only on the smaller group does result in fewer total referrals (without loss of equity), there are still more deferrals than a naive approach (*viz.* without post-processing).

2.4) The authors aren't sure why this happens. In the two-threshold method, it's clear that when the classifier is unsure (*i.e.* the classifier score is middling) that the classifier defers. I suspect this result happens because of the minimization procedure. Removing extremely high- or low-scoring cases would swing the distribution, since there are fewer cases scoring in the extrema. I would guess that the minimization algorithm is picking up on this, and this is causing the "shoe-in" cases to be deferred.

CAS CS 640 Homework 1

Alec Hoyland (U83403624)

September 26, 2019

Part II

Rule-based Systems

Question 3.

Use a forward-chaining algorithm to determine what the book is named. The set of assertions in the working memory are:

A1: Book X has two copies
A2: Book X is missing pages
A3: Book X has tea stains
A4: Book X is about science
A5: Book X is donated by Prof. Betke
A6: Book X is heavy

Rule matched	**With assertion(s)**	**Assertion addition(s)**
R7	A5, A6	A8: Book X is heavy
R11	A8, A1, A2, A3, A4	A9: Book X is *AI*

Question 4.

Use a backward-chaining algorithm to prove that the book is *Artificial Intelligence* using a depth-first search order.

The set of assertions in the working memory are:

BCA1: Book X is *Artificial Intelligence*.

Rule matched	**With assertion(s)**	**Assertion addition(s)**
R11	BCA1	BCA2: Book X is textbook
		BCA3: Book X has two copies
		BCA4: Book X is missing pages
		BCA5: Book X has tea stains
		BCA6: Book X is about science

Now that we have exhausted the only rule where the consequence is "is *Artificial Intelligence*", we consider rules one at a time, where the consequence is a "backward-chaining" assertion.

Rule matched	**With assertion(s)**	**Assertion addition(s)**
R6	BCA2	BCA7: Book X is donated by Prof. Betke
		BCA8: Book X has notes

We can't prove BCA8, so we backtrack.

Rule matched	**With assertion(s)**	**Assertion addition(s)**
R7	BCA2	BCA7: Book X is donated by Prof. Betke
		BCA9: Book X is heavy

We now have confirmed all of our working memory assertions via back-chaining.

Question 5.

A1: Book X has two copies
A2: Book X is missing pages
A3: Book X has tea stains
A4: Book X is about science
A5: Book X is donated by Prof. Betke
A6: Book X is heavy
A7: Book X has doodles

5.1)

Rule matched	**With assertion(s)**	**Assertion addition(s)**
R5	A5, A7	A8: Book X is fiction
R7	A5, A6	A9: Book X is textbook
R9	A8, A3, A4	Book X is *I, Robot*

This is a different answer than before, because of the precedence of the rules. I followed the rules in order, $1 - n$. When a rule was shown to be true, given the assertions, I added the consequence to the working memory assertion list, and started over at Rule 1, skipping any rules that were already proven.

5.2) (a) With backward-chaining, we begin with R11, and show that A1, A2, A3, and A4 are satisfied. We need to confirm that "A8: Book X is textbook" is true though.

(b) R6 and R7 have the consequence, "Book X is textbook". We begin with R6, and find that A5 is satisfied. We must prove now that "A9: Book X has notes" is true though.

(c) We cannot prove if Book X has notes. There is no rule for that. So we default to trying to prove R7.

(d) R7 is used when A5 and A7 are satisified. A5 and A7 are in our working memory, so we are finished.

Backward-chaining says that the book is *Artificial Intelligence*.

Question 6.

We cannot add to, or remove from the assertion list, and we cannot add to or remove any the rules we initially had. We can do the following: (a) add new rules, and (b) reorder rules.

The problem is that the rule that leads to *I, Robot* has a set of claims that are a subset of the claims that lead to *Artificial Intelligence*.

By moving the precedence of R7 up to above R5, we could solve the problem, at least in this particular instance. Then, we would conclude that the book is AI first, without reaching the conclusion that the book could be fiction.

We could also add a rule that precludes a book from being both a textbook and fiction. We would need two rules: $\text{Fiction}(X) \Rightarrow \neg\text{Textbook}(X)$ and $\text{Textbook}(X) \Rightarrow \neg\text{Fiction}(X)$. These rules would have to have high precedence, and wouldn't protect us if we were using a depth-first search, since we would conclude that the book is *I, Robot* if we didn't restart at the top of the rule list. This might make the problem undecidable though, since we would get to a contradiction, and then have to restart entirely.

CAS CS 640 Homework 1

Alec Hoyland (U83403624)

September 26, 2019

Part III

Logic and Planning

Question 7.

Convert block-world assertions into axioms in 1st-order logic without any free variables.

7.1) $\text{On}(b, a)$

7.2) $\text{On}(A, \mathit{table})$

7.3) $\text{On}(c, \mathit{table}) \vee \text{On}(c, a)$

7.4) $\text{BlueBlock}(d) \wedge \text{On}(d, \mathit{table})$

7.5) $\exists z[\text{RedBlock}(z)] \wedge \exists w[\text{BlueBlock}(w)]$

7.6) $\forall x, y[(\text{On}(x, y) \wedge \text{Equal}(y, \mathit{table})) \Rightarrow \text{On}(x, \mathit{table})]$

7.7) $\forall x, y[\text{Pyramid}(x) \Rightarrow \neg\text{On}(y, x)]$

7.8) $\forall x, y[\text{Pyramid}(x) \Rightarrow \neg\text{On}(y, x)]$

Question 8.

Prove the following theorem: $\text{CanProgram}(\mathit{Alex})$.

We have the axioms:

1. $\text{CSSenior}(\mathit{Alex})$
2. $\forall x[\text{CSSenior}(x) \Rightarrow \text{CanProgram}(x)]$

The proof of the theorem is straightforward.

We first replace implications in axiom 2.

$$\forall x[\neg\text{CSSenior}(x) \vee \text{CanProgram}(x)]$$

By specialization:

$$\neg\text{CSSenior}(Alex) \vee \text{CanProgram}(Alex)$$

We note that by axiom 1, CSSenior($Alex$), so by binary resolution,

$$\text{CanProgram}(Alex)$$

Question 9.

Transform the following expressions into clause form

9.1) Initial expression

$$\exists student : [\text{FriendsWith}(Anna, student)]$$

Skolemize to remove existential quantifiers by instantiating a Skolem constant $a \leftarrow f()$

$$\text{FriendsWith}(Anna, a)$$

9.2) Initial expression

$$\forall tire, \forall rim : [\text{RobotHas}(tire) \Rightarrow \text{RobotCanMountOn}(tire, rim)]$$

Eliminate implications

$$\forall tire, \forall rim : [\neg\text{RobotHas}(tire) \vee \text{RobotCanMountOn}(tire, rim)]$$

Eliminate universal quantifiers

$$\neg\text{RobotHas}(tire) \vee \text{RobotCanMountOn}(tire, rim)$$

9.3) Initial expression

$$\forall x : [\exists y : \text{FriendsWith}(x, y) \Rightarrow \neg\text{Misanthropist}(x)]$$

Eliminate implications

$$\forall x : [\exists y : \neg\text{FriendsWith}(x, y) \vee \neg\text{Misanthropist}(x)]$$

Skolemize by $y \leftarrow f(x)$

$$\forall x : [\neg\text{FriendsWith}(x, f(x)) \vee \neg\text{Misanthropist}(x)]$$

Eliminate universal quantifiers

$$\neg\text{FriendsWith}(x, f(x)) \vee \neg\text{Misanthropist}(x)$$

9.4) Initial expression

$$\forall x : [\text{Dog}(x) \Rightarrow [\exists y : [\text{Loves}(x, y) \wedge \neg\text{Cat}(y)]]]$$

Eliminate implications

$$\forall x : [\neg\text{Dog}(x) \vee [\exists y : [\text{Loves}(x, y) \wedge \neg\text{Cat}(y)]]]$$

Skolemize by $y \leftarrow f(x)$

$$\forall x : [\neg\text{Dog}(x) \vee [\text{Loves}(x, f(x)) \wedge \neg\text{Cat}(f(x))]]$$

Distribute disjunctions across conjunctions

$$\forall x : [[\neg\text{Dog}(x) \vee \text{Loves}(x, f(x))] \wedge [\neg\text{Dog}(x) \vee \neg\text{Cat}(f(x))]]$$

Break across conjunctions

$$\forall x : [\neg\text{Dog}(x) \vee \text{Loves}(x, f(x))]$$

$$\forall x : [\neg\text{Dog}(x) \vee \neg\text{Cat}(f(x))]$$

Eliminate universal quantifiers

$$\neg\text{Dog}(x) \vee \text{Loves}(x, f(x))$$

$$\neg\text{Dog}(x) \vee \neg\text{Cat}(f(x))$$

Question 10.

Construct a resolution proof with situation variables that serves as a plan for moving a robot through an open door using Green's trick. Let R denote the robot, D the door, S the start situation, and s_f the final situation.

Axioms:

1. $\text{Open}(D, S)$
2. $\text{Behind}(D, R, S)$
3. $\forall x, y, s[\text{Behind}(x, y, s) \wedge \text{Open}(x, s) \Rightarrow \text{InFront}(x, y, \text{Advance}(y, s))]$

Prove the theorem:

$$\exists s_f[\text{InFront}(D, R, s_f)]$$

Let's consider a proof by contradiction. Start with assuming the negation of the theorem:

$$\neg\text{InFront}(D, R, s_f)$$

and add an "answer" term (Green's trick).

$$\neg\text{InFront}(D, R, s_f) \vee \text{Answer}(s_f)$$

Let's begin: $\text{Open}(D, S)$ and $\text{Behind}(D, R, S)$ imply

$$\text{Behind}(D, R, S) \wedge \text{Open}(D, S) \Rightarrow \text{InFront}(D, R, \text{Advance}(R, S))$$

$\text{InFront}(D, R, \text{Advance}(R, S))$ implies

$$\text{InFront}(D, R, \text{Advance}(R, S)) \vee \text{Answer}(\text{Advance}(R, S))$$

Since $\neg\text{InFront}(D, R, s_f)$, by *modus ponens*,

$$\text{Answer}(\text{Advance}(R, S))$$

CAS CS 640 Homework 1

Alec Hoyland (U83403624)

September 26, 2019

Part IV

ROC Analysis

Question 11.

11.1) The true positive rate is computed as the true positives divided by the total positives, $\text{TPR} = \frac{\text{TP}}{\text{P}}$. The false negative rate is $\text{FNR} = 1 - \text{TPR}$. The false positive rate is $\text{FPR} = \frac{\text{FP}}{N}$. The true negative rate is $\text{TNR} = 1 - \text{FPR}$.

For model A, the TPR is $26/(26+30) = 0.46$. The FNR is $1 - \text{TPR} = 0.54$. The FPR is $18/(18+26) = 0.41$. The TNR is $1 - \text{FPR} = 0.59$.

For model B, the TPR is $13/(13+32) = 0.29$. The FNR is 0.71. The FPR is $38/(38+17) = 0.69$. The TNR is 0.21.

11.2) Accuracy is defined as $\text{ACC} = \frac{\text{TP}+\text{TN}}{P+N}$. For model A, the accuracy is $(26+26)/(26+30+18+26) = 0.52$. For model B, the accuracy is $(13+17)/(13+32+38+17) = 0.30$.

Precision is defined as $\text{TP}/(\text{TP}+\text{FP})$. For model A, the precision is $26/(26+18) = 0.59$. For model B, the precision is $13/(13+38) = 0.25$.

Recall is the same as the true positive rate, since it measures the fraction of relevant instances that have been retrieved.

The F-1 score is the harmonic mean of precision and sensitivity.

$$\text{F}_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} \cdot \text{TPR}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

For model A, it's $2 \cdot 26/(2 \cdot 26 + 18 + 30) = 0.52$. For model B, it's $2 \cdot 13/(2 \cdot 13 + 38 + 32) = 0.27$.

Question 12.

Model 1 is the better one. A good ROC curve is above the diagonal. This is because perfect classification means that all positives are classified as positives, (i.e. TPR of 1), and no negatives are classified as positives (i.e.

FPR of 0). This point is at $(0, 1)$ in the top-left corner. Model A likely produced ROC curve #1, since it has a higher TPR and lower FPR.

Question 13.

If you want to minimize the number of false positives, you want the threshold to be high, so that only strong quakes would trip the system. Too sensitive (low threshold), and anything could set it off.

Colophon

This document was typeset using the LaTeX 2_ε document processing system originally developed by Leslie Lamport, based on the TeX typesetting system created by Donald Knuth. The class is `latex-homework-class` by Jake Zimmerman, released under the MIT license. All above work was done by the authors.