CSE547T Class 8

Jeremy Buhler

February 13, 2017

1 DFAs Can't Count, Part Deux

I want to go back to a statement I made earlier and see if I can make it more precise.

- When we looked at the language 0^n1^n, I declared that "DFAs can't count"
- Basic idea: DFAs have finite memory, and counting requires remembering arbitrarily large integers
- Can we be a little more specific?

Think about a DFA M that would try to accept $0^n1^n\ldots$

- Somehow, the states of M need to encode how "unbalanced" current string is.
- For example, if $x = 00$, we need to see two 1's; if $x = 000$, we need to see three 1's.
- In general, for strings of the form $x = 0^m$, we need to remember that we have to see exactly m 1's to succeed.
- But there are infinitely many possible values of m, so how can we remember the right one with a (finite) DFA?

I want a way to make this kind of intuitive argument precise and use it to prove non-regularity.

2 Fancy Notation

- **Defn**: the **quotient** of a language L by x, denoted L/x, is the set of strings
$$\{z \in \Sigma^* \mid xz \in L\}.$$
- For example, if L is all even-length strings of 0's, and $x = 000$, then L/x is the set of all odd-length strings of 0's.
- You may prefer to read L/x as "the completions of x in L".
- Every language L induces a relation I_L on Σ^*, defined by
$$xI_Ly \text{ iff } L/x = L/y.$$
- Note that I_L is always an *equivalence relation*. (The equivalence property reduces to equality of sets.)
- For string $x \in \Sigma^*$, we denote the equivalence class of x under I_L by $[x]$.

3 A New Characterization of RLs

Thm (Myhill-Nerode): A language L is regular iff I_L induces a finite number of equivalence classes on Σ^*.

- ($\leftarrow$) Suppose I_L induces a finite number of equivalence classes on Σ^*.
- We will construct a DFA M that recognizes L.
- *Idea*: eqv classes of I_L correspond exactly to states of the DFA! "All strings with same quotient" = "All strings that get us to same state of DFA."
- Construct $M = (Q, \Sigma, q_0, A, \delta)$ as follows.
- For every class $[x]$ of I_L, Q contains a state $\langle [x] \rangle$.
- $q_0 = \langle [\varepsilon] \rangle$.
- $A = \{\langle [x] \rangle \mid x \in L\}$
- For $a \in \Sigma$, $\delta(\langle [x] \rangle, a) = \langle [x \cdot a] \rangle$.

Whoa, hold on! Are A and δ well-defined in our construction?

- Definitions seem to depend on our choice of class labels.
- Need to check that we get same automaton regardless of how we name the eqv classes.
- **Lemma 1**: if $[x] = [y]$, then $[xa] = [ya]$.
- **Pf**: If $[x] = [y]$, then for any $z \in \Sigma^*$, $xz \in L$ iff $yz \in L$.
- Consider strings of the form $z = aw$. Above implies that, for any $w \in \Sigma^*$, $(xa)w \in L$ iff $(ya)w \in L$.
- Conclude that $[xa] = [ya]$. QED
- **Lemma 2**: $x \in L$ iff for all $y \in [x]$, $y \in L$.
- ($\leftarrow$) is trivial because $x \in [x]$.
- ($\rightarrow$) if $y \in [x]$, then for all z, $yz \in L$ iff $xz \in L$.
- Because $x = x\varepsilon \in L$, it must be that $y\varepsilon = y \in L$. QED

Whew! OK, back to easy proof. We claim that

$$\delta^*(\langle [x] \rangle, z) = \langle [xz] \rangle .$$

- Proceed by induction on $|z|$.
- **Bas**: when $z = \varepsilon$, observe that

$$\begin{aligned} \delta^*(\langle [x] \rangle, \varepsilon) &= \langle [x] \rangle \\ &= \langle [x\varepsilon] \rangle . \end{aligned}$$

- **Ind**: suppose $z = ya$. We have that

$$\begin{aligned} \delta^*(\langle [x] \rangle, ya) &= \delta(\delta^*(\langle [x] \rangle, y), a) \\ &= \delta(\langle [xy] \rangle, a) \\ &= \langle [x \cdot ya] \rangle \end{aligned}$$

  where the second line follows by the IH.

OK, let's wrap up.

- Claim that $L(M) = L$.
- We have $z \in L(M)$ iff $\delta^*(q_0, z) \in A$.
- Now $q_0 = \langle [\varepsilon] \rangle$, so

$$\delta^*(q_0, z) = \delta^*(\langle [\varepsilon] \rangle, z) = \langle [\varepsilon z] \rangle = \langle [z] \rangle .$$

- Conclude that $z \in L(M)$ iff $\langle [z] \rangle \in A$.
- But, by our definition of A and Lemma 2, $\langle [z] \rangle \in A$ iff $z \in L$.
- Hence $z \in L(M)$ iff $z \in L$. QED

That was half the proof. Now for other half (by contrapositive). (This is the formalization of our first intuition.)

- ($\rightarrow$) Let L be a language for which I_L induces *infinitely many* equivalence classes.
- Suppose we could construct a DFA M for L.
- Define the relation I_M on pairs of strings by

$$x I_M y \text{ iff } \delta^*(q_0, x) = \delta^*(q_0, y).$$

  I_M is also an equivalence relation.

- We claim that for strings x and y, if $x I_M y$, then $x I_L y$.
- Indeed, if $x I_M y$, then for any $z \in \Sigma^*$, we have that

$$\begin{aligned} \delta^*(q_0, xz) &= \delta^*(\delta^*(q_0, x), z) \\ &= \delta^*(\delta^*(q_0, y), z) \\ &= \delta^*(q_0, yz). \end{aligned}$$

- Hence, M accepts xz iff it accepts yz, and so $xz \in L$ iff $yz \in L$. Conclude that $x I_L y$ by defn.
- We have just shown that I_M is a *refinement* of I_L. That is, each class of I_L is composed of one or more classes from I_M, and no two classes of I_L can share a class of I_M.

- Now there are infinitely many (nonempty!) distinct classes of I_L, hence infinitely many distinct classes of I_M.
- But M has only finitely many states, so I_M cannot be infinite. $\rightarrow\leftarrow$
- Conclude that no DFA recognizes L, and so L is not regular. QED

4 Application of Myhill-Nerode

You can use the Myhill-Nerode theorem to prove that a language is non-regular.

- We showed: if I_L has infinitely many eqv classes, L is not regular.
- How can we show that I_L has infinitely many eqv classes for a *given* L?
- Let x and y be two strings. If we can find a z for which $xz \in L$ but $yz \notin L$ (or vice versa), we say that x and y are *distinguishable under* L. x and y cannot be in same eqv class of I_L.
- If we can find infinitely many strings that are all pairwise distinguishable, L must have infinitely many eqv classes.
- **Example**:
$$L = \{x \in \{0,1\}^* \mid x = 0^n 1^n\}.$$
- Consider the strings $x = 0^n$ and $y = 0^m$, for $n \neq m$. Let $z = 1^n$; then $xz \in L$, but $yz \notin L$.
- Conclude that 0^n and 0^m are distinguishable for any $n \neq m$. Hence L is not regular. QED
- **Example**:
$$L = \{x \in \{0,1\}^* \mid x = yy\}.$$
- Consider the strings $x = 10^n$ and $y = 10^m$, for $n \neq m$. Let $z = 10^n$; then $xz \in L$, but $yz \notin L$.
- Conclude that 10^n and 10^m are distinguishable for any $n \neq m$. Hence L is not regular. QED

5 One More Thing...

Claim: the DFA produced for a regular language L in the proof of the Myhill-Nerode Theorem has the fewest states among all DFAs recognizing L.

- Suppose not; i.e. suppose we can find a smaller DFA M' with $L(M') = L$. Let A' be the accepting set of M'.
- Now M has as many states as there are equivalence classes for L, but M' has fewer states.

- Hence, for some state q of M', there must be strings x, y with $[x] \neq [y]$ and

$$\delta^*(q_0, x) = \delta^*(q_0, y) = q.$$

- Let z be a string that distinguishes x from y. WLOG, suppose $xz \in L$ but $yz \notin L$.
- If $xz \in L$, then in M', $\delta^*(q, z) \in A'$.
- But this implies that

$$\delta^*(\delta^*(q_0, y), z) = \delta^*(q_0, yz) \in A',$$

  which means that M' incorrectly accepts yz.
- Hence, $L(M') \neq L$, and so M' cannot exist. QED