Netflix intro

What data should we encode about each Netflix account holder to help us make effective recommendations?

In machine learning, clustering can be used to group similar data for prediction and recommendation. For example, each Netflix user's viewing history can be represented as a n-tuple indicating their preferences about movies in the database, where n is the number of movies in the database. People with similar tastes in movies can then be clustered to provide recommendations of movies for one another. Mathematically, clustering is based on a notion of distance between pairs of n-tuples.

Data types

Term	Examples:	
	(add additional	examples from class)
set	$7 \in \{43, 7, 9\}$	$2 \notin \{43, 7, 9\}$
unordered collection of elements		
repetition doesn't matter		
Equal sets agree on membership of all elements		
n-tuple		
ordered sequence of elements with n "slots" $(n > 0)$		
repetition matters, fixed length		
Equal n-tuples have corresponding components equal		
string		

ordered finite sequence of elements each from specified set (called the alphabet over which the string is defined) repetition matters, arbitrary finite length Equal strings have same length and corresponding characters equal

Special cases:

When n = 2, the 2-tuple is called an **ordered pair**.

A string of length 0 is called the **empty string** and is denoted λ .

A set with no elements is called the **empty set** and is denoted $\{\}$ or \emptyset .

Ratings encoding

In the table below, each row represents a user's ratings of movies: \checkmark (check) indicates the person liked the movie, \checkmark (x) that they didn't, and \bullet (dot) that they didn't rate it one way or another (neutral rating or didn't watch). Can encode these ratings numerically with 1 for \checkmark (check), -1 for \checkmark (x), and 0 for \bullet (dot).

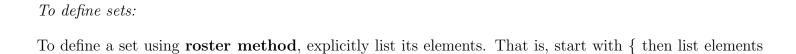
Person	Dune	Oppenheimer	Barbie	Nimona	Ratings written as a 4-tuple
P_1	X	•	✓		
D					
P_2	/	√	X		
P_3		1	1		
1 3	•	•	•		
P_4	•	×	✓		
You					

Definitions set prereqs

Term	Notation Example(s)	We say in English
all reals	\mathbb{R}	The (set of all) real numbers (numbers on the number
		line)
all integers	$\mathbb Z$	The (set of all) integers (whole numbers including neg-
		atives, zero, and positives)
all positive integers	\mathbb{Z}^+	The (set of all) strictly positive integers
all natural numbers	N	The (set of all) natural numbers. Note : we use the
		convention that 0 is a natural number.

Defining sets

of the set separated by commas and close with \{\}.



To define a set using **set builder definition**, either form "The set of all x from the universe U such that x is ..." by writing

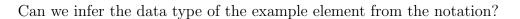
$$\{x \in U \mid ...x...\}$$

or form "the collection of all outputs of some operation when the input ranges over the universe U" by writing

$$\{...x...\mid x\in U\}$$

We use the symbol \in as "is an element of" to indicate membership in a set.

Example sets: For each of the following, identify whether it's defined using the roster method or set builder notation and give an example element.





 $\{0, 0\}$

 $\{-1,0,1\}$

$$\{(x, x, x) \mid x \in \{-1, 0, 1\}\}\$$

{}

$$\{x \in \mathbb{Z} \mid x \ge 0\}$$

$$\{x \in \mathbb{Z} \mid x > 0\}$$

 $\{\smile, \heartsuit\}$

 $\{A,C,U,G\}$

{AUG, UAG, UGA, UAA}

Definitions functions prereqs

Term	Notation Example(s)	We say in English
sequence	x_1, \ldots, x_n	A sequence x_1 to x_n
summation	x_1, \dots, x_n $\sum_{i=1}^n x_i \text{ or } \sum_{i=1}^n x_i$	The sum of the terms of the sequence x_1 to x_n
piecewise rule definition	$f(x) = \begin{cases} \text{rule 1 for } x & \text{when COND 1} \\ \text{rule 2 for } x & \text{when COND 2} \end{cases}$	Define f of x to be the result of applying rule 1 to x when condition COND 1 is true and the result of applying rule 2 to x when condition COND 2 is true. This can be generalized to having more than two conditions (or cases).
function applica-	f(7)	f of 7 or f applied to 7 or the image of 7 under f
01011	f(z)	f of z or f applied to z or the image of z under f
	f(g(z))	f of g of z or f applied to the result of g applied to z
absolute value	-3	The absolute value of -3
square root	$\sqrt{9}$	The non-negative square root of 9

Pro-tip: the meaning of two vertical lines | | depends on the data-types of what's between the lines. For example, when placed around a number, the two vertical lines represent absolute value. We've seen a single vertial line | used as part of set builder definitions to represent "such that". Again, this is (one of the many reasons) why is it very important to declare the data-type of variables before we use them.

Defining functions

New! Defining functions A function is defined by its (1) domain, (2) codomain, and (3) rule assigning each element in the domain exactly one element in the codomain.

The domain and codomain are nonempty sets.

The rule can be depicted as a table, formula, piecewise definition, or English description.

The notation is

"Let the function FUNCTION-NAME: DOMAIN \rightarrow CODOMAIN be given by FUNCTION-NAME(x) = ... for every $x \in DOMAIN$ ".

or

"Consider the function FUNCTION-NAME: DOMAIN \rightarrow CODOMAIN defined as FUNCTION-NAME(x) = ... for every $x \in DOMAIN$ ".

Example: The absolute value function

Domain

Codomain

Rule

Defining functions ratings

Recall our representation of Netflix users' ratings of movies as n-tuples, where n is the number of movies in the database. Each component of the n-tuple is -1 (didn't like the movie), 0 (neutral rating or didn't watch the movie), or 1 (liked the movie).

Consider the ratings $P_1 = (-1, 0, 1, 0), P_2 = (1, 1, -1, 0), P_3 = (1, 1, 1, 0), P_4 = (0, -1, 1, 0)$

Which of P_1 , P_2 , P_3 has movie preferences most similar to P_4 ?

One approach to answer this question: use **functions** to quantify difference among user preferences.

For example, consider the function $d_0: \{-1,0,1\}^4 \times \{-1,0,1\}^4 \to \mathbb{R}$ given by

$$d_0(\ (\ (x_1,x_2,x_3,x_4),(y_1,y_2,y_3,y_4)\)\) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2+(x_4-y_4)^2}$$

Defining functions recursively

When the domain of a function is a recursively defined set, the rule assigning images to domain elements (outputs) can also be defined recursively.

Recall: The set of RNA strands S is defined (recursively) by:

Basis Step: $A \in S, C \in S, U \in S, G \in S$

Recursive Step: If $s \in S$ and $b \in B$, then $sb \in S$

where sb is string concatenation.

Definition (Of a function, recursively) A function rnalen that computes the length of RNA strands in S is defined by:

 $rnalen: S \rightarrow \mathbb{Z}^+$

Basis Step: If $b \in B$ then rnalen(b) = 1Recursive Step: If $s \in S$ and $b \in B$, then rnalen(sb) = 1 + rnalen(s)

The domain of rnalen is

The codomain of rnalen is

Example function application:

$$rnalen(\mathtt{ACU}) =$$

Example: A function basecount that computes the number of a given base b appearing in a RNA strand s is defined recursively: