Rohan Nagar
ran679

EE 379K Lab 3 Written Questions

1) Take the columns of the characteristic matrix for sets S_1 and S_2. There are 3 possibilities for each row:
   (1) rows have a 1 in both columns
   (2) rows have a 1 in one column and a 0 in the other
   (3) rows have a 0 in both columns

Type 1 and 2 determine the similarity and the probability that $h(S_1) = h(S_2)$. Let there be x rows of type 1 and y rows of type 2. We then know that x is the size of $S_1 \cap S_2$ and $(x+y)$ is the size $S_1 \cup S_2$.

So, the Jaccard Similarity is $\frac{x}{x+y}$.

Now, if we imagine a random permutation, the probability that we will meet a type 1 row before a type 2 row is $\frac{x}{x+y}$. This is because there are x type 1 rows and $(x+y)$ total type 1 and type 2 rows. We can ignore type 3 because if it shows up first, both entries are 0, so we continue looking for the first instance of a 1.

So, since $P(\text{type 1 before type 2}) = \frac{x}{x+y}$, and if that occurs we know that $h(S_1) = h(S_2)$ because the row has both columns with a value 1. Then, $P(h(S_1) = h(S_2)) = \frac{x}{x+y}$. This is the same as the Jaccard Similarity. Thus,

$$P(h(S_1) = h(S_2)) = \frac{\|S_1 \cap S_2\|}{\|S_1 \cup S_2\|}$$