Problem 1

Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T?

Lowest non-zero JS for two sets of size m is: $JS_{min} = JS_1 = \frac{1}{2m-1}$, which occurs when only 1 element is shared between S and T. On the other hand, $JS_{max} = JS_m = \frac{m}{m} = 1$ when $S = T$. In general, $JS_k = \frac{k}{2m-k}$ for S and T, where $k = |S \cap T|$.

Expected value of Jaccard Similarity is:

$$E[JS] = \Sigma_{k=1}^{m} JS_k(S,T) \times P[JS_k(S,T)]$$

Probability that *some* element x_i from U is in S:

$$P(x_i \in S) = \frac{1}{n}$$

Probability that *some* element x_i from U is not in S:

$$P(x_i \notin S) = 1 - \frac{1}{n}$$

Probability that *some* element x_i from U is in both S and T:

$$P(x_i \in T | x_i \in S) = \frac{1}{n^2}$$

Probability that *some* element x_i from U is in T given that it is not in S:

$$P(x_i \in T | x_i \notin S) = \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2}$$

Probability that *some* k elements $\tilde{S}_k$ from U are in S (given $k << n$):

$$P(\tilde{S}_k \subset S) \simeq \left(\frac{1}{n}\right)^k$$

Probability that *some* k elements $\tilde{S}_k$ from U are in T given that they are in S (given $k << n$):

$$P(\tilde{S}_k \subset T | \tilde{S}_k \subset S) \simeq \left(\frac{1}{n}\right)^{2k}$$

Probability that *some* $m-k$ elements $\tilde{S}'_k$ from U are not in T given that they are in S (given $k << n$):

$$P(\tilde{S}'_k \not\subset T | \tilde{S}'_k \subset S) \simeq \left(\frac{n-1}{n^2}\right)^{m-k}$$

Probability that given S, *only some* k elements $\tilde{S}_k$ are also in T:

$$P(\tilde{S}_k = S \cap T,\ \tilde{S}'_k \not\subset T) \simeq \left(\frac{n-1}{n}\right)^{m-k}\left(\frac{1}{n}\right)^{2k}$$

Extending this to *any* k in U:

$$P(\tilde{S}_k = S \cap T,\ \tilde{S}'_k \not\subset T) \simeq \left(\frac{n-1}{n}\right)^{m-k}\left(\frac{1}{n}\right)^{2k} \times \binom{n}{k} = P(JS_k)$$

Thus, Expected Jaccard Similarity for sets S and T each of size m from U is:

$$E[JS] = \Sigma_{k=1}^{m}\left\{\frac{k}{2m-k} \times \left(\frac{n-1}{n}\right)^{m-k}\left(\frac{1}{n}\right)^{2k} \times \binom{n}{k}\right\} \tag{1}$$

Problem 2

Consider a similarity function s(x, y) which returns a similarity measure in the range [0, 1] between two items x and y. We say that s is LSHable if there is a family of hash functions H so that the following holds: For any set of items x, y, z, w with s(x, y) $\geq$ s(z, w), if h is chosen from H uniformly at random, then Pr[h(x) = h(y)] $\geq$ Pr[h(z) = h(w)]. In other words, a higher similarity between items implies a higher probability of being matched under a randomly chosen hash function from H.

Part 1

Prove that if the function s is LSHable then the function d(x, y) = 1 - s(x, y) satisfies the triangle inequality i.e. for any three items x, y and z, d(x, y) + d(y, z) $\geq$ d(x, z).

Let d(x,y) = 1 - s(x,y) be the distance between x and y, such that for any set of items x, y, z, w with d(x, y) $\leq$ d(z, w), if h is chosen from H uniformly at random, then Pr[h(x) = h(y)] $\geq$ Pr[h(z) = h(w)]. In other words, a higher distance between items implies a lower probability of being matched under a randomly chosen hash function from H.

Proof by Contradiction

Suppose that d is LSHable but:

$$d(x, y) + d(y, z) < d(x, z)$$

Then,

$$1 - s(x, y) + 1 - s(y, z) < 1 - s(x, z)$$

$$s(x,y) + s(y,z) - 1 > s(x,z)$$
$$s(x,y) + s(y,z) - 1 < s(x,y)$$
$$s(x,y) > s(x,z)$$
$$P[h(x) = h(y)] > P[h(x) = h(z)]$$
$$s(x,y) + s(y,z) - 1 < s(y,z)$$
$$s(y,z) > s(x,z)$$
$$P[h(y) = h(z)] > P[h(x) = h(z)]$$
$$s(y,z) + s(x,y) > 2s(x,z) \tag{2}$$

Since for each time h(x)$\neq$ h(z), either h(x) has to be $\neq$ h(y) or h(y) has to be $\neq$ h(z) (otherwise d would be not LSHable), it follows that:

$$P[h(x) \neq h(y)] + P[h(y) \neq h(z)] \geq P[h(x) \neq h(z)]$$

$$1 - P[h(x) = h(y)] + 1 - P[h(y) = h(z)] \geq 1 - P[h(x) = h(z)]$$
$$P[h(x) = h(y)] + P[h(y) = h(z)] - 1 \leq P[h(x) = h(z)] \tag{3}$$

Now from 2 it follows that

$$P[h(x) = h(y)] + P[h(y) = h(z)] \geq P[h(x) = h(z)]$$

which contradicts 3. □

Part 2

> *Consider the following similarity measure for sets called Srensens Dice Similarity coefficient : dsc(A, B) =* $\frac{2|A \cap B|}{|A|+|B|}$. *Use the above to conclude that dsc is not LSHable.*

(Since *dsc* is a similarity measure in the range [0,2] (and not [0,1] as required by the above definition of LShability), it is not LSHable, but let us still consider a counterexample for fun.)

Let $d(A,B) = 1 - dsc(A,B)$. If $d(A,B) + d(B,C) < d(A,C)$, then dsc is not LSHable.

Let $A = \Big\{1,2,3\Big\}$, $B = \Big\{1,0,3\Big\}$, $C = \Big\{1,0,4\Big\}$.

Then: d(A,B) + d(B,C) = 0 and d(A,C) = 0.6 > d(A,B) + d(B,C). □

Problem 3

Let v be any (fixed) vector in R^d . Consider a random gaussian vector $g = (g_1, ..., g_d) \in R^d$ where each component g_i of g is chosen independently from the normal distribution $N(0, 1)$ with mean 0 and variance 1. What is the distribution of the random variable $x = g \cdot v$?

For some fixed scalar A and gaussian $j \in N(0, 1)$:

$$E[Aj] = E[A].E[j] = A.E[j]$$

$$Var[A.j] = A^2.Var[j]$$

$$A.j = \tilde{j} \ where \ \tilde{j} \in N(0, x^2)$$

Then:

$$\begin{pmatrix} v_1 \\ v_2 \\ . \\ . \\ . \\ v_d \end{pmatrix} \cdot \begin{pmatrix} g_1 \\ g_2 \\ . \\ . \\ . \\ g_d \end{pmatrix} = \begin{pmatrix} g_1.v_1 \\ g_2.v_2 \\ . \\ . \\ . \\ g_d.v_d \end{pmatrix} = \begin{pmatrix} N(0, v_1^2) \\ N(0, v_2^2) \\ . \\ . \\ . \\ N(0, v_d^2) \end{pmatrix}$$