02-425/725 HW1

February 5, 2021

1 Molecular Graph Kernels

Consider the following three molecules.

- Penicillin A: http://www.chemspider.com/Chemical-Structure.16736211.html
- Penicillin G: http://www.chemspider.com/Chemical-Structure.6014.html
- Caffeine: http://www.chemspider.com/Chemical-Structure.2424.html

Using the save button download each molecular structure. ChemSpider provides you with a V2000 MOL file. This represents a molecular graph using two tables: one for atoms and one for bonds. Below is the V2000 MOL file for ethane.

```



  2  1  0  0000  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    1.3300    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
M  END

> <StdInChI>
InChI=1S/C2H6/c1-2/h1-2H3

> <StdInChIKey>
OTMSDBZUPAUEDD-UHFFFAOYSA-N

> <AuxInfo>
1/0/N:1,2/E:(1,2)/rA:2nCC/rB:s1;/rC:;1.33,0,0;

> <Formula>
C2 H6

> <Mw>
30.06904

> <SMILES>
CC

> <CSID>
6084

$$$$
```

The first three lines are intended for information about the molecule and can safely be ignored.

The next line is the counts line, which contains information about how many atoms and bonds are in the molecule. The first number is the number of atoms in the molecule and the second number is the number of bonds. The remainder of the counts line largely exists for backwards compatibility and can be ignored. In the ethane example there are 2 atoms and 1 bond.

Next there is a line for each atom in the molecule, as many as specified in the counts line. Each atom line contains (in order):

- x, y, and z coordinates for the atom (0,0,0 and 1.33,0,0)
- the atomic symbol for the atom (C and C)

The remaining fields specify things like charge and mass difference from the periodic table. For our purposes they are not necessary. Each atom is assigned an implicit 1-based index, in the order they appear in the atom table. For ethane the first carbon (xyz coordinates 0,0,0) gets assigned index 1. Note that ethane only has two atoms. This is because hydrogens are typically implicit and can be easily added after the heavy (non-hydrogen) atoms have been fully specified.

Next there is a line for each bond in the atom, as many as specified in the counts line. The relevant fields, in order, are

- The index of first atom involved in bond
- The index of second atom involved in bond
- The bond type (1 = single, 2 = double, 3 = triple)

The remaining fields can be safely ignored. For ethane the only bond that exists says that the two carbons in the molecule are connected via a single bond.

The `M END` tag tells us that the connectivity of the atom is done. The remaining lines in the file are combinations of property names and their values, which we won't make use of here.

If interested in a more rigorous specification for MOL files please see `http://c4.cabrillo.edu/404/ctfile.pdf`.

1.1 Implicit Hydrogens

1. Consider the molecular graph (with hydrogens omitted) below

$$\mathrm{C} = \mathrm{C} - \mathrm{O} - \mathrm{C}$$

   List the number of hydrogens each atom should have (from left to right).

2. Consider Penicillin A. Using the implicit atom indices for each heavy atom list the number of hydrogens attached to that atom. You may assume that sulfur (S) always has valence 2 and nitrogen (3) always has valence 3. There should be 26 hydrogens in total.

1.2 Computing Kernels

1. Write a script to compute the adjacency matrix for a molecular graph given a V2000 MOL file. Indices should correspond to the implicit indices in the MOL file (for the ethane example, the first row/column in the adjacency matrix should represent the first carbon in the MOL file). You may temporarily ignore hydrogens that are not explicit (you don't need to add implicit hydrogens yet). Show the adjacency matrix your code outputs for Caffeine.

2. Make all implicit hydrogens explicit. You may assume nitrogen and sulfur always have valences 3 and 2 respectively. Give your hydrogens indices according to the indices of the heavy atom they're connected to. For example, in ethane the hydrogens connected to carbon 1 will get indices 3, 4, and 5. Similarly carbon 2's hydrogens will get indices 6, 7, and 9. If any hydrogens are already explicit in the MOL file you do not need to reassign their indices. Show the new adjacency matrix for Caffeine.

3. Using your new adjacency matrix compute the following mappings discussed in lecture for Penicillin A, Penicillin G, and Caffeine.

   - Molecular formula (consider atoms C, O, N, H and S)
   - Label paired with length = 3
   - Depth first search (all cycles, double traverse, no compression, depth = 2). This should be a binary (0/1) map

4. Using your mappings compute the following kernels $k(G_1, G_2)$ for all $\binom{3}{2}$ pairs of Penicillin A, G, and Caffeine.

   - Molecular formula
   - Moleculr formula + MinMax
   - Label paired
   - Label paired + MinMax
   - Depth first search
   - Depth first search + Tanimoto

2 Kernel Properties

Let k_1 and k_2 be valid kernels.

$$k_1(x, y) = \phi_1(x)^T \phi_1(y) \tag{1}$$

$$k_2(x, y) = \phi_2(x)^T \phi_2(y) \tag{2}$$

Should that kernel k where $k = k_1 + k_2$ is a valid kernel by explicitly constructing a corresponding feature map ϕ such that $k(x, y) = \phi(x)^T \phi(y)$