BMI 503 Homework 1 (20 Pts)

* Due Time: Nov. 10th 23:59 (3 weeks)
* If you cannot finish all, submit solutions for some problems to get partial credits.
* Given data files from UCI Repository (breast cancer and promoter, files attached)

Problem 1: data loading (10 pts)

* Convert the data into a list of lists
  + A list of data, each datum is a patient or other person/object instances
  + Each datum is a list of attributes (same length, same correspondence)
  + If continuous value, keep the value
  + If categorical value/nominal value, construct a set/dictionary with the attribute being the key
  + If missing value “?”, use None
  + You can keep the patient ID and the protein name as a separate attribute for information. It will not be used for learning in our next homework though.

Problem 2: data curation (10 pts)

* Compute basic statistics of each attribute
  + All statistics of the attributes form a list (one attribute per item)
  + If a continuous-valued attribute: save max, min, mean, standard deviation (use dictionary here would be easier)
  + If a categorical: construct a dictionary to count frequencies
  + Skip missing value ”None” when you calculate these statistics
* Generate a normalized data
  + Filling missing value:
    - for continuous-valued attributes: use mean
    - For categorical-valued: use maximal frequency value
  + Normalize the data
    - For continuous-valued attributes: new val = (org val – mean) / std
    - For categorical: map values to 0, 1, 2, …., L, have a dictionary mapping org val to new val
* Write the curated data into a new text file, using the normalized and filled values
* Float numbers needs only 3 decimal numbers.