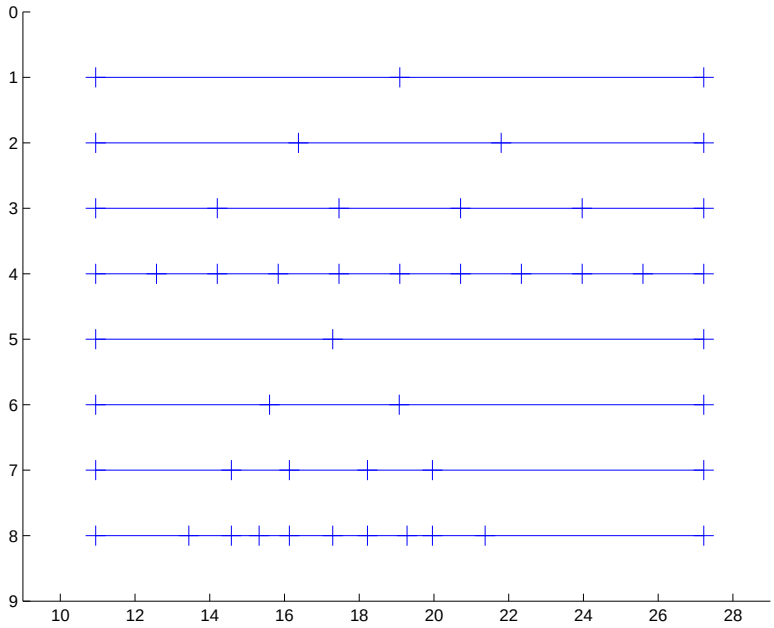
Homework 2 Sample Solutions
COMP 766-001 – Machine Learning for Bioinformatics

[1] (10 pts) The plot below depicts the different ways of discretizing cell radius. On each horizontal line, the vertical ticks indicate the boundaries of bins. Lines 1 through 4 are equal-width (EW) discretizations. Lines 5 through 8 are equal count (EC) discretizations.



The exact bin boundaries are, not including the end points:

Scheme	Bin boundaries
EW-2	19.0850
EW-3	16.3733 21.7967
EW-5	14.2040 17.4580 20.7120 23.9660
EW-10	12.5770 14.2040 15.8310 17.4580 19.0850 20.7120 22.3390 23.9660 25.5930
EC-2	17.2900
EC-3	15.6000 19.0700
EC-5	14.5800 16.1300 18.2200 19.9600
EC-10	13.4400 14.5800 15.3200 16.1300 17.2900 18.2200 19.2800 19.9600 21.3700

These result in the discretized contingency tables:

```
EW-2   cell radius (discretized)
norec    44   107
recur    22    25

EW-3   cell radius (discretized)
norec    11    70    70
recur     5    28    14

EW-5   cell radius (discretized)
```

```
norec     3    15    47    59    27
recur     3     6    22    10     6

EW-10   cell radius (discretized)
norec     0     3     7     8    26    21    24    35    19     8
recur     2     1     1     5    13     9     4     6     5     1

EC-2   cell radius (discretized)
norec    68    83
recur    32    15

EC-3   cell radius (discretized)
norec    44    54    53
recur    23    12    12

EC-5   cell radius (discretized)
norec    27    28    31    34    31
recur    14    11    10     4     8

EC-10   cell radius (discretized)
norec    14    13    13    15    13    18    18    16    15    16
recur     7     7     6     5     7     3     1     3     5     3
```

[2] (10 pts) The results of chi-square tests are shown below.

Scheme	D.O.F.	χ^2-value	p-value
EW-2	1	5.03578977032549	$\leq$0.025
EW-3	2	4.08373256305481	$\leq$0.2
EW-5	4	8.90692156689424	$\leq$0.1
EW-10	9	15.3407662106295	$\leq$0.1
EC-2	1	7.61957780378602	$\leq$0.01
EC-3	2	6.27617277202071	$\leq$0.05
EC-5	4	6.7812309107615	$\leq$0.2
EC-10	9	10.4960655831491	$\leq$1

[3] (10 pts) The table below contains estimated mutual information for each discretization (based on the maximum likelihood estimate of the joint distribution), as well as a p-value for rejecting the null hypothesis that the variables are independent. The p-value is estimated by a permutation test with 10,000 repeats. For each repeat, the second column of the data matrix was randomly permuted and the mutual information was estimated again. The p-value is the fraction of those 10,000 repeats for which the estimated mutual information was at least as large as the estimated mutual information based on the original (discretized) data matrix.

```
                    EW-2   EW-3   EW-5   EW-10  EC-2   EC-3   EC-5   EC-10
Est. mutual info. | 0.0177 0.0153 0.0320 0.0544 0.0283 0.0221 0.0265 0.0423
p-value           | 0.0365 0.133  0.0763 0.1298 0.0077 0.0503 0.1302 0.2713
```

[4] (10 pts) The p-values for rejecting the null hypothesis of independence are similar for the χ^2-test and the

mutual information score. The surprise to me was how quickly the significance of association between the variables decreased with increasing number of bins for cell radius. Only for the EW-2 and EC-2 discretizations is there a strong case for an association. p-values for the EC-3 model also suggest an association, though the significance is marginal at around 0.05, and one must keep in mind that when testing lots of different discretizations, some may appear significant by chance. In hindsight, it would have been interesting to try 4 bins, as this would result in dividing the EW-2 or EC-2 bins in half. Overall, I favor the EC-2 because: (1) it produces a significant result according to either p-value (χ^2-test or mutual information), (2) with only two bins for cell radius, that result is likely to be real, and not a result of overfitting the data, and (3) the EC-2 discretization has, by definition, more balanced bins than EW-2. But EW-2 is also good.

[5] (10 pts) This is a little more complicated than the other problems, but the basic idea is that we want to use equal count bins. That is, suppose we have N samples of a continuous random variable, and we want to discretize into K bins or intervals. After discretization, let N_i be the number of samples in interval i. We basically want all N_i to be approximately K/N. If N happens to be a multiple of K, then we can do exactly that. Otherwise, some bins would have $\lfloor K/N \rfloor$ samples and some will have $\lceil K/N \rceil$. Intuitively, the reason that this maximizes the information content of the discretized variable is that a random variable with K possible outcomes has maximum information content when all outcomes are equally likely. (If you had forgotten that, then you can prove it by choosing the probabilities of the K outcomes, p_1, p_2, …, p_K, to maximize the information, subject to the constraint that $\sum_i p_i = 1$.)

Now, for a real proof. I will assume that all of the N samples are distinct. In this case, by choosing the intervals appropriately we can assign each N_i to be in the set $\{0, 1, \ldots, N\}$, so long as we respect the condition $\sum_i N_i = N$. (Intervals are mutually exclusive and cover the whole data set.)

Claim: Assume we know a discretization that maximizes the entropy and let N_1, N_2, …, N_K be the number of samples in the K bins. Each of the N_i is at most one greater than each other N_i.

Proof by contrapositive: Assume that N_a is at least two greater than N_b for some indexes a and b. Consider what happens to the entropy if we shift the intervals around so that bin a loses one sample, bin b gains one sample, and all other bins have the same number of samples. (This is always possible to do if the samples are distinct.) If the entropy increases, then we have shown that the original assignment is not optimal, proving the claim by contrapositive.

The original entropy is

$$\sum_i -\frac{N_i}{N} \log_2 \frac{N_i}{N}$$

and the new entropy is

$$\sum_{i \neq a,b} -\frac{N_i}{N} \log_2 \frac{N_i}{N} - \frac{N_a - 1}{N} \log_2 \frac{N_a - 1}{N} - \frac{N_b + 1}{N} \log_2 \frac{N_b + 1}{N} .$$

The change in entropy upon making the switch is thus

$$\Delta H = -\frac{N_a - 1}{N} \log_2 \frac{N_a - 1}{N} - \frac{N_b + 1}{N} \log_2 \frac{N_b + 1}{N} + \frac{N_a}{N} \log_2 \frac{N_a}{N} + \frac{N_b}{N} \log_2 \frac{N_b}{N} .$$

It is not immediately obvious whether ΔH is positive or not. Let

$$f(z) = -\frac{N_a - z}{N} \log_2 \frac{N_a - z}{N} - \frac{N_b + z}{N} \log_2 \frac{N_b + z}{N} .$$

Then

$$\Delta H = f(1) - f(0) = \int_{z=0}^{1} \frac{\partial f}{\partial z} dz \, ,$$

where the second equality is just an application of the Fundamental Theorem of Calculus. What is $\frac{\partial f}{\partial z}$? If you go through a bunch of algebra, which I'll omit here, you can find that

$$\frac{\partial f}{\partial z} = \frac{1}{N} \log_2 \frac{N_a - z}{N_b + z} \, .$$

It is important to notice that for $z \in [0, 1)$ this is strictly positive. For $z = 1$, the partial derivative is zero only if $N_a = N_b + 2$, and is positive if $N_a > N_b + 2$. (We have assumed that $N_a \geq N_b + 2$, so we don't need to consider the case $N_a < N_b + 2$.) Because the partial derivative is strictly positive, except possibly at $z = 1$, then the integral above is strictly positive, so ΔH is strictly positive. This proves the claim.

Now, back to the original problem. If N is a multiple of K, then the optimal solution must be $N_i = N/K$ for all i. Otherwise, at least one of the N_i would be strictly less than N/K and one would be strictly greater, and the difference of at least two samples between those two $N_i's$ would imply the solution is not optimal. If N is not a multiple of K, then, by similar reasoning, all the N_i must be $\lfloor N/K \rfloor$ or $\lceil N/K \rceil$. And because $\sum_i N_i = N$, any such discretization has the same number of N_i being $\lfloor N/K \rfloor$ and the same number being $\lceil N/K \rceil$. So, a discretization optimizes the entropy if and only if all N_i are equal to $\lfloor N/K \rfloor$ or $\lceil N/K \rceil$ and $\sum_i N_i = N$.