問題１　n 個のレコードの集合の中に m 個の正のレコードが存在する場合を考える。いま条件 T を満たすレコードの数を x、その中の正のレコードの数を y とするとき、クラス分類器としての条件 T の良否を次に定めるエンロトピーゲイン Ent(x, y) により評価する。

$$\mathrm{Ent}(x,y)=\mathrm{ent}\left(\frac{m}{n}\right)-\left(\frac{x}{n}\right)\mathrm{ent}\left(\frac{y}{x}\right)-\left(\frac{n-x}{n}\right)ent\left(\frac{m-y}{n-x}\right)$$

$$\mathrm{ent}(p)=\begin{cases}-p\log_2 p-(1-p)\log_2(1-p) & 0<p<1\\ 0 & p=0\text{ または }1\end{cases}$$

1. ent$(p) = 1$ となる p の値を答えよ。
2. Ent(x, y) が $\mathrm{ent}\left(\frac{m}{n}\right)$ および 0 となる x, y の値を各々答えよ。
3. Ent(x, y) は $m/n = y/x$ のとき最小値をとり、 さらに次の性質を満たす凸関数であることが知られている。

   $$\mathrm{Ent}(\lambda x_1+(1-\lambda)x_2,\lambda y_1+(1-\lambda)y_2)\le\lambda\mathrm{Ent}(x_1,y_1)+(1-\lambda)\mathrm{Ent}(x_2,y_2)$$

   （ただし $0\le\lambda\le1$)

   n=16, m=8 のとき Ent(8,8), Ent(11,8), Ent(16,8) の大小関係を理由とともに述べよ。
   同様に Ent(8,4), End(8,3), Ent(8,0) の大小関係も理由とともに示せ。
4. エントロピーゲインは、コストの小さい決定木を生成する際に利用されるが、コストが最小の木を効率的に生成できるとは限らない（コストは根から各葉ノードにいたる道の長さの総和とする)。この背後にある NP 困難性について簡単に述べよ。

問題２　長さ G のゲノム配列を、ランダムに切断した長さ L の断片（ゲノムから無作為に選択した位置から始まる長さ L の部分配列）が N 個与えられたとする。

1. ２つの断片が長さ $L\theta$ ($0<\theta\le1$)だけ端と端が重なるとき１つの配列として結合する操作を繰り返して断片をつなげていったとき生成される連続な配列をコンティグ(contig)と呼ぶ。コンティグ総数の期待値は以下の値で近似できることを示せ。

   $$Ne^{-(1-\theta)\frac{LN}{G}}$$

2. このコンティグ総数は理想的な値である。現実にはこの数を大きく上回り、短いコンティグが多数生成されることが殆どである。その原因について述べよ。
3. 近年のシークエンシング技術の高速化について述べよ。

問題３　近年、遺伝子発現量解析等で利用される Support Vector Machine について以下の問いに答えよ。

1. Fisher の線形判別法と Support Vector Machine が顕著に異なるクラス分類をする例を挙げ、その違いを述べよ。
2. ソフトマージンとカーネルトリックについて述べよ。