下記の問題は過去に本講義のテスト問題として使用したが、本年度の範囲外の問題も含まれている。

問題

N個のデータの集合 $S=\{(x_{1i}, x_{2i}, x_{3i}, y_i) \mid x_{1i}, x_{2i}, x_{3i}, y_i \in \{0,1\}, i=1, \ldots, N\}$ において、テスト属性 T_1, T_2, T_3 の値を x_{1i}, x_{2i}, x_{3i}、目標属性の値を y_i とする。テスト属性から目標属性の値を予測する決定木を求めることを考える。Entropy gain を最大化するテスト属性を選択する貪欲(greedy)アルゴリズムが、必ずしも根ノードから葉ノードに至るテスト数の総和を最小にしない集合 S と決定木の例を示せ。

問題

k-means 法で、誤差二乗平均の最小化に失敗する例を示せ。

問題

遺伝子発現量を解析する際には様々な機械学習手法が用いられる。これらの学習法に関する以下の問いに対して簡潔に答えよ(各問題 200～500 文字程度)。

(a) トップダウンにデータを分割してゆく階層型クラスタリング diana 法について述べよ。

(b) 決定木の大きさを小さくしようとするときの問題点と、解決案について述べよ。

(c) 2つの異なる細胞から各々 cDNA を無作為に抽出して各 cDNA の発現頻度を調べ、どちらかの細胞に有意に偏って発現している遺伝子を判定する際に、注意しなければならない問題点と解決案について述べよ。

(d) Fisher 法と Support Vector Machine の違いについて述べよ。

問題

クラス分類に関する以下の問いに答えよ。

(1) クラス分類問題とは何か説明せよ。

(2) クラス分類の手法である決定木について述べ、決定木の良否を判断する基準を示せ。

(3) クラス分類における overfitting 問題とは何か説明せよ。

(4) クラス分類のための手法である AdaBoost を説明せよ。また最終のクラス分類器がエラーする率の上界について成り立つ性質を解説せよ。

Problem

Consider to assemble two genomic regions that contain the same repetitive region by using the whole genome shotgun approach. Describe a case when mate-pair information is useful in resolving the ambiguity between two possible layouts of shotgun fragments.

Problem

Consider to generate a decision tree that always outputs the correct objective value to any record in the training dataset. Design such a training dataset that the cost of any decision tree is eight. (The cost of a decision tree is the sum of the lengths of the paths from the root to each leaf node.)

Problem

Consider AdaBoost algorithm listed below:

AdaBoost

Let $(\vec{x}_1, y_1),...,(\vec{x}_N, y_N)$ denote the input records.

Let w_i^t be the weight of *i*-th record at the *t*-th round.

Set $w_i^1 = 1 / N$ in the initial step.

Repeat the following three steps for each $t = 1, ..., T$.

1. $$p_i^t = \begin{cases} w_i^t \Big/ \sum_{i=1}^N w_i^t & \text{if } \sum_{i=1}^N w_i^t > 0 \\ 0 & \text{otherwise} \end{cases}$$
2. Generate such a hypothesis h_t that its weighted error is less than 1/2, *i.e.*,
$$\varepsilon_t = \sum_{i=1}^N p_i^t \left| h_t(\vec{x}_i) - y_i \right| < 1/2$$
3. Update the weight of each record as follows, and increment *t*.
$$\beta_t = \varepsilon_t / (1 - \varepsilon_t) \qquad w_i^{t+1} = w_i^t \beta_t^{1-|h_t(\vec{x}_i)-y_i|}$$

(1) Design such a training dataset that AdaBoost fails to output the first hypothesis.

(2) Design such a training dataset that for *t*=2, ..., *T,* any hypothesis is qualified; that is, its weighted error ratio is less than 1/2.