確認問題 (1)

1. 数値データをカテゴリデータに変換する必要があるのはどのような場合か考察せよ。また、このような変換の具体例を示せ。
2. カテゴリデータをニューラルネットワークの入力とする際に必要な処理について考えよ。
   （ヒント：都道府県の情報はどう扱えばよいか）

確認問題（１） 解答例

1. 別のデータと突き合わせることによって、個人が特定できてしまう場合がある。そのような場合は、数値をぼやかしてカテゴリで表現することによって、そのデータが表す個人を特定できないようにする。具体的には、数値データの値を特定の範囲で離散化する。たとえば年齢を表す数値データを 20 代、 30 代 ... のようにカテゴリデータに変換する。

確認問題（１） 解答例

2. ニューラルネットワークは数値データを入力とすることが前提の学習アルゴリズムである。このような場合、 0 または 1 の二値をとるダミー変数をカテゴリーの数だけ用意する。たとえば都道府県名を表すカテゴリデータは、 47 次元の one-hot ベクトル (1 つの次元の値だけが 1 で、残りは 0) となる。

確認問題 (2)

1.決定木の学習に関する以下の記述の空欄 (A) ～ (J) を埋めよ。

事例番号	規模	収益	成長性	株価
1	小	普通	高	上昇
2	大	少ない	低	下降
3	大	少ない	高	下降
4	大	多い	高	上昇
5	小	少ない	高	下降
6	大	多い	普通	上昇
7	小	普通	普通	下降
8	大	普通	普通	下降

確認問題 (2)　解答例

A 0.95

B 0.92

C 0.97

D 0.95

E 0.34

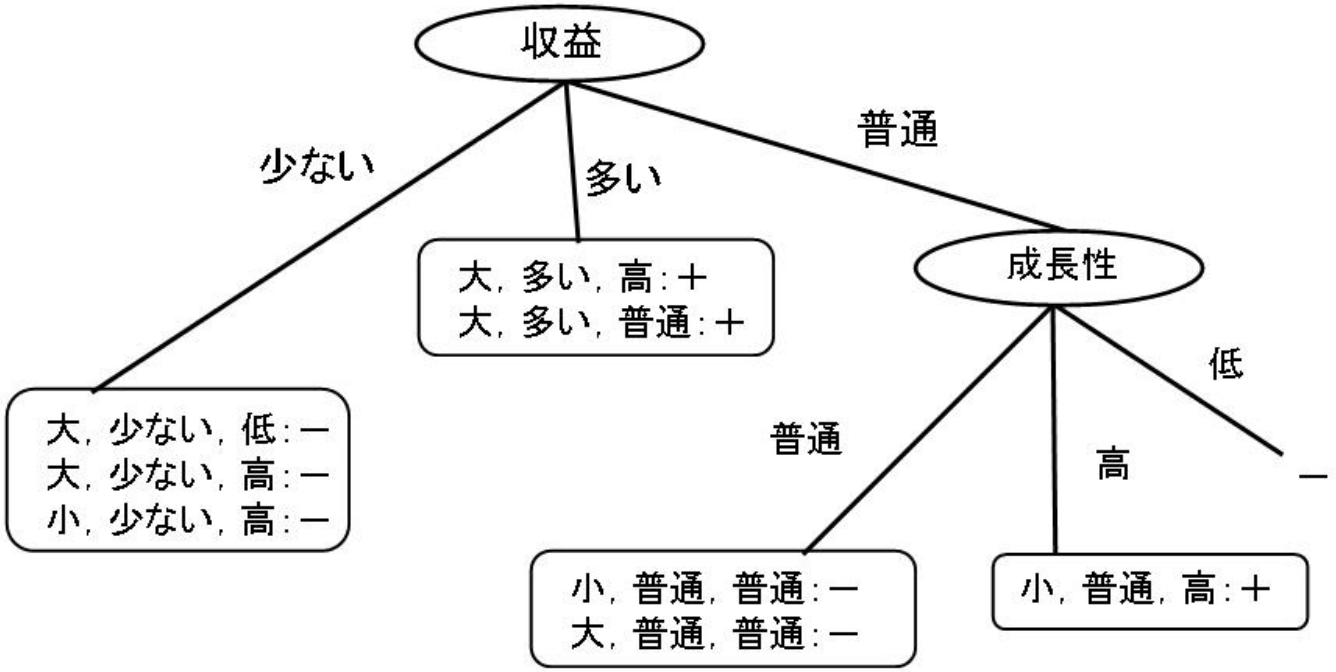
F　収益

G　普通

H　成長性

I　低

J　下降



確認問題 (3)

1. ある病気の検査法は、その病気の患者には 99% 、そうでない人には 3% の確率で陽性反応を示す。また、その病気の患者の割合は 0.1% であるとする。この検査で陽性反応が出たとき、その病気である確率をグラフィカルモデルを書いて求めよ。
2. 同じ病気に対する別種の検査は、その病気の患者には 98% 、そうでない人には 2% の確率で陽性反応を示す。 1 に続いて、この別種の検査でも陽性が出たときに、その病気である確率をグラフィカルモデルを書いて求めよ。

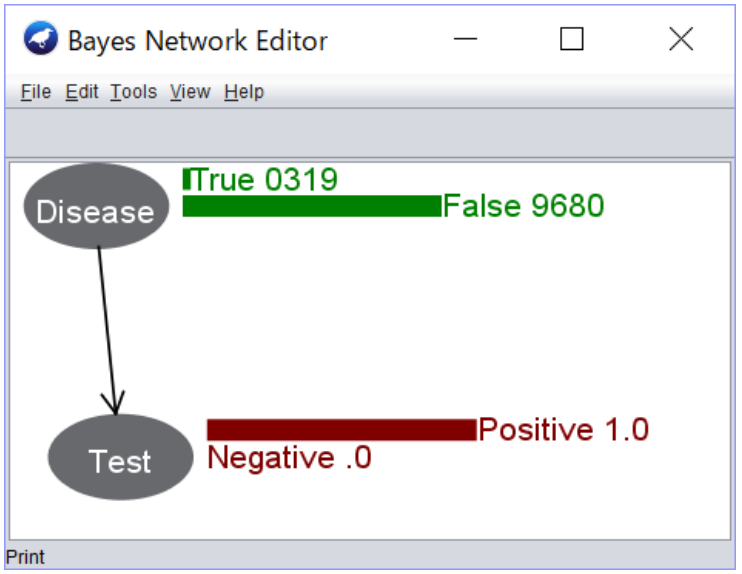
確認問題 (3)　解答例

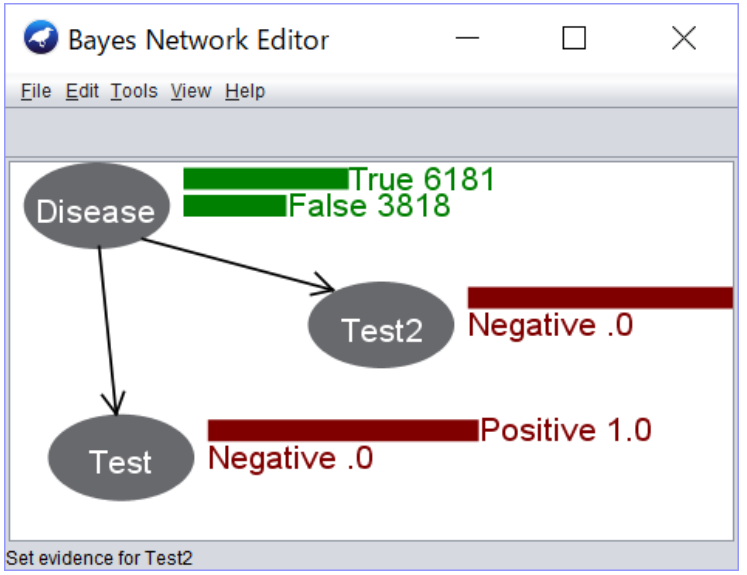
1.

$$P(\text{病気} \mid \text{陽性}) = \frac{0.99 \times 0.001}{0.999 \times 0.03 + 0.001 \times 0.99} = 0.032$$

2.

$$P(\text{病気} \mid \text{陽性}) = \frac{0.98 \times 0.032}{0.968 \times 0.02 + 0.032 \times 0.98} = 0.618$$





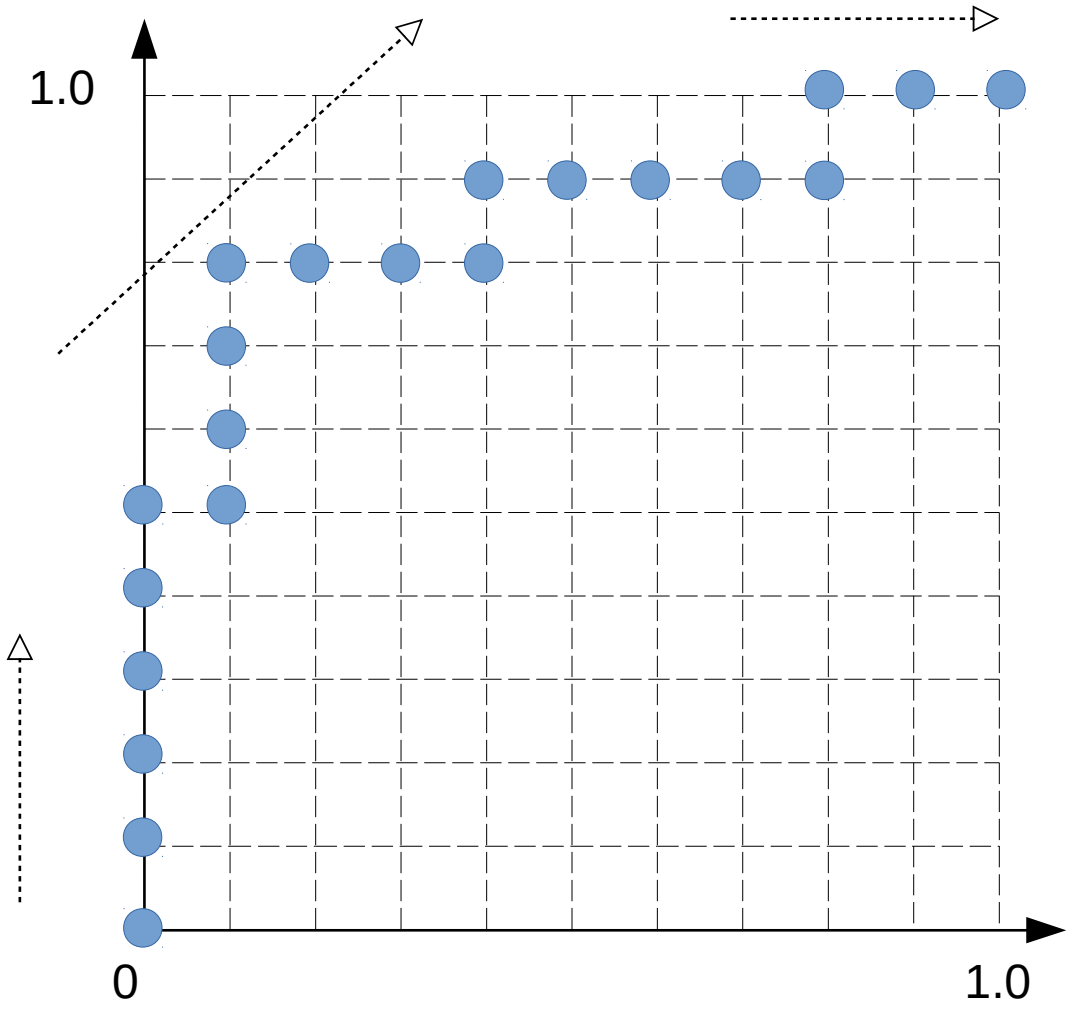
確認問題 (4)

1.統計的識別手法を用いると、ある検査の結果から一定確率以上で病気が疑われる場合に再検査を実施するなどの判断ができる。そのような判断に用いるROC 曲線（教科書 p.31 ）を右のデータから作成せよ。

No.	正解	確率
1	1	0.97
2	1	0.91
3	1	0.89
4	1	0.86
5	1	0.85
6	0	0.70
7	1	0.69
8	1	0.68
9	1	0.59
10	0	0.52
11	0	0.49
12	0	0.48
13	1	0.38
14	0	0.29
15	0	0.25
16	0	0.22
17	0	0.18
18	1	0.15
19	0	0.11
20	0	0.10

確認問題 (4)　解答例

$$TPR = \frac{TP}{N_{positive}}$$



$$FPR = \frac{FP}{N_{negative}}$$

確認問題 （５）

1.カテゴリ特徴のデータに対する回帰問題を考える。下記のスピーカー価格データを用いて回帰木を作成するとき、どの特徴を根とするべきか。

model	condition	leslie	price
B3	excellent	no	4513
T202	fair	yes	625
A100	good	no	1051
T202	good	no	270
M102	good	yes	870
A100	excellent	no	1770
T202	fair	no	99
A100	good	yes	1900
E112	fair	no	77

確認問題（５）　解答例

- model=[A100, B3, E112, M102, T202]
  - [1051, 1770, 1900] ,4513, 77, 870, [99, 270, 625]

分散：　139407

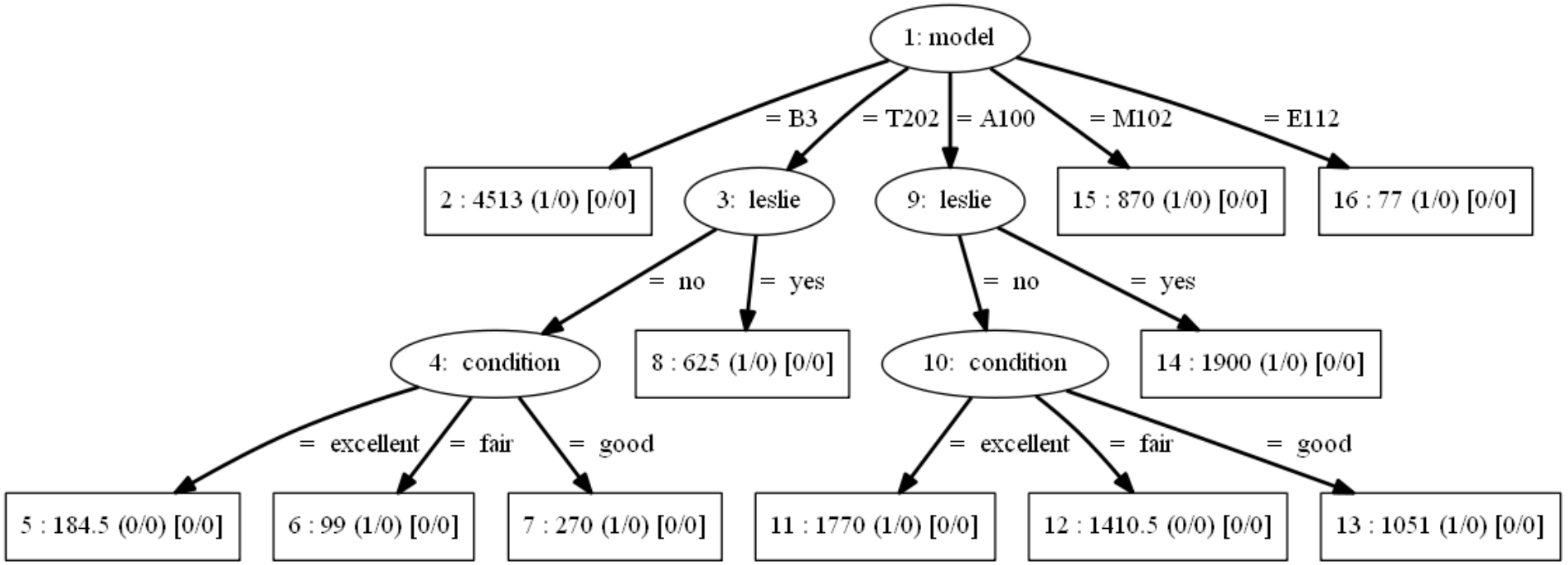
分散：　47994

$$\frac{3}{9} \times 139407 + \frac{3}{9} \times 47994 = 62467$$

- condition: 590538
- leslie: 1724527
- 重み付き分散和が最小のものは model 特徴

確認問題（５）　解答例

- 作成した回帰木



確認問題（６）

1. 文書分類において、入力文中の単語を次元とするベクトルを特徴とすると、助詞（「の」「は」「を」など）・助動詞（「です」など）・指示代名詞（「これ」、「それ」など）など、正例・負例のいずれにも頻繁に現れる特徴がいくつか存在する。この特徴の影響をなくす（あるいは少なくする）方法を考案せよ。

確認問題（６）　解答例

- ストップワード方式
  - 助詞・助動詞・代名詞など、どの文書にも頻出する単語は予め除外辞書（ストップワード）を作成しておき、ベクトルの次元に加えない
  - tf ・ idf 方式
    - $tf(w,d)$: ある文書 d に単語 w が出現した回数
    - $idf(w)$: 総文書数 D を単語 w が出現した文書数 D' で割ったものの対数

$$idf(w) = \log \frac{D}{D'}$$

    - tf ・ idf: これらの積