14.11

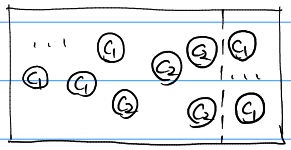
ツリーAの分割

予測値=C_1 予測値=C_2

C_1: 300個 C_1: 100個

C_2: 100個 C_2: 300個

ツリーBの分割



予測値=領域内で最多のクラス

予測値=C_2 予測値=C_1

C_1: 200個 C_1: 200個

C_2: 400個 C_2: 0個

ツリーAの左の葉ノードの誤分類率は

$$E_L^A = \frac{1}{N_{R_L}^A} \sum_{x_i \in R_L^A} I(t_i \neq C_1) = \frac{100}{400} = 0.25$$

ツリーAの右の葉ノードの誤分類率は

$$E_R^A = \frac{1}{N_{R_R}^A} \sum_{x_i \in R_R^A} I(t_i \neq C_2) = \frac{100}{400} = 0.25$$

ツリーBの左の葉ノードの誤分類率は

$$E_L^B = \frac{1}{N_{R_L}^B} \sum_{x_i \in R_L^B} I(t_i \neq C_2) = \frac{200}{600} = 0.33$$

ツリーBの右の葉ノードの誤分類率は

$$E_R^B = \frac{1}{N_{R_R}^B} \sum_{x_i \in R_R^B} I(t_i \neq C_1) = \frac{0}{200} = 0$$

問題文には反するが

とする。ツリーA, Bで各ノードの誤分類率が特に等しいということはない。

(ルリーAについて)

左右の葉ノードのデータ点割合

$p^A_{L C_1} = \frac{1}{N^A_L} \sum_{x_i \in R^A_L} I(t_i = C_1) = \frac{300}{400} = 0.75$

$p^A_{L C_2}$ = 上と同様 $= \frac{100}{400} = 0.25$

$p^A_{R C_1}$ = 上と同様 $= \frac{100}{400} = 0.25$

$p^A_{R C_2}$ = 上と同様 $= \frac{300}{400} = 0.75$

交差エントロピー誤差関数は

$Q^{A\,cross}_L = -\sum_{k \in \{C_1, C_2\}} p^A_{Lk} \ln p^A_{Lk} = -(p^A_{LC_1} \ln p^A_{LC_1} + p^A_{LC_2} \ln p^A_{LC_2})$

$= -(0.75 \ln 0.75 + 0.25 \ln 0.25) = 0.56$

$Q^{A\,cross}_R = -(p^A_{RC_1} \ln p^A_{RC_1} + p^A_{RC_2} \ln p^A_{RC_2})$

$= -(0.25 \ln 0.25 + 0.75 \ln 0.75) = 0.56$

このとき枝刈り基準は (注)(14.31)式は N_τ が掛けられる

$C^{A\,cross}(T) = \sum_{\tau=1}^{|T|} N_\tau Q_\tau(T) + \lambda |T| = 400 \cdot Q^{A\,cross}_L + 400 \cdot Q^{A\,cross}_R + \lambda \cdot 2 = 448 + 2\lambda$

ジニ係数は

$Q^{A\,Gini}_L = \sum_{k \in \{C_1, C_2\}} p^A_{Lk}(1 - p^A_{Lk}) = p^A_{L_1}(1 - p^A_{LC_1}) + p^A_{LC_2}(1 - p^A_{LC_2})$

$= 0.75 \times 0.25 + 0.25 \times 0.75 = 0.38$

$Q^{A\,Gini}_R = p^A_{RC_1}(1 - p^A_{RC_1}) + p^A_{RC_2}(1 - p^A_{RC_2}) = 0.25 \times 0.75 + 0.75 \times 0.25 = 0.38$

このとき枝刈り基準は

$C^{A\,Gini}(T) = 400 \cdot Q^{A\,Gini}_L + 400 \cdot Q^{A\,Gini}_R + \lambda \cdot 2 = 304 + 2\lambda$

(リリーBについて)

左右の葉ノードのデータ比割合

$$p^B_{LC_1} = \frac{1}{N^B_L} \sum_{x_i \in R^B_L} I(y_i = C_1) = \frac{200}{600} = 0.33$$

$$p^B_{LC_2} = 上と同様 = \frac{400}{600} = 0.67$$

$$p^B_{RC_1} = 上と同様 = \frac{200}{200} = 1$$

$$p^B_{RC_2} = 上と同様 = \frac{0}{200} = 0$$

交差エントロピー誤差関数は

$$Q^{B\,cross}_L = -\sum_{k \in \{C_1, C_2\}} p^B_{Lk} \ln p^B_{Lk} = -\left(p^B_{LC_1} \ln p^B_{LC_1} + p^B_{LC_2} \ln p^B_{LC_2} \right)$$

$$= -(0.33 \ln 0.33 + 0.67 \ln 0.67) = 0.63$$

$$Q^{B\,cross}_R = -\left(p^B_{RC_1} \ln p^B_{RC_1} + p_{RC_2} \ln p_{RC_2} \right)$$

$$= -(1 \ln 1 + 0 \ln 0) = 0$$

このとき枝刈り基準は

$$C^{B\,cross}_{(T)} = \sum_{\tau=1}^{|T|} N_\tau Q_\tau(T) + \lambda |T| = 600 \cdot 0.63 + 200 \cdot 0 + \lambda \cdot 2 = 378 + 2\lambda$$

ジニ係数は

$$Q^{B\,Gini}_L = \sum_{k \in \{C_1, C_2\}} p^B_{Lk} (1 - p^B_{Lk}) = p^B_{LC_1} (1 - p^B_{LC_1}) + p^B_{LC_2} (1 - p^B_{LC_2})$$

$$= 0.33 \cdot 0.67 + 0.67 \cdot 0.33 = 0.44$$

$$Q^{B\,Gini}_R = p^B_{RC_1} (1 - p^B_{RC_1}) + p^B_{RC_2} (1 - p^B_{RC_2}) = 0$$

このとき枝刈り基準は

$$C^{B\,Gini}_{(T)} = \sum_{\tau=1}^{|T|} N_\tau Q_\tau(T) + \lambda |T| = 600 \cdot 0.44 + 200 \cdot 0 + \lambda \cdot 2 = 264 + 2\lambda$$

まとめると

交差エントロピーを使った枝刈り基準は

ツリーAにおいて　448＋2λ

ツリーBにおいて　378＋2λ

なのでツリーBのほうが小さくなっている

ジニ係数を使った枝刈り基準は

ツリーAにおいて　304＋2λ

ツリーBにおいて　264＋2λ

なのでツリーBのほうが小さくなっている