1.8. Decision Trees

1.8.7. Mathematical formulation
1.8.7.1. Classification criteria
1.8.7.2. Regression criteria

11T4056H　　永田 純平

1.8.7. Mathematical formulation

- トレーニングベクトル$x_i \in R^n$,ラベルベクトル$y \in R^I$
  決定木は再帰的に分岐する

$$Q_{left}(\theta) = (x, y)|x_j <= \ t_m$$
$$Q_{right} = Q \setminus Q_{left}(\theta)$$

- 分岐した領域の不純度は関数H()で計算される

$$\mathrm{G}(\mathrm{Q}, \theta) = \frac{n_{left}}{N_m} H\big(Q_{left}(\theta)\big) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

- 不純度を最小にするパラメータθを選ぶ

$$\theta^* = arg \min_{\theta} G(Q, \theta)$$

1.8.7.1. Classification criteria

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

- Gini:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

- Cross-Entropy:

$$H(X_m) = \sum_k p_{mk} \log(p_{mk})$$

- Misclassification:

$$H(X_m) = 1 - \max(p_{mk})$$

1.8.7.2. Regression criteria

- 対象が連続値である場合

- 最小二乗誤差は、不純度を最小化する

$$c_m = \frac{1}{N_m} \sum_{i \in N_m} y_i$$

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - c_m)^2$$