Machine Learning, Tutorial 7
University of Bern

Abdelhak Lemkhenter, abdelhak.lemkhenter@inf.unibe.ch

06/11/2017

Decision Trees

1. Indicate whether the following statements about decision trees are True of False. Justify your answers:
   - Decision trees are prone to overfitting **[TRUE/FALSE]**
   - Decision trees are suitable for linear problems **[TRUE/FALSE]**

   Solution

   - True, the number of region grows exponentially with depth
   - False. c.f. course notes

2. Given the data set given by table 1, find the first split of the data that maximizes information gain. Note that all feature in table 1 are categorical.

Rain	Coat	Wind speed	Umbrella
None	Yes	High	No
Light	No	Low	Yes
Light	Yes	Low	No
None	Yes	High	No
Heavy	Yes	Low	Yes
Heavy	Yes	High	No
None	No	Low	No
None	No	High	No

Table 1: Use of an umbrella

Solution
$R_1 = \{Wind\ speed = High\}$ and $R_2 = \{Wind\ speed = Low\}$ or $R_1 = \{Rain = None\}$ and $R_2 = \{Rain \neq None\}$. In order to maximize the information gain, one needs to minimize the weighted cross-entropy of the split, $entropy : p \mapsto -p \times log(p) - (1-p)log(1-p)$

R_1	weighted cross-entropy
Rain = None	$0.5 \times entropy(0.5) \sim 0.34$
Rain = Light	$1/4 \times entropy(0.5) + 3/4 \times entropy(1/6) \sim 0.51$
Rain = Heavy	$1/4 \times entropy(0.5) + 3/4 \times entropy(1/6) \sim 0.51$
Wind speed = Low	$0.5 \times entropy(0.5) \sim 0.34$
Coat = Yes	$5/8 \times entropy(1/5) + 3/8 \times entropy(1/3) \sim 0.55$

3. The Gini index is defined as $G = 1 - \sum_{i=1}^{c} p_i^2$ where c is the number of classes. In the case of binary classification:
   (a) Show that $G = 2p(1-p)$, where p is probability of the positive class.
   (b) Show that G is strictly concave in p.

(a) $G = 1 - p^2 - (1-p)^2 = 2p(1-p)$.

(b) $\frac{d^2G}{dp^2} = -2 < 0$.

Regression Trees

1. We consider the following regression problem $y_i = f(x_i) + \epsilon_i$ where $\{\epsilon_i\}_i$ are I.I.D, $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ and $\forall i$ x_i and ϵ_i are independent.

   (a) Show that $L_{squared}(R) = \frac{\sum_{i \in R}(y_i - \hat{y})^2}{|R|} = Var(y)$, where $\hat{y} = \frac{\sum_{i \in R} y_i}{|R|}$.

   (b) Given the data set shown on figure 1, show that the optimal split is at $x = 0.5$. f is given by equation 1

$$f(x) = \begin{cases} 6, & \text{if } x > 0.5 \\ 4, & \text{else} \end{cases} \tag{1}$$

Hint: $Var(x+y) = Var(x) + Var(y)$ if x and y are independent.

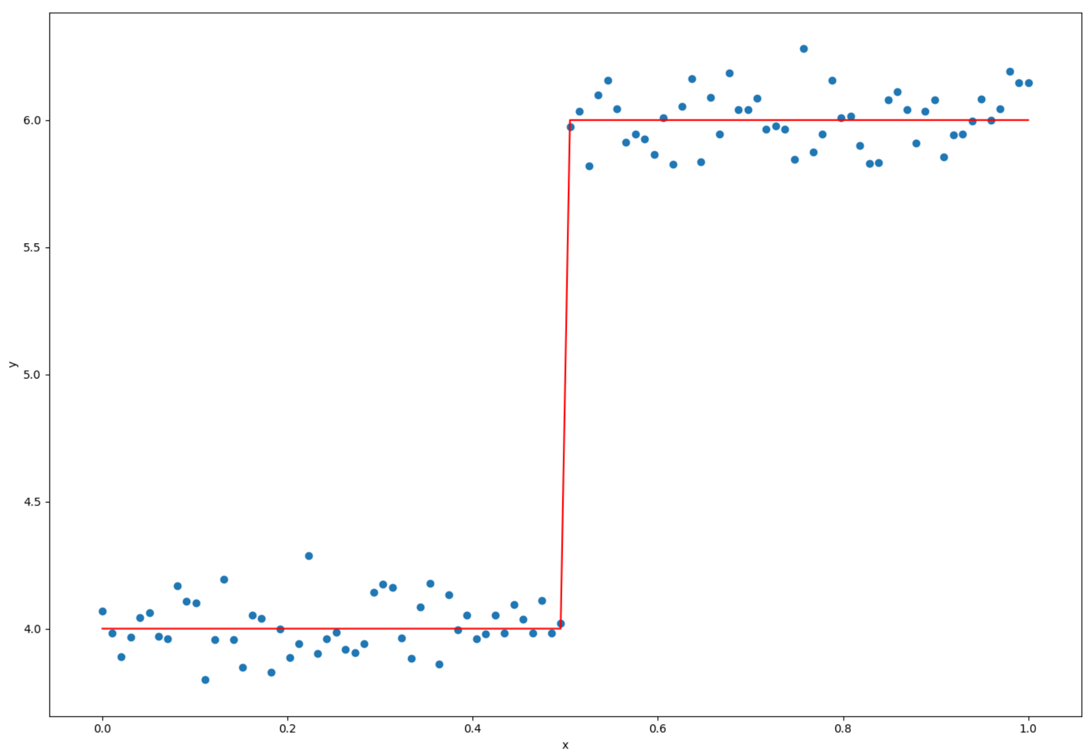

Figure 1: The red line represent the function f.

(a) $Var(y) = E[(y - E[y])^2]$.

(b) $Var(y) = Var(f(x)) + Var(e) = Var(f(x)) + \sigma^2$. In order to minimize the variance in each split, we need to minimize $Var(f(x))$. Since the variance of a constant is zero, the optimal split is at $x = 0.5$.