February 6, 2018

# Problem Set 2

Sanchit Singhal

INF 385T – Introduction to Machine Learning with Danna Gurari

Spring 2018

School of Information

① Entropy $= -\sum_{i=1}^{n} p_i \log_2 p_i$ , $C_1 =$ Yes $C_2 =$ No , $P_{c1} = 5/8$ $P_{c2} = 3/8$

a)

current entropy $= -\left(P_{c1}\log_2 P_{c1} + P_{c2}\log_2 P_{c2}\right)$

$= -\left(\cancelto{-0.424}{\frac{5}{8}\log_2\frac{5}{8}} + \cancelto{-0.531}{\frac{3}{8}\log_2\frac{3}{8}}\right) = 0.955$

**If we split on Feature Sick:**

left tree: entropy when Yes $= -\left(\cancelto{-0.5}{\frac{2}{4}\log_2\frac{2}{4}} + \cancelto{-0.5}{\frac{2}{4}\log_2\frac{2}{4}}\right) = 1$

right tree: entropy when No $= -\left(\cancelto{-0.311}{\frac{3}{4}\log_2\frac{3}{4}} + \cancelto{-0.5}{\frac{1}{4}\log_2\frac{1}{4}}\right) = 0.811$

$\Rightarrow$ Information Gain $= 0.955 - \left(\cancelto{0.5}{\frac{4}{8}(1)} + \cancelto{0.4055}{\frac{4}{8}(0.811)}\right)$

$= 0.955 - 0.9055 = 0.0495$

**If we split on Feature Rain:**

<u>Ordered Attribute values</u>

$\{0, 0.2, 0.4, 0.4, 0.7, 1, 2, 3\}$

① ② ③ ④ ⑤ ⑥ ⑦

① Greater than 0.1 cm

left tree: entropy when No $= -\left(\cancelto{0}{\frac{0}{1}\log_2\frac{0}{1}} + \cancelto{0}{\frac{1}{1}\log_2\frac{1}{1}}\right) = 0$

right tree: entropy when Yes $= -\left(\cancelto{-0.461}{\frac{4}{7}\log_2\frac{4}{7}} + \cancelto{-0.524}{\frac{3}{7}\log_2\frac{3}{7}}\right) = 0.985$

$\Rightarrow$ Information Gain $= 0.955 - \left(\cancelto{0}{\frac{1}{8}(0)} + \cancelto{0.862}{\frac{7}{8}(0.985)}\right)$

$= 0.955 - 0.862 = 0.093$

② Greater than 0.3 cm

left tree: entropy when NO $= -\left(\cancelto{0}{\frac{0}{2}\log_2\frac{0}{2}} + \cancelto{0}{\frac{2}{2}\log_2\frac{2}{2}}\right) = 0$

right tree: entropy when Yes $= -\left(\cancelto{-0.5}{\frac{3}{6}\log_2\frac{3}{6}} + \cancelto{-0.5}{\frac{3}{6}\log_2\frac{3}{6}}\right) = 1$

$\Rightarrow$ Information Gain $= 0.955 - \left(\cancelto{0}{\frac{2}{8}(0)} + \cancelto{0.75}{\frac{6}{8}(1)}\right)$

$= 0.955 - 0.75 = 0.205$

③ Greater than 0.4 cm

left tree: entropy when NO $= -\left(\cancelto{0}{\frac{0}{3}\log_2\frac{0}{3}} + \cancelto{0}{\frac{3}{3}\log_2\frac{3}{3}}\right) = 0$

right tree: entropy when Yes $= -\left(\cancelto{-0.442}{\frac{3}{5}\log_2\frac{3}{5}} + \cancelto{-0.529}{\frac{2}{5}\log_2\frac{2}{5}}\right) = 0.971$

$\Rightarrow$ Information Gain $= 0.955 - \left(\cancelto{0}{\frac{3}{8}(0)} + \cancelto{-0.607}{\frac{5}{8}(0.971)}\right)$

$= 0.955 - 0.607 = 0.348$

④ Greater than 0.55 cm

left tree: entropy when NO $= -\left(\cancelto{0}{\frac{0}{4}\log_2\frac{0}{4}} + \cancelto{0}{\frac{4}{4}\log_2\frac{4}{4}}\right) = 0$

right tree: entropy when Yes $= -\left(\cancelto{-0.311}{\frac{3}{4}\log_2\frac{3}{4}} + \cancelto{-0.5}{\frac{1}{4}\log_2\frac{1}{4}}\right) = 0.811$

$\Rightarrow$ Information Gain $= 0.955 - \left(\cancelto{0}{\frac{4}{8}(0)} + \cancelto{0.406}{\frac{4}{8}(0.811)}\right)$

$= 0.955 - 0.406 = 0.549$

⑤ Greater than 0.85 cm

left tree: entropy when no $= -\left( \overset{-0.464}{\frac{1}{5}\log_2\frac{1}{5}} + \overset{-0.256}{\frac{4}{5}\log_2\frac{4}{5}} \right) = 0.72$

right tree: entropy when Yes $= -\left( \overset{-0.39}{\frac{2}{3}\log_2\frac{2}{3}} + \overset{-0.528}{\frac{1}{3}\log_2\frac{1}{3}} \right) = 0.92$

$\Rightarrow$ Information Gain $= 0.955 - \left( \overset{0.45}{\frac{5}{8}(0.72)} + \overset{0.345}{\frac{3}{8}(0.92)} \right)$

$= 0.955 - 0.795 = 0.16$

⑥ Greater than 1.5 cm

left tree: entropy when no $= -\left( \overset{-0.43}{\frac{1}{6}\log_2\frac{1}{6}} + \overset{-0.22}{\frac{5}{6}\log_2\frac{5}{6}} \right) = 0.65$

right tree: entropy when Yes $= -\left( \overset{0}{\frac{2}{2}\log_2\frac{2}{2}} + \overset{0}{\frac{0}{2}\log_2\frac{2}{2}} \right) = 0$

$\Rightarrow$ Information Gain $= 0.955 - \left( \overset{0.4875}{\frac{6}{8}(0.65)} + \overset{0}{\frac{2}{8}(0)} \right)$

$= 0.955 - 0.4875 = 0.4625$

⑦ Greater than 2.5 cm

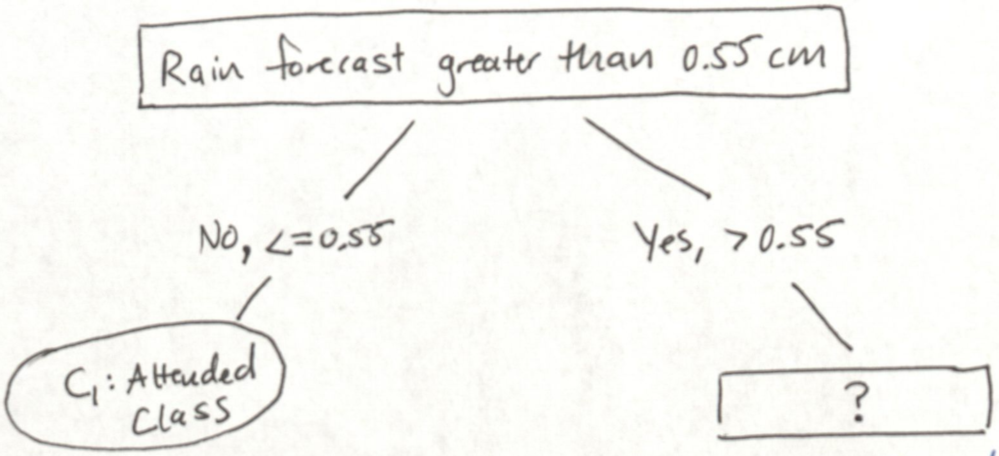
left tree: entropy when no $= -\left( \overset{-0.516}{\frac{2}{7}\log_2\frac{2}{7}} + \overset{-0.347}{\frac{5}{7}\log_2\frac{5}{7}} \right) = 0.863$

right tree: entropy when Yes $= -\left( \overset{0}{\frac{1}{1}\log\frac{1}{1}} + \overset{0}{\frac{0}{1}\log_2\frac{0}{1}} \right) = 0$

$\Rightarrow$ Information Gain $= 0.955 - \left( \overset{0.755}{\frac{7}{8}(0.863)} + \overset{0}{\frac{1}{8}(0)} \right)$

$= 0.955 - 0.755 = 0.195$

The greatest information gain is with Feature Rain at 0.55 cm:




⇒ This branch has 3 $C_2$ & 1 $C_1$

Therefore, we must find split that maximizes information gain again!

$$\text{current entropy} = -\left(\cancelto{-0.5}{\frac{1}{4}\log_2\frac{1}{4}} + \cancelto{-0.311}{\frac{3}{4}\log_2\frac{3}{4}}\right) = 0.81125$$

If we split on Feature Sick:

$$\text{left tree: entropy when Yes} = -\left(\cancelto{0}{\frac{0}{2}\log_2\frac{0}{2}} + \cancelto{0}{\frac{2}{2}\log_2\frac{2}{2}}\right) = 0$$

$$\text{right tree: entropy when No} = -\left(\cancelto{-0.5}{\frac{1}{2}\log_2\frac{1}{2}} + \cancelto{-0.5}{\frac{1}{2}\log_2\frac{1}{2}}\right) = 1$$

$$\Rightarrow \text{Information Gain} = 0.81125 - \left(\cancelto{0}{\frac{2}{4}(0)} + \cancelto{0.5}{\frac{2}{4}(1)}\right)$$

$$= 0.31125$$

## If we split on Feature Rain:

**Ordered Attribute values**

$$\{0.7, \overset{\uparrow}{\underset{8}{}} 1, \overset{\uparrow}{\underset{9}{}} 2, \overset{\uparrow}{\underset{10}{}} 3\}$$

8) Greater than 0.85 cm

left tree: entropy when no $= -\left(\cancelto{0}{\frac{0}{1}\log_2\frac{0}{1}} + \cancelto{0}{\frac{1}{1}\log_2\frac{1}{1}}\right) = 0$

right tree: entropy when Yes $= -\left(\cancelto{-0.53}{\frac{1}{3}\log_2\frac{1}{3}} + \cancelto{-0.39}{\frac{2}{3}\log_2\frac{2}{3}}\right) = 0.92$

$\Rightarrow$ Information Gain $= 0.81125 - \left(\cancelto{0}{\frac{1}{4}(0)} + \cancelto{0.69}{\frac{3}{4}(0.92)}\right) = \boxed{0.12125}$

9) Greater than 1.5 cm

left tree: entropy when no $= -\left(\cancelto{-0.5}{\frac{1}{2}\log_2\frac{1}{2}} + \cancelto{-0.5}{\frac{1}{2}\log_2\frac{1}{2}}\right) = 1$

right tree: entropy when Yes $= -\left(\cancelto{0}{\frac{0}{2}\log_2\frac{0}{2}} + \cancelto{0}{\frac{2}{2}\log_2\frac{2}{2}}\right) = 0$

$\Rightarrow$ Information Gain $= 0.81125 - \left(\cancelto{0.5}{\frac{2}{4}(1)} + \cancelto{0}{\frac{2}{4}(0)}\right) = \boxed{0.31125}$

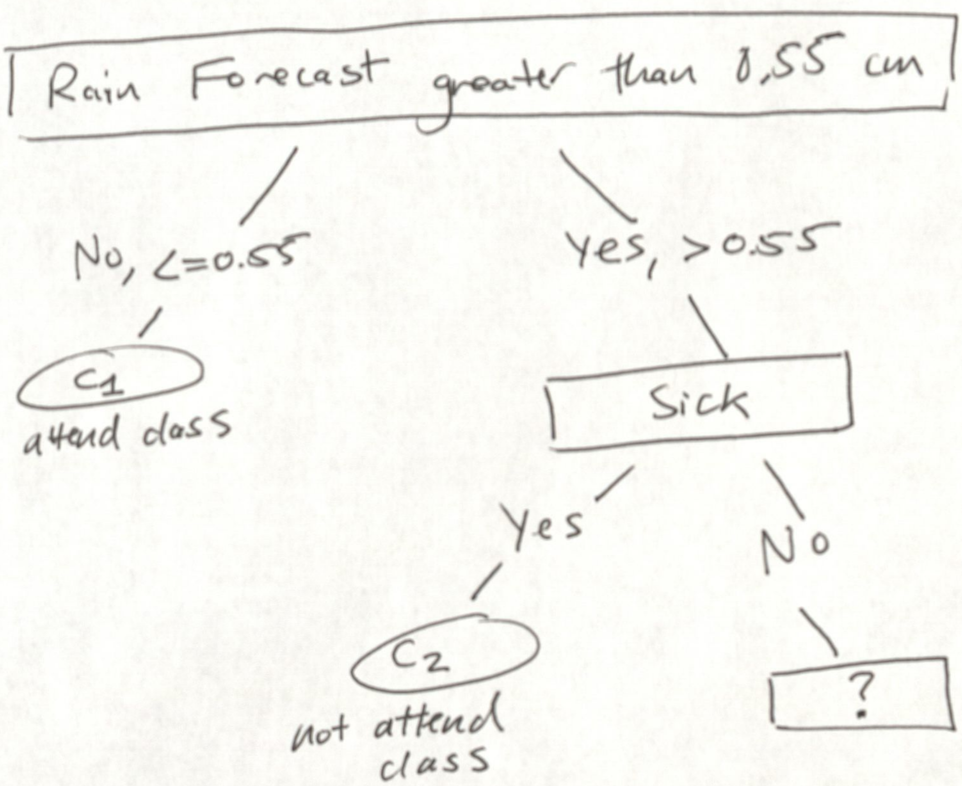
10) Greater than 2.5 cm

left tree: entropy when no $= -\left(\cancelto{-0.39}{\frac{2}{3}\log_2\frac{2}{3}} + \cancelto{-0.53}{\frac{1}{3}\log_2\frac{1}{3}}\right) = 0.92$

right tree: entropy when Yes $= -\left(\cancelto{0}{\frac{1}{1}\log_2\frac{1}{1}} + \cancelto{0}{\frac{0}{1}\log_2\frac{0}{1}}\right) = 0$

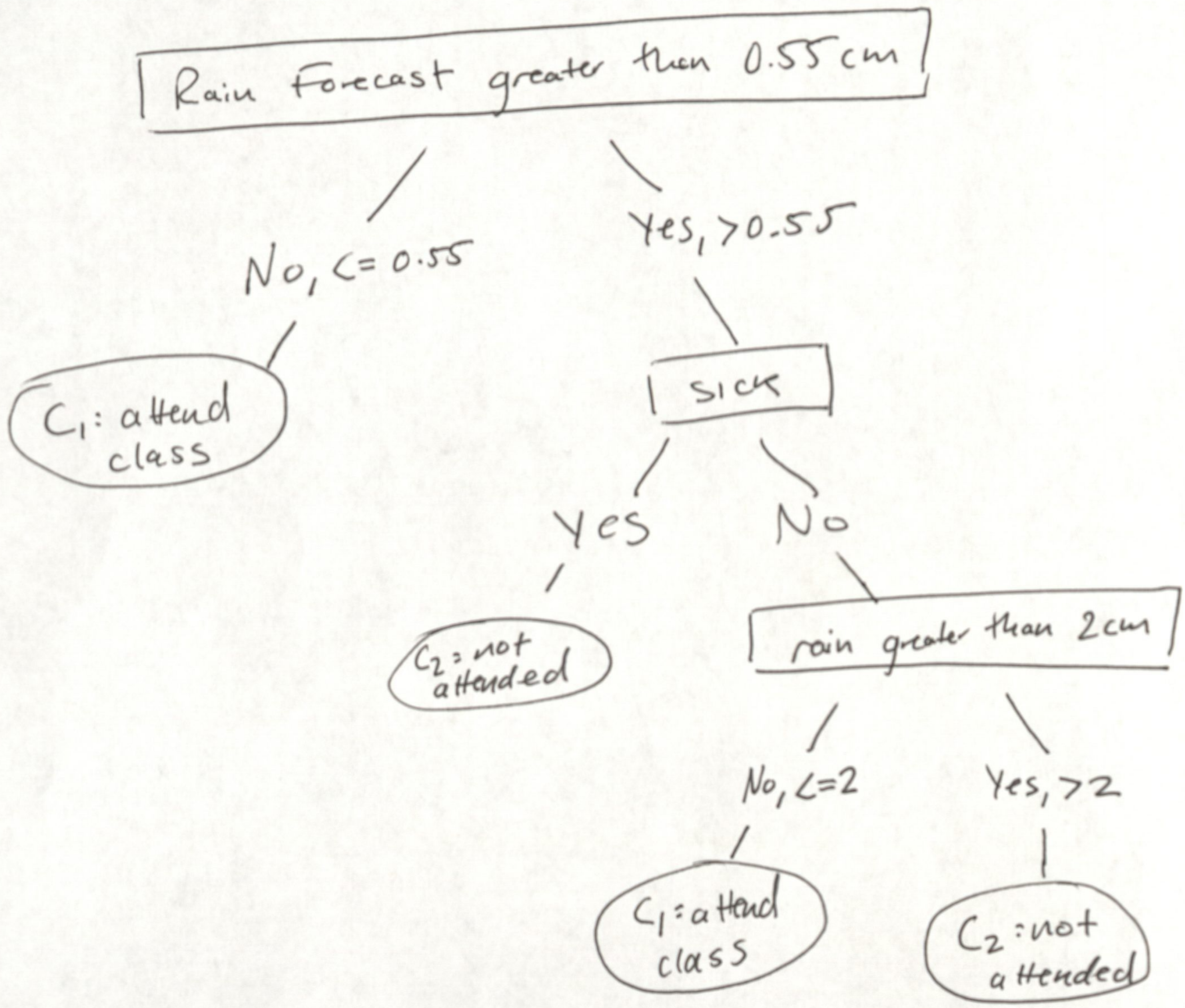
$\Rightarrow$ Information Gain $= 0.81125 - \left(\cancelto{0.69}{\frac{3}{4}(0.92)} + \cancelto{0}{\frac{1}{4}(0)}\right) = \boxed{0.12125}$

Because the information gain is the same when splitting by the feature Sick & feature rain at 1.5 cm at 0.31125, we can choose either one. I believe the model will be able to model a more complex relationship with 2 parameters and therefore I chose to include the Sick feature. It must be said that this adds difficulty for future data sets used on this model:



Since the 2 remaining datapoints have different classes, the only remaining info gain would be to split evenly between Feature Rain at 2 cm:

Decision Tree



b) test data

⇒ p9 : rain = 1.5 cm

Predict : Yes
Actual : Yes

⇒ p10 : rain = 1.7 cm

Predict : NO
Actual : No

⇒ p11 : rain = 0.5 cm

predict : Yes
actual : Yes

⇒ p12 : rain = 0.8 cm

predict : No
actual : No

c)

$$\text{Precision} = P = \frac{T_P}{T_P + F_P} = \frac{2}{2+0} = 1 = 100\%$$

$$\text{Recall} = R = \frac{T_P}{T_P + F_N} = \frac{2}{2+0} = 1 = 100\%$$

$$F_1 = 2\frac{PR}{(P+R)} = \frac{2(1)(1)}{(1+1)} = \frac{2}{2} = 100\%$$

$$\left.\begin{array}{l} \Rightarrow \text{high precision} \\ \quad\ \text{high recall} \end{array}\right\} \text{ideal system}$$

⇒ many results w/ correct labels

d)

$n = 4$

Actual

| Predicted | Yes | no |
|---|---|---|
| Yes | $T_P = 2$ | $F_P = 0$ |
| no | $F_N = 0$ | $T_N = 2$ |

②

a)

$P(C_1) \Rightarrow P(\text{Attended}) = \frac{5}{8} = 0.625$

$P(C_2) \Rightarrow P(\text{Not Attended}) = \frac{3}{8} = 0.375$

$P(x_1 | C_1)$

P ( Rain | Attended )

- mean = 0.4 ($\mu$)
- standard deviation = 0.37 ($\sigma$)
- variance = 0.14 ($\sigma^2$)

$$* \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$P(x_1 | C_2)$

P ( Rain | Not Attended )

- mean = 1.9 ($\mu$)
- standard deviation = 1.15 ($\sigma$)
- variance = 1.33 ($\sigma^2$)

b) test data

$\Rightarrow$ p9 : rain = 1.5 cm

$\boxed{P(\text{Rain} | \text{Not Attended}) = 0.32654}$

P (Rain | Attended) = 0.01298

Predict : No
Actual : Yes

$\Rightarrow$ p10 : rain = 1.7 cm

$\boxed{P(\text{rain} | \text{not attended}) = 0.34170}$

P(rain | attended) = 0.00225

Predict : No
Actual : No

$\Rightarrow$ p11 : rain = 0.5 cm

$P(\text{rain} \mid \text{not attended}) = 0.16534$    Predict : Yes

$\boxed{P(\text{rain} \mid \text{attended}) = 1.0396}$    Actual : Yes

$\Rightarrow$ p12 : rain = 0.8 cm

$P(\text{rain} \mid \text{not attended}) = 0.21955$    Predict : Yes

$\boxed{P(\text{rain} \mid \text{attended}) = 0.60106}$    Actual : No

c) $$\text{Precision} = \frac{T_P}{T_P + F_P} = \frac{1}{1+1} = \frac{1}{2} = 50\%$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} = \frac{1}{1+1} = \frac{1}{2} = 50\%$$

$$F_1 = 2\frac{PR}{(P+R)} = 2\frac{1/2 \cdot 1/2}{(1/2 + 1/2)} = 2(1/4) = 1/2 = 50\%$$

d) Confusion Matrix

$n = 4$

| Predicted \ Actual | Yes | No |
|---|---|---|
| Yes | $T_P = 1$ | $F_P = 1$ |
| No | $F_N = 1$ | $T_N = 1$ |

3. Classification Models

Generative and Discriminative models follow different approaches to building the distribution. A generative model use the joint probability of all attributes when creating the distribution however a discriminative model applies conditional probability such that attributes are “given” other attributes. The generation of data plays an important role in the difference between the two models. Generative models use data generation to categorize the data to model the distribution of each class and thus usually work better on smaller datasets. Discriminative models just distinguish between the classes and therefore do not use data generation as a categorization metric. Hence, discriminative models end up making lesser assumption on the structure of the input than a generative model.

Example of both models:

Generative – Naïve Bayes – assumes conditional independence of the features

Discriminative – logistic regression – does not assume independence of the features