Assignment 4

Jiarong Ye

October 17, 2018

Q1

Logistic Regression:

1. Consider a logistic regression model parameterized by $\theta_0 = 14$ and $\theta_1 = -0.14$ and the following sample data.

Samples	X	y
x_1	80	1
x_2	20	1
x_3	120	0

- a) (10%) Calculate the probability that $y = 1$ for each x_i of the data set $(h_\theta(x))$.
- b) (10%) Calculate the value of objective (cost) function based on idea of SSE,
- c) (10%) Calculate the value of objective (cost) function based on log of Sigmoid.

For logistic regression learning:

(a)

$$h_\theta(x_1) = P(y = 1|x_1, \theta) = \frac{1}{1 + e^{-\theta^T x}} = \frac{1}{e^{-\theta_0 - \theta_1 \cdot x_1}}$$

$$= \frac{1}{e^{-14+0.14 \cdot 80}} = 0.943$$

$$h_\theta(x_2) = P(y = 1|x_2, \theta) = \frac{1}{1 + e^{-\theta^T x}} = \frac{1}{e^{-\theta_0 - \theta_1 \cdot x_2}}$$

$$= \frac{1}{e^{-14+0.14 \cdot 20}} = 1.000$$

$$h_\theta(x_3) = P(y = 1|x, \theta) = \frac{1}{1 + e^{-\theta^T x}} = \frac{1}{e^{-\theta_0 - \theta_1 \cdot x_3}}$$

$$= \frac{1}{e^{-14+0.14 \cdot 120}} = 0.057$$

(b)

Object Function based on the idea of SSE:

$$J(\theta) = J(\theta_0, \theta_1, ..., \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2, \text{ where } m = 3$$

Thus

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 = \frac{1}{6}[(0.943 - 1)^2 + (1.00 - 1)^2 + (0.057 - 0)^2] = 0.001083$$

(c)

Object Function based on the idea of log of sigmoid:

$$Err(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), if\ y = 1 \\ -\log(1 - h_\theta(x)), if\ y = 0 \end{cases} \tag{1}$$

So

$$\begin{aligned} J(h_\theta(x), y) &= \frac{1}{m}\sum_{i=1}^{m} Err(h_\theta(x), y) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log(h_\theta(x^i)) + (1 - y^i)(\log(1 - h_\theta(x^i))\right] \\ &= -\frac{1}{m}[1 \cdot \log(0.943) + 1 \cdot \log 1 + 1 \cdot \log(1 - 0.057)] \\ &= 0.0564 \end{aligned} \tag{2}$$

Q2

Decision Tree

2. Consider the following data set for binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	T	-
T	T	-
T	F	-

- a. (15%) Calculate the information gain (based on entropy) when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
- b. (15%) Calculate the gain (based on the Gini index) when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(a)

	A=T	A=F
C1=+	4	0
C2=-	3	3

	B=T	B=F
C1=+	3	1
C2=-	2	4

- entropy before split:
  - $$E = -0.4\log(0.4) - 0.6\log(0.6) = 0.971$$
  - entropy after splitting on A:
    - $$E(A_T) = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.985$$
    - $$E(A_F) = -\frac{3}{3}\log\frac{3}{3} = 0$$

    * info gain(A):

$$E - \frac{7}{10}E(A_T) - \frac{3}{10}E(A_F) = 0.971 - 0.7 \cdot 0.985 = 0.2815$$

  * entropy after splitting on B:
    *
$$E(B_T) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$
    *
$$E(B_F) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = 0.722$$
    * info gain(B):

$$E - \frac{5}{10}E(B_T) - \frac{5}{10}E(B_F) = 0.971 - 0.5 \cdot 0.971 - 0.5 \cdot 0.722 = 0.1245$$

Since $info\,gain(A) = 0.2815 > info\,gain(B) = 0.1245$, so if split on A, the reduction of entropy, i.e. the reduction of uncertainty is larger than if split on B. Thus **A should be selected for the Decision Tree.**

(b)

- Gini before split:
  *
$$G = 1 - 0.4^2 - 0.6^2 = 0.48$$
  * Gini after splitting on A:
    *
$$G(A_T) = 1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.4898$$
    *
$$G(A_F) = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0$$
    * info gain(A):

$$G - \frac{7}{10}G(A_T) - \frac{3}{10}G(A_F) = 0.48 - 0.7 \cdot 0.4898 = 0.137$$

  * Gini after splitting on B:
    *
$$G(B_T) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$
    *
$$G(B_F) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.32$$
    * info gain(B):

$$G - \frac{5}{10}G(B_T) - \frac{5}{10}G(B_F) = 0.48 - 0.5 \cdot 0.48 - 0.5 \cdot 0.32 = 0.08$$

Since $info\,gain(A) = 0.137 > info\,gain(B) = 0.08$, so if split on A, the reduction of Gini, i.e. the reduction of uncertainty is larger than if split on B. Thus **A should be selected for the Decision Tree.**

Q3

The following table summarizes a data set with three attributes A, B, C and two class labels +, − . Build a two-level decision tree.

A	B	C	+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	10
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	20

- a. (15%) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table (i.e., count matrix) and the gains in classification error rate.
- b. (15%) Repeat for the two children of the root node.
- c. (10%) How many instances are misclassified by the resulting decision tree?

(a)

	A=T	A=F
C1=+	25	25
C2=-	0	50

	B=T	B=F
C1=+	30	20
C2=-	20	30

	C=T	C=F
C1=+	25	25
C2=-	30	20

- Error before split:
  *

$$Err = 1 - \max(\frac{50}{100}, \frac{50}{100}) = 0.5$$

* Error after splitting on A:
  *
  $$Err(A_T) = 1 - \max(\frac{25}{25}, \frac{0}{25}) = 0$$
  *
  $$Err(A_F) = 1 - \max(\frac{25}{75}, \frac{50}{75}) = 0.333$$
  *
  $$\Delta Err(A) = Err - \frac{25}{100} \cdot Err(A_T) - \frac{75}{100} \cdot Err(A_F) = 0.5 - 0.75 \cdot 0.333 = 0.25025$$
* Error after splitting on B:
  *
  $$Err(B_T) = 1 - \max(\frac{30}{50}, \frac{20}{50}) = 0.4$$
  *
  $$Err(B_F) = 1 - \max(\frac{20}{50}, \frac{30}{50}) = 0.4$$
  *
  $$\Delta Err(B) = Err - \frac{50}{100} \cdot Err(B_T) - \frac{50}{100} \cdot Err(B_F) = 0.5 - 0.5 \cdot 0.4 \cdot 2 = 0.1$$
* Error after splitting on C:
  *
  $$Err(C_T) = 1 - \max(\frac{25}{55}, \frac{30}{55}) = 0.455$$
  *
  $$Err(C_F) = 1 - \max(\frac{25}{45}, \frac{20}{45}) = 0.444$$
  *
  $$\Delta Err(C) = Err - \frac{55}{100} \cdot Err(C_T) - \frac{45}{100} \cdot Err(C_F) = 0.5 - 0.55 \cdot 0.455 - 0.45 \cdot 0.444 = 0.04995$$

Since $\Delta Err(A) > \Delta Err(B) > \Delta Err(C)$, thus attribute A should be chosen as the first splitting attribute.

(b)

- first child
  * $A = T$ is pure and no further split is required
  * $A = F$

	B=T	B=F
C1=+	25	0
C2=-	20	30

	C=T	C=F
C1=+	0	25
C2=-	30	20

- Error before split:
  *
$$Err = 1 - \max(\frac{25}{75}, \frac{50}{75}) = 0.333$$

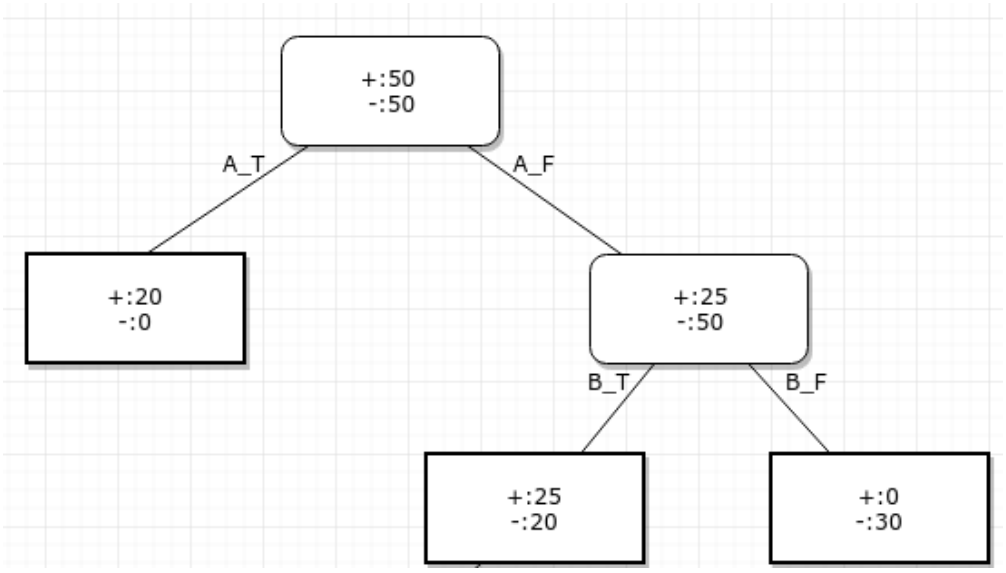
* Error after splitting on B:
  *
$$Err(B_T) = 1 - \max(\frac{25}{45}, \frac{20}{45}) = 0.444$$
  *
$$Err(B_F) = 1 - \max(\frac{0}{30}, \frac{30}{30}) = 0$$
  *
$$\Delta Err(B) = Err - \frac{45}{75} \cdot Err(B_T) - \frac{30}{75} \cdot Err(B_F) = 0.333 - 0.6 \cdot 0.444 - 0.4 \cdot 0 = 0.0666$$

* Error after splitting on C:
  *
$$Err(C_T) = 1 - \max(\frac{0}{30}, \frac{30}{30}) = 0$$
  *
$$Err(C_F) = 1 - \max(\frac{25}{45}, \frac{20}{45}) = 0.444$$
  *
$$\Delta Err(C) = Err - \frac{30}{75} \cdot Err(C_T) - \frac{45}{75} \cdot Err(C_F) = 0.333 - 0.4 \cdot 0 - 0.6 \cdot 0.444 = 0.0666$$

(c)

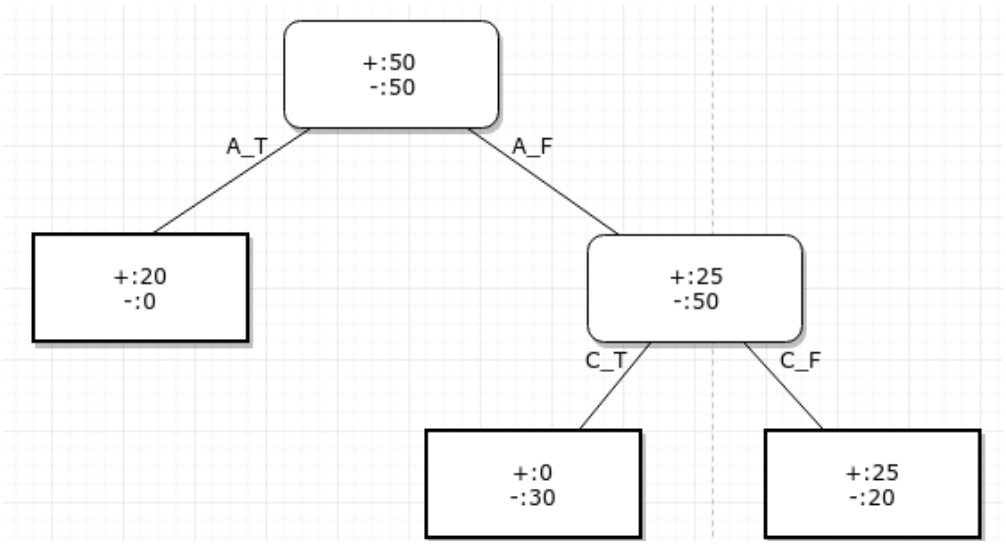
from part B:

- if B is the second child



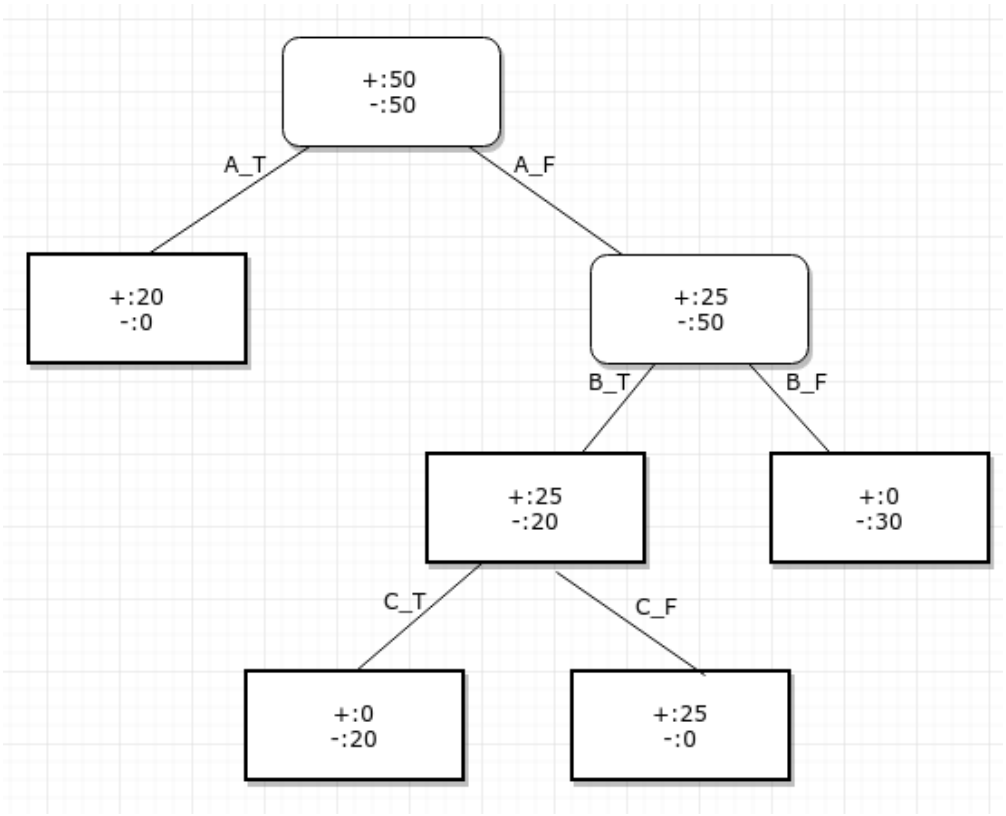
misclassified instances = 20

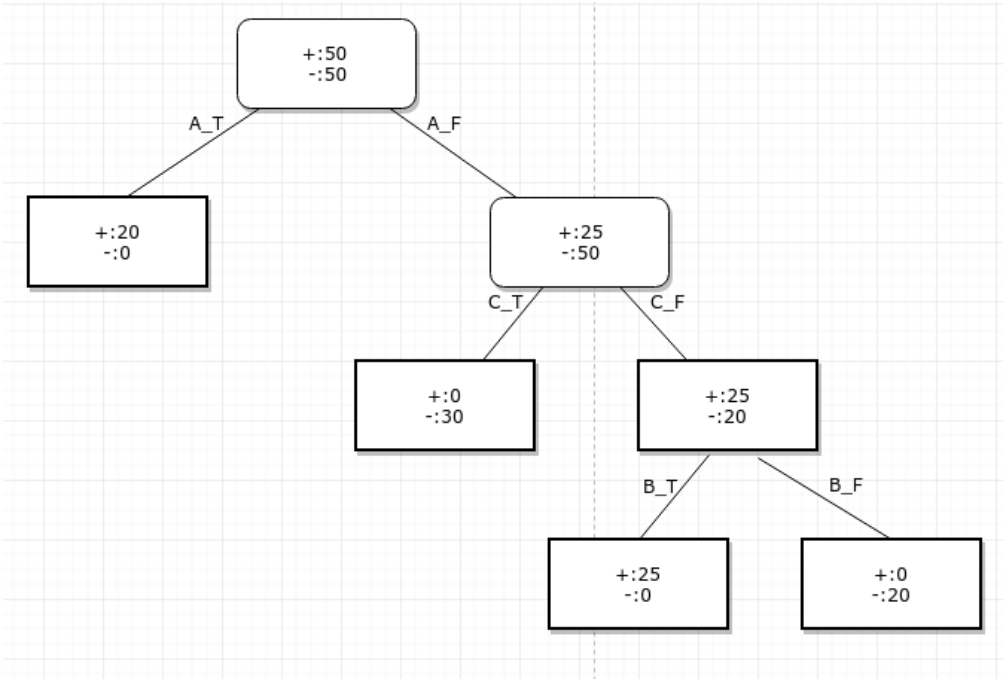
- if C is the second child



misclassified instances = 20

So, in conclusion, there are 20 misclassified instances from this decision tree.
But if the splitting continues,


+:50
-:50
A_T
A_F
+:20
-:0
+:25
-:50
B_T
B_F
+:25
-:20
+:0
-:30
C_T
C_F
+:0
-:20
+:25
-:0


+:50
-:50
A_T
A_F
+:20
-:0
+:25
-:50
C_T
C_F
+:0
-:30
+:25
-:20
B_T
B_F
+:25
-:0
+:0
-:20

Then we might be able to reduce the misclassified instance to 0.