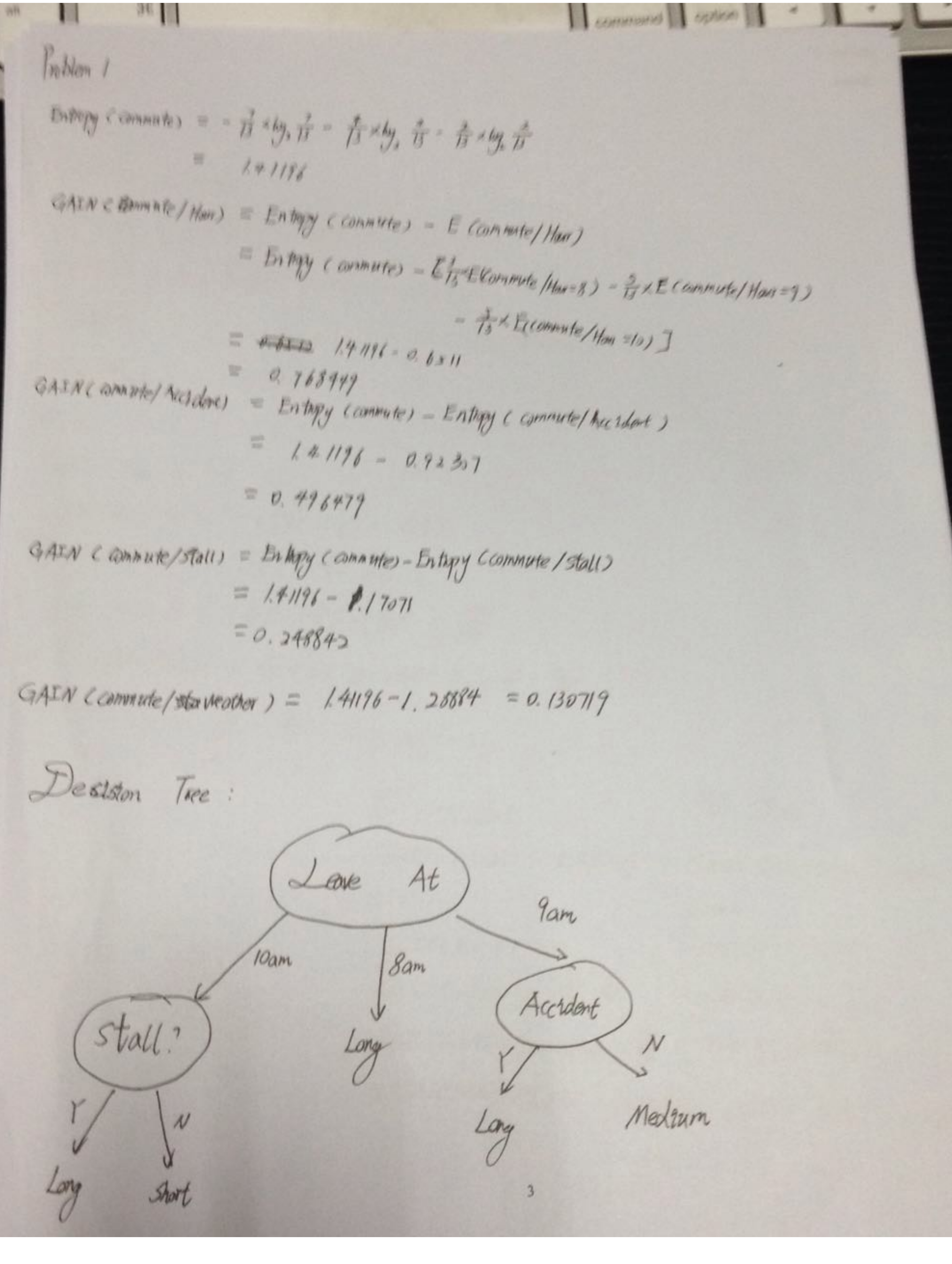


Problem 1

Entropy (commute) $= -\frac{7}{13} \times \log_2 \frac{7}{13} - \frac{4}{13} \times \log_2 \frac{4}{13} - \frac{2}{13} \times \log_2 \frac{2}{13}$

$= 1.41196$

GAIN (commute/Hour) = Entropy (commute) − E (commute/Hour)

= Entropy (commute) − E[$\frac{1}{13}$ × E(commute/Hour=8) − $\frac{5}{13}$ × E(commute/Hour=9)

− $\frac{3}{13}$ × E(commute/Hour=10)]

= ~~0.6511~~ 1.41196 − 0.6511

= 0.768449

GAIN (commute/Accident) = Entropy (commute) − Entropy (commute/Accident)

= 1.41196 − 0.92307

= 0.496477

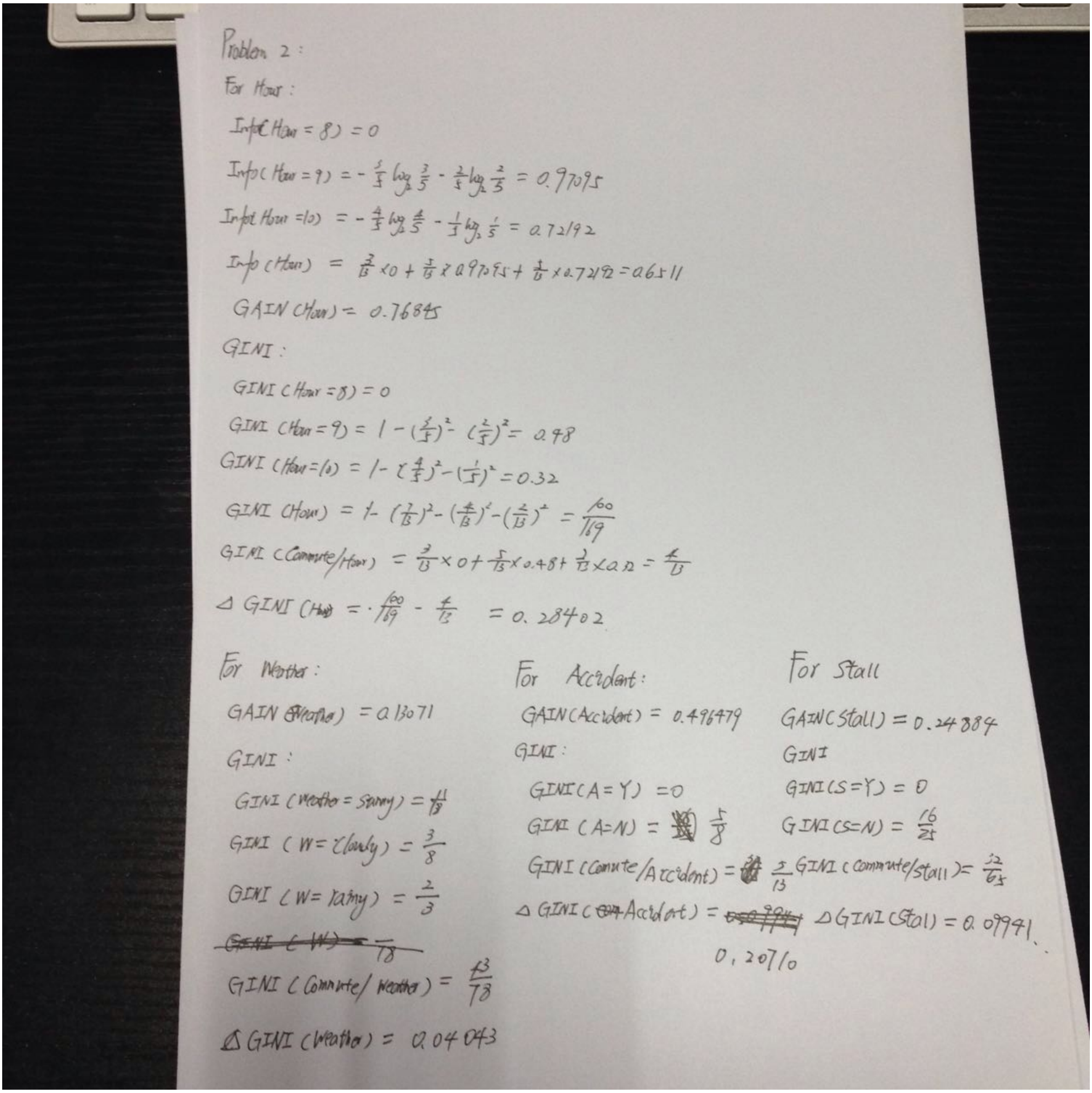
GAIN (commute/stall) = Entropy (commute) − Entropy (commute/stall)

= 1.41196 − 1.17071

= 0.248842

GAIN (commute/~~sta~~ weather) = 1.41196 − 1.28884 = 0.130719

Decision Tree :



Problem 2:

For Hour:

Info(Hour = 8) = 0

Info(Hour = 9) = $-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97095$

Info(Hour = 10) = $-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72192$

Info(Hour) = $\frac{3}{13}\times 0 + \frac{5}{13}\times 0.97095 + \frac{5}{13}\times 0.72192 = 0.6511$

GAIN(Hour) = 0.76845

GINI:

GINI(Hour = 8) = 0

GINI(Hour = 9) = $1-(\frac{3}{5})^2-(\frac{2}{5})^2 = 0.48$

GINI(Hour = 10) = $1-(\frac{4}{5})^2-(\frac{1}{5})^2 = 0.32$

GINI(Hour) = $1-(\frac{7}{13})^2-(\frac{4}{13})^2-(\frac{2}{13})^2 = \frac{100}{169}$

GINI(Commute/Hour) = $\frac{3}{13}\times 0 + \frac{5}{13}\times 0.48 + \frac{5}{13}\times 0.32 = \frac{4}{13}$

Δ GINI(Hour) = $\frac{100}{169} - \frac{4}{13} = 0.28402$

For Weather:

GAIN(Weather) = 0.13071

GINI:

GINI(Weather = Sunny) = $\frac{11}{18}$

GINI(W = Cloudy) = $\frac{3}{8}$

GINI(W = rainy) = $\frac{2}{3}$

~~GINI(W) = $\frac{}{78}$~~

GINI(Commute/Weather) = $\frac{13}{78}$

Δ GINI(Weather) = 0.04043

For Accident:

GAIN(Accident) = 0.496479

GINI:

GINI(A = Y) = 0

GINI(A = N) = $\frac{5}{8}$

GINI(Commute/Accident) = $\frac{5}{13}$

Δ GINI(Accident) = 0.20710

For Stall

GAIN(Stall) = 0.24884

GINI

GINI(S = Y) = 0

GINI(S = N) = $\frac{16}{25}$

GINI(Commute/Stall) = $\frac{32}{65}$

Δ GINI(Stall) = 0.09941

According to "**Theoretical Comparison between the Gini Index and Information Gain Criteria**" *(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.9764&rep=rep1&type=pdf)*, the frequency of disagreement of them is only 2%,therefore, it's hard to say which one is better.

But there are still some differences between the two algorithm. For ID3 algorithm and GAIN, log-likelihood might give higher scores to balanced portions when there are too many classes though.However, for CART and GINI, it can be nicer because it doesn’t have logarithms and you can find the closed form for its expected value and variance under random split assumption.