Problem 1

1. This is a classification problem.
2. There are 4 features.
3. 500 samples in total.

Problem 2

Yes, this is a disaster. An area over 0.9 can be excellent, while an area at around 0.5 indicates its randomness. If the area is only 0.01, it seems that everything goes to the opposite.

Problem 3

False. If there are less training data, it is likely to over fit noise. So that less training data, rather than more data, is likely to lead to over fit.

Problem 4

False. Since recall = TP / (TP + FN), where TP means true positive and FN means false negative, in a situation where FN (detecting positive instance as negative) is unacceptable, FN should be as less as possible. So I would prefer higher recall.

Problem 5

False. Precision = TP / (TP + FP), and in this case there are also a lot of false positive, which are actually negative detected as positive, so precision cannot be optimal.

Problem 6

a. High variance. Since training data is huge, trying an overfitting approach may be good to gain higher precision.

b. High bias. Since training data is relatively less, it is better to try an under fitting approach to avoid fitting for noise.

c. High variance. Overfitting approach may be helpful for non-linear relationship for more precise representation.

d. High variance. Since joint possibility density can be inferred, it may be better to use an overfitting approach for better precision.

Problem 7

a. True. Linear regression will eventually find a best solution for optimization.

b. True. Logistic regression uses gradient ascent on likelihood, which could gradually lead to an optimal solution.

c. False. For decision tree with ID3 algorithm, it finds optimal solution only for that single step, so it is possible that there exists a tree which has less depth than the solution.

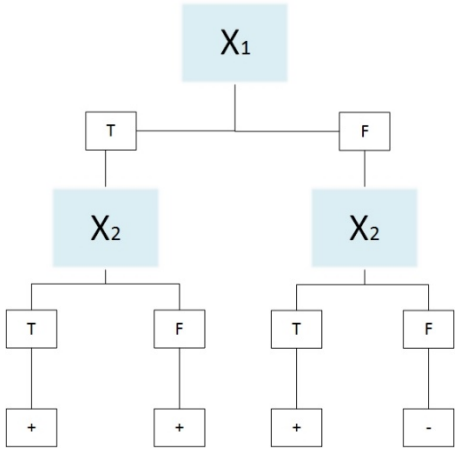
Problem 8

a. $H(Y) = -P(Y+)\log_2 P(Y+) - P(Y-)\log_2 P(Y-) = 0.9852$

b. $IG(X_1) = H(Y) - \left[\frac{8}{21}H\left(\frac{7}{8},\frac{1}{8}\right) + \frac{13}{21}H\left(\frac{5}{13},\frac{8}{13}\right)\right] = 0.1831$

$IG(X_2) = H(Y) - \left[\frac{10}{21}H\left(\frac{7}{10},\frac{3}{10}\right) + \frac{11}{21}H\left(\frac{5}{11},\frac{6}{11}\right)\right] = 0.0449$

c. Since IG(X_1) is higher, I choose X_1 as first branch. Following graph shows the decision tree:



Problem 9

In this problem, known conditions are: $P(Ebola) = 0.01$, $P(headache|Ebola) = 0.7$, and we can assume $P(headache|NoEbola) = 0.3$.

$P(Ebola, headache) = P(headache|Ebola) \times P(Ebola) = 0.007$

$P(NoEbola, headache) = P(headache|NoEbola) \times P(NoEbola) = 0.297$

According to Bayes' rule: $P(Ebola|headache) = \frac{P(headache|Ebola)P(Ebola)}{P(Ebola,headache)+P(NoEbola,headache)} = 0.0191$

So the probability of a patient has Ebola given she has severe headache is 0.0191.

Problem 10

$P(Yes) = 0.6, P(No) = 0.4$

Model for correct answers:

$\mu_{char,yes} = \frac{1}{3} \times (216 + 69 + 60) = 208$

$\sigma^2_{char,yes} = \frac{1}{3} \times ((216 - 208)^2 + (69 - 208)^2 + (60 - 208)^2) = 5114$

$\mu_{len,yes} = \frac{1}{3} \times (5.68 + 4.78 + 3.16) = 4.54$

$\delta^2_{len,yes} = \frac{1}{3} \times ((5.68 - 4.54)^2 + (4.78 - 4.54)^2 + (3.16 - 4.54)^2) = 1.0872$

Model for wrong anwers:

$\mu_{char,no} = \frac{1}{2} \times (302 + 393) = 347.5$

$\sigma^2_{char,no} = \frac{1}{2} \times ((302 - 347.5)^2 + (393 - 347.5)^2) = 2070.5$

$\mu_{len,no} = \frac{1}{2} \times (2.31 + 4.20) = 3.255$

$\delta^2_{len,no} = \frac{1}{2} \times ((2.31 - 3.255)^2 + (4.20 - 3.255)^2) = 0.8930$

Inferring:

$$P(Char_{242}|yes) = \frac{1}{\sqrt{2\pi\sigma^2_{char,yes}}} exp - \frac{1}{2}\left(\frac{\left(Char_{242} - \mu_{char,yes}\right)^2}{\sigma^2_{char,yes}}\right) = 1.1526 \times 10^{-3}$$

$$P(Len_{4.56}|yes) = \frac{1}{\sqrt{2\pi\sigma^2_{len,yes}}} exp - \frac{1}{2}\left(\frac{\left(Len_{4.56} - \mu_{len,yes}\right)^2}{\sigma^2_{len,yes}}\right) = 0.3825$$

$$P(Char_{242}|no) = \frac{1}{\sqrt{2\pi\sigma_{char,no}^2}} exp - \frac{1}{2}\left(\frac{\left(Char_{242} - \mu_{char,no}\right)^2}{\sigma_{char,no}^2}\right) = 5.9644 \times 10^{-4}$$

$$P(Len_{4.56}|no) = \frac{1}{\sqrt{2\pi\sigma_{len,no}^2}} exp - \frac{1}{2}\left(\frac{\left(Len_{4.56} - \mu_{len,no}\right)^2}{\sigma_{len,no}^2}\right) = 0.1627$$

$$P(Answer|yes) = P(Char_{242}|yes) \times P(Len_{4.56}|yes) = 4.4091 \times 10^{-5}$$

$$P(Answer|no) = P(Char_{242}|no) \times P(Len_{4.56}|no) = 9.7038 \times 10^{-5}$$

$$P(yes|Answer) = \frac{P(Answer|yes) \times P(yes)}{P(Answer|yes) \times P(yes) + P(Answer|no) \times P(no)} = 0.8720$$

So this answer is likely (87.2%) to be correct.