Q1.

(a)

(1) False

K-NN is a instanced based learning model, which does not "train".

(2) False.

Precision : How accurate it is among all instances classified as "fatigue".

Recall : How many "fatigue" are correctly classified.

To detect "fatigue" as much as possible, we should use Recall.

(3) False.

The linear regression model is not complex enough and will be under-fitted.

It should have high bias and low variance.

Q1.

(b).

(1) Yes, since we have unlimited amont of data and logistic regression optimizes $P(y|x)$ directly. It is possible to achieve perfect training accuracy.

(2) The logistic regression is under-fitted. standardizing the data set may improve a little bit but not too much. Because the main problem is it does not learn the data very well.

(c)

(1) Embedded approaches build a model to find out the best feature, such as regression with regularization

Filter methods calculted the relativity between features and the label.

(2) For small data set, with small number of features. such that is not enough to train a model, I prefer use filtering strategy

Q1.(d)

(1) Naive Bayes,

$$P(X_1, X_2, X_3, Y) = P(Y)\,P(X_1|Y)\,P(X_2|Y)\,P(X_3|Y).$$

Perceptron:

can't compute probability since it is not probabilistic. It use parameters to directly calculte "Y" using active function.

(e)

(1) The model is under fitting.

(2) There are not enough number of hidden layers.

(3) ① add more hidden layers.

② add more unit for each hidden layer.

(g)

(1).(2) Yes, increase weights.

(h) (1) decrease D,

(2) increase p.

Q2.

(a)

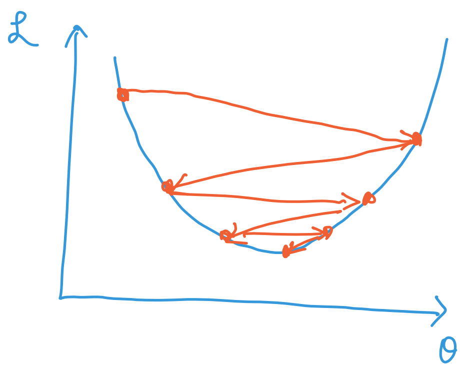
$$m_k \leftarrow m_k - \eta \frac{dL}{\partial m_k}$$

$$\frac{\partial L}{\partial m_k} = -\sum 2(x_i - m_k)$$

$$\Rightarrow m_k \leftarrow m_k + \eta \sum_{x_i \in C_k} 2(x_i - m_k)$$

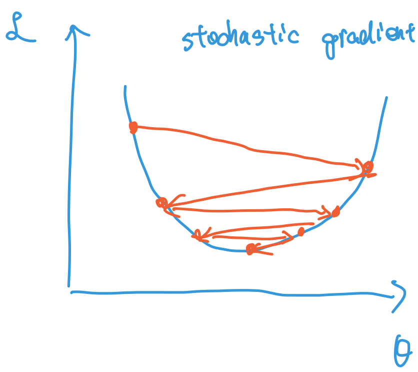
(b)

Gradient descent




Each arrow is for one batch

stochastic gradient descent.




Each arrow is for one instance

GD calculate $\theta$ for all instances and update once after iterating all instances.

SGD calculate $\theta$ for all instances and update once per instances

Q2.

(c).

If there are two many instances in the data set, SGD is better.

(d)

(a) If "point" moves between "left" and "right", then the learning rate is too big.

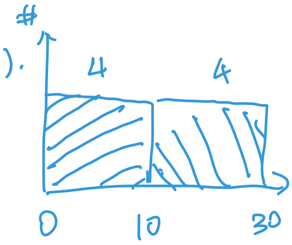
If "point" moves along one side but needs to update too many times, it is too small.

(b) too big: may not converge

too small: running time to long and may converge to local minimum.

Q3.

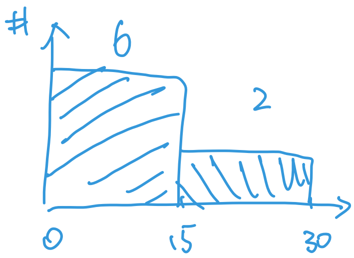
(a).




low: $\{2, 6, 7, 8\}$ if salary $\le 10$

high: $\{1, 3, 4, 5\}$ if salary $> 10$.

(b)




low: $\{1, 2, 5, 6, 7, 8\}$, Salary $\le 15$

High: $\{3, 4\}$, $15 < $ Salary $< 30$

(c)

$P(Y=1) = \frac{5}{8}$, $P(Y=0) = \frac{3}{8}$

$P(\text{low}) = \frac{1}{2}$, $P(\text{high}) = \frac{1}{2}$

| | low | high | total |
|---|---|---|---|
| y=1 | 1 | 4 | 5 |
| y=0 | 3 | 0 | 3. |
| total | 4 | 4 | 8 |

$P(\text{low}, 1) = \frac{1}{8}$,

$P(\text{low}, 0) = \frac{3}{8}$

$P(\text{high}, 1) = \frac{4}{8}$

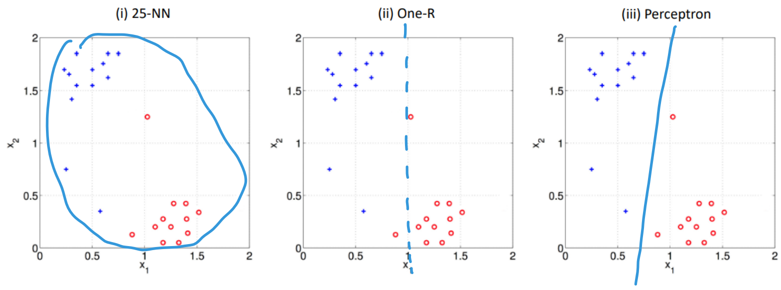
$P(\text{high}, 0) = 0$

$$MI = \frac{1}{8}\log_2 \frac{\frac{1}{8}}{\frac{1}{2}\times\frac{5}{8}} + \frac{3}{8}\log_2 \frac{\frac{3}{8}}{\frac{1}{2}\times\frac{3}{8}} + \frac{4}{8}\log_2 \frac{\frac{4}{8}}{\frac{1}{2}\times\frac{3}{8}} + 0 = 0.92$$

(d) For the scenario that the data doesn't have high variance, I prefer equal-frequency bin.

# Q4.

your answer sheet, rather than annotating the exercise sheet directly if that is easier.)




(i) with majority voting, all instance will be classified as "red".

(ii) If $x < 1$, then "blue".
If $x > 1$, the "red".

(iii) Above a straight line : "blue",
Below a straight line : "red".

Q5.

(a)

C, since C is the nearest point to the boundary, which means it is the least confident one.

(b) A, since A is farest point from the boundary line, which means it is the most confident one.

Q6.

(a) $a_1^{(1)} = x_1 \cdot 1 + x_2 \times 1 + x_0 \times 1$

$= 0.7 + 0.5 + 1 = 2.2$

$a_2^{(1)} = x_1 \cdot 1 + x_2 \times 1 + x_0 \times 1$

$= 2.2$

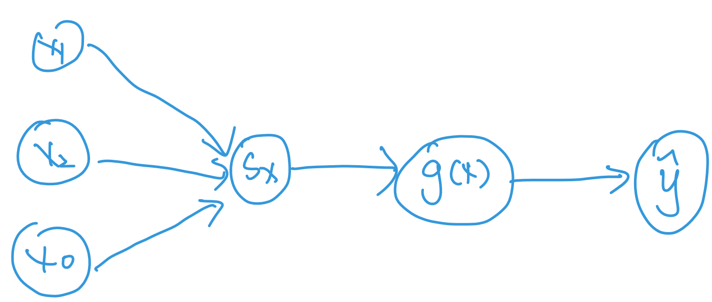
$a^{(2)} = 2.2 \times 1 + 2.2 \times 1 = 4.4$

output $= \dfrac{1}{1+e^{-2.2}} = 0.9879 > 0.5 = 1$

(b) The parameters are not updated while training.

Apply backpropagation.

(c)



Q6. (d)

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla L$$

$$\nabla L = \frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial g} \frac{\partial g}{\partial \theta}$$

$$= \frac{\partial L}{\partial g} \frac{\partial g}{\partial s} \frac{\partial s}{\partial \theta}$$

$$= (Y - a^{(2)}) \frac{1}{1+e^{-sx}} \left(1 - \frac{1}{1+e^{-sx}}\right) \theta^{(k-1)}$$

$$\Rightarrow \theta^{(k+1)} = \theta^{(k)} - \eta (Y - a^{(2)}) \frac{1}{1+e^{-sx}} \left(1 - \frac{1}{1+e^{-sx}}\right) \theta^{(k-1)}$$

Q7.

(a)

$$H(Y) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}$$

$$= 1.59$$

$$H(X=1) = 1 \times \log_2(1) = 0$$

$$H(X=2) = -\left(\frac{1}{2}\times\log_2\frac{1}{2} + \frac{1}{2}\times\log\frac{1}{2}\right) = 1$$

$$H(X=3) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

$$H(X=4) = \log_2(1) = 0$$

$$H(X=5) = \log_2(1) = 0$$

$$H(X=6) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

MeanInfo

$$= \frac{1}{14}\times 0 + \frac{2}{14}\times 1 + \frac{2}{14}\times 1 + \frac{1}{14}\times 0$$

$$+ \frac{1}{14}\times 0 + \frac{2}{14}\times 1 = \frac{6}{14}$$

$$GR = \frac{1.1564}{2.02} = 0.57$$

$$IG = H(Y) - \text{MeanInfo} = 1.1564$$

$$SI = \left(\frac{1}{14}\times\log_2\frac{1}{14} + \frac{2}{14}\times\log\frac{2}{14} + \frac{2}{14}\times\log\frac{2}{14} + \frac{1}{14}\log\frac{1}{14} + \frac{1}{14}\log\frac{1}{14} + \frac{2}{14}\log\frac{2}{14}\right) = 2.02$$

(b)

low = { 3, 4, 5 } $H(low) = -(\frac{1}{3} \times \log_2 \frac{1}{3} + \frac{2}{3} \times \log_2 \frac{2}{3})$

$= 0.92$

med = { 0, 7, 9 } $H(med) = -(\frac{1}{3} \times \log_2 \frac{1}{3} + \frac{2}{3} \times \log_2 \frac{2}{3})$

$= 0.92$

high = { 1, 2, 8 }

$H(high) = -(\frac{2}{3} \times \log_2 \frac{2}{3} + \frac{1}{2} \times \log_2 \frac{1}{3})$

$= 0.92$

Mean Info $= \frac{1}{3} \times 0.92 + \frac{1}{3} \times 0.92 + \frac{1}{3} \times 0.92$

$= 2.76$

IG $= 1.59 - 2.76 = -1.17.$

SI $= -(\frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{3} \log_2 \frac{1}{3})$

$= 1.59$

GR $= \frac{-1.17}{1.59} = -0.74.$

(c) Information gain is biased to the feature with more values.

Q8.

(a)

DT: proper, since have labelled data.

CNB: proper, supervised

LR: not proper, data are not linear.

50-NN: proper, supervised.

k-means: not proper, unsupervised.

(b)

b.1. input = 4.

output = 1.

b.2. Yes, since it's non-linear.

b.3. softmax.

b.4. backpropagation. update weights for each instance.

(c). using weighted f-score.

(d)

(e) gender bias.

data is inbalanced to "like"

Doesn't use gender as a feature.

resample dataset using up-sample or down sample to balance distribution.