Question 1. Logistic Regression

①. training set: $D = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(N)}, t^{(N)})\}$

input: $x^{(i)} = (x_1^{(i)} \ldots x_D^{(i)})$

$t^{(i)} \in \{0, 1\}$.

assume: $p(t=1 \mid x^{(i)}, w, w_0) = \sigma(w^T x^{(i)} + w_0) = \dfrac{1}{1+\exp(-\sum_{d=1}^{D} w_d x_d^{(i)} - w_0)}$

and data is i.i.d.

Gaussian Prior: $P(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$

Multivariate Normal Distribution: $p(w) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$

(where k is dimension of data set (features in this case)

$\mu = 0 \quad \Sigma = \alpha^{-1}$)

$= \dfrac{1}{(2\pi)^{\frac{k}{2}} \alpha^{(-\frac{k}{2})}} \exp(-\frac{\alpha}{2} x^2)$ $\quad$ x is weights in this case

$= \dfrac{1}{(2\pi)^{\frac{k}{2}} \alpha^{(-\frac{k}{2})}} \exp(w^T I w)^{-\frac{\alpha}{2}}$

loss function: $L(w) = \prod_{i=1}^{N} P(t^{(i)} \mid x^{(i)}) \cdot P(w)$

$= \prod_{i=1}^{N} P(t=1 \mid x^{(i)})^{t^{(i)}} \cdot P(t=0 \mid x^{(i)})^{(1-t^{(i)})} \cdot \dfrac{1}{(2\pi)^{\frac{k}{2}} \alpha^{(-\frac{k}{2})}} \exp(w^T I w)^{-\frac{\alpha}{2}}$

$-\log L(w) = -\sum_{i=1}^{N} t^{(i)} \log(P(t=1 \mid x^{(i)})) - \sum_{i=1}^{N} (1-t^{(i)}) \log(P(t=0 \mid x^{(i)})) - \sum_{i=1}^{N} (\log(\exp(w^T I w)^{-\frac{\alpha}{2}} + \frac{k}{2}(\log(\frac{\alpha}{2\pi}))$

where. $P(t=1 \mid x^{(i)}) = \dfrac{1}{1+\exp(-z)}$ $\quad$ $P(t=0 \mid x^{(i)}) = \dfrac{\exp(-z)}{1+\exp(-z)}$

$= -\sum_{i=1}^{N} t^{(i)} \log(\frac{1}{1+\exp(-z)}) - \sum_{i=1}^{N} (1-t^{(i)}) \log(\frac{\exp(-z)}{1+\exp(-z)}) + \sum_{i=1}^{N} (\frac{\alpha}{2}(w^T I w) - \frac{k}{2} \log\frac{\alpha}{2\pi})$

$= \sum_{i=1}^{N} t^{(i)} \cancel{\log(\frac{1}{1+\exp(-z)})} - \sum_{i=1}^{N} [(-z - \log(1+\exp(-z)) + t^{(i)} z + t^{(i)} \cancel{\log(1+\exp(-z))}) + \frac{\alpha}{2}(w^T I w) - \frac{k}{2} \log\frac{\alpha}{2\pi}$

$= \sum_{i=1}^{N} z + \log(1+\exp(-z)) - t^{(i)} z + \frac{\alpha}{2}(w^T I w) - \frac{k}{2} \log\frac{\alpha}{2\pi}$

Loss function $= \sum_{i=1}^{N} \log(1+\exp(-z)) + z(1-t^{(i)}) + \frac{\alpha}{2}(w^T I w) - \frac{k}{2} \log\frac{\alpha}{2\pi}$

②. Gradient:

$\dfrac{\partial loss}{\partial w_j} = \sum_{i=1}^{N} \dfrac{\exp(-z)}{1+\exp(-z)} (-x^{(i)}) + x^{(i)}(1-t^{(i)}) + \alpha w_i = \sum_{i=1}^{N} x^{(i)} (\frac{1}{1+\exp(-z)} - t^{(i)}) + \alpha w_i$

$\dfrac{\partial Loss}{\partial w_j} = \sum_{i=1}^{N} [x_j^{(i)} (\frac{1}{1+\exp(-z)} - t^{(i)})] + \alpha w_j$

$\dfrac{\partial loss}{\partial w_0} = \sum_{i=1}^{N} \dfrac{\exp(-z)}{1+\exp(-z)} (-1) + (1-t^{(i)}) \quad = \sum_{i=1}^{N} (\frac{1}{1+\exp(-z)} - t^{(i)})$

$\dfrac{\partial loss}{\partial w_0} = \sum_{i=1}^{N} (\frac{1}{1+\exp(-z)} - t^{(i)})$

③ Pseudocode:

λ: learning rate n: number of iterations k: number of features

w^i: ith iteration w.

for i in range n:

$$w_0^{i+1} = w_0^i - \lambda \frac{\partial loss}{\partial w_0}$$

for j in range k:

$$w_j^{i+1} = w_j^i - \lambda \frac{\partial loss}{\partial w_j}.$$

Question 2: Decision Tree:

①. overcooked pasta.

$\frac{3}{5}$ Yes → $\frac{1}{3}$ Yes, $\frac{2}{3}$ No

$\frac{2}{5}$ No → 1 Yes, 0 No.

$$IG(Y|X) = H(Y) - H(Y|X) = H(Y) - \Sigma P(x) H(Y|X=x)$$

$$H(Y) = -\sum_{x \in X} P(x) \log_2 P(x)$$

$$= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.97$$

$$IG_{overcooked} = 0.97 - \frac{3}{5} \times \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}\left(-\log_2 1 - 0\log_2 0\right)$$

$$= 0.97 - 0.55 = 0.42.$$

rude waiter:

$\frac{1}{5}$ Yes → 1 Yes, 0 No

$\frac{4}{5}$ No → $\frac{1}{2}$ Yes, $\frac{1}{2}$ No

$$IG_{rude\ waiter} = 0.97 - \frac{4}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= 0.97 - 0.8 = 0.17$$

waiting time.

$\frac{3}{5}$ Yes → $\frac{2}{3}$ Yes, $\frac{1}{3}$ No

$\frac{2}{5}$ No → $\frac{1}{2}$ Yes, $\frac{1}{2}$ No.

$$IG_{waiting\ time} = 0.97 - \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\right.$$

$$= 0.97 - 0.55 - 0.4 = 0.02.$$

⇒ pick overcooked pasta as first node: and split on overcooked pasta.

waiting time

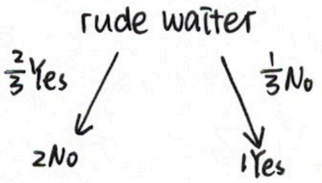
$\frac{2}{3}$ Yes → $\frac{1}{2}$ Yes, $\frac{1}{2}$ No.

$\frac{1}{3}$ No → 1 No 0 Yes

$IG_{waiting\ time | overcooked} =$

$$-\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.918 - 0.667$$

$$= 0.25$$

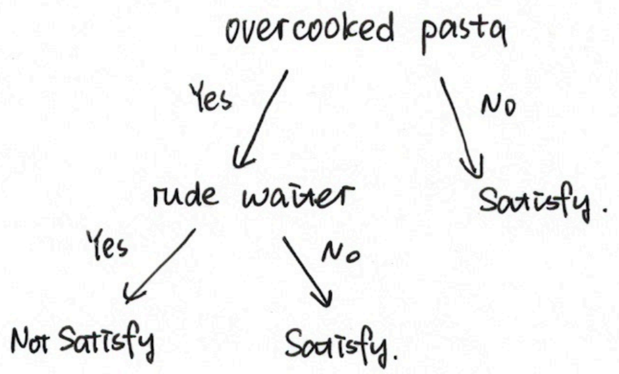


IG(rude waiter | overcooked pasta)

$= 0.918 - 0 = 0.918$

$\Rightarrow$ pick rude waiter as second node.

Construct the decision tree:



②. Prediction:

Person ID	Satisfied?
6	Yes
7	No
8	Yes.