2.

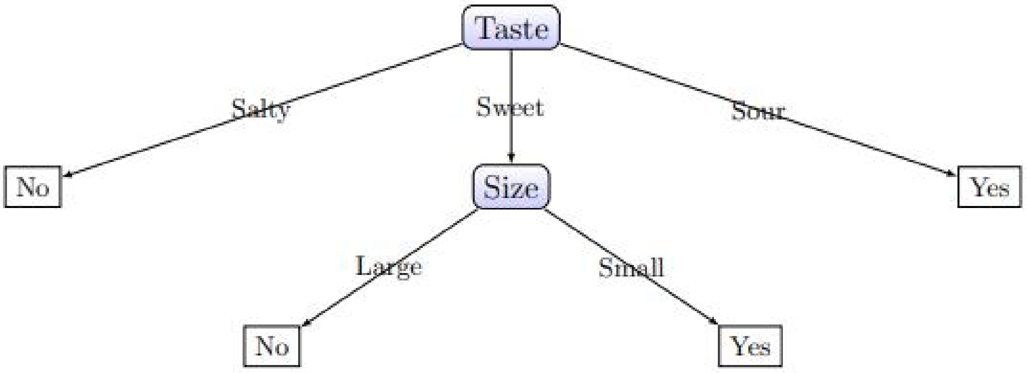
1. The initial entropy of "Appealing" is

$$H(t) = -\sum_{c\in\{\text{Yes,No}\}} p(c|t)\log_2 p(c|t) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1.$$

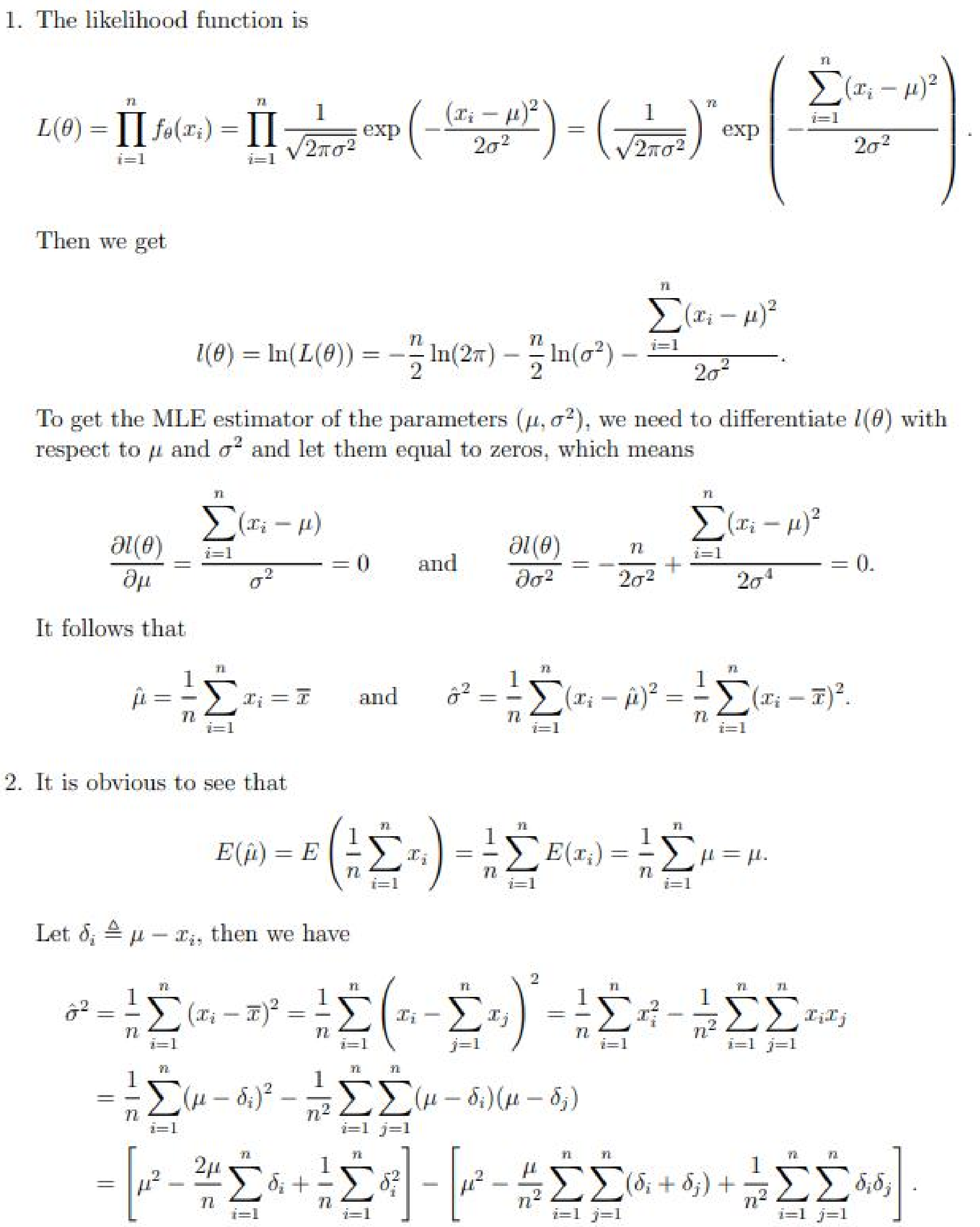
2. The information gain associated with "Taste" is

$$\begin{aligned}
\text{InfoGain}_{\text{split}} &= H(t) - \sum_{k\in\{\text{Salty,Sweet,Sour}\}} \frac{n_k}{n} H(k) \\
&= 1 - \sum_{k\in\{\text{Salty,Sweet,Sour}\}} \frac{n_k}{n}\left(-\sum_{c\in\{\text{Yes,No}\}} p(c|k)\log_2 p(c|k)\right) \\
&= 1 - \left[-\frac{3}{10}(1\log_2 1) - \frac{4}{10}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{3}{10}(1\log_2 1)\right] \\
&= \frac{3}{5}.
\end{aligned}$$

3.



3.



1. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

Then we get

$$l(\theta) = \ln(L(\theta)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}.$$

To get the MLE estimator of the parameters (μ, σ^2), we need to differentiate $l(\theta)$ with respect to μ and σ^2 and let them equal to zeros, which means

$$\frac{\partial l(\theta)}{\partial \mu} = \frac{\sum\limits_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0 \qquad \text{and} \qquad \frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0.$$

It follows that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

2. It is obvious to see that

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Let $\delta_i \triangleq \mu - x_i$, then we have

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \sum_{j=1}^{n} x_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mu-\delta_i)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mu-\delta_i)(\mu-\delta_j) \\
&= \left[\mu^2 - \frac{2\mu}{n}\sum_{i=1}^{n}\delta_i + \frac{1}{n}\sum_{i=1}^{n}\delta_i^2\right] - \left[\mu^2 - \frac{\mu}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_i+\delta_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\right].
\end{aligned}$$

Use the fact that $E(\delta_i) = 0$, $E(\delta_i^2) = \sigma^2$ and $E(\delta_i \delta_j) = 0$ (independent), we get

$$\begin{aligned}
&E(\hat{\sigma}^2) \\
&= \left[\mu^2 - \frac{2\mu}{n}\sum_{i=1}^{n} 0 + \frac{1}{n}\sum_{i=1}^{n}\sigma^2\right] - \left[\mu^2 - \frac{\mu}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(0+0) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} 0 + \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

It follows that

$$E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \sigma^2.$$

This means that $\hat{\mu}$ is an unbiased estimator of μ, but $\hat{\sigma}^2$ is a biased estimator of σ^2.

4.

Let

$$P(y_i) = \prod_{k=1}^{c} p_k^{I(y_i=k)},$$

then the likelihood function is

$$L(p_1, p_2, \ldots, p_c) = \prod_{i=1}^{n} P(y_i) = \prod_{i=1}^{n} \prod_{k=1}^{c} p_k^{I(y_i=k)}.$$

Then we get

$$l(p_1, p_2, \ldots, p_c) = \ln(L(p_1, p_2, \ldots, p_c)) = \sum_{i=1}^{n} \sum_{k=1}^{c} I(y_i = k) \ln(p_k).$$

By the Lagrangian multiplier method and the restriction $\sum_{k=1}^{c} p_k = 1$,

$$l(p_1, p_2, \ldots, p_c, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{c} I(y_i = k) \ln(p_k) - \lambda \left(\sum_{k=1}^{c} p_k - 1 \right).$$

To get the MLE estimator of the parameter p_k, we need to differentiate $l(p_1, p_2, \ldots, p_c, \lambda)$ with respect to p_k and let it equal to zero, which means

$$\frac{\partial l(p_1, p_2, \ldots, p_c, \lambda)}{\partial p_k} = \sum_{i=1}^{n} \frac{I(y_i = k)}{p_k} - \lambda = 0 \quad \Rightarrow \quad \hat{p}_k = \frac{\sum_{i=1}^{n} I(y_i = k)}{\lambda}.$$

According to the restriction $\sum_{k=1}^{c} p_k = 1$, we would get $\lambda = n$, it follows that

$$\hat{p}_k = \frac{\sum_{i=1}^{n} I(y_i = k)}{n}, \qquad k = 1, \ldots, c.$$

Similarly, let

$$P(x_i|y_i) = \prod_{s=1}^{t} \prod_{k=1}^{c} p_{sk}^{I(x_i=s, y_i=k)},$$

then the likelihood function is

$$L(p_{11}, \ldots, p_{tc}) = \prod_{i=1}^{n} P(x_i|y_i) = \prod_{i=1}^{n} \prod_{s=1}^{t} \prod_{k=1}^{c} p_{sk}^{I(x_i=s, y_i=k)}.$$

Then we get

$$l(p_{11},\ldots,p_{tc}) = \ln(L(p_{11},\ldots,p_{tc})) = \sum_{i=1}^{n}\sum_{s=1}^{t}\sum_{k=1}^{c} I(x_i = s, y_i = k)\ln(p_{sk}).$$

By the Lagrangian multiplier method and the restriction $\sum\limits_{s=1}^{t} p_{sk} = 1$,

$$l(p_{11},\ldots,p_{tc},\lambda) = \sum_{i=1}^{n}\sum_{s=1}^{t}\sum_{k=1}^{c} I(x_i = s, y_i = k)\ln(p_{sk}) - \lambda\left(\sum_{s=1}^{t} p_{sk} - 1\right).$$

To get the MLE estimator of the parameter p_{sk}, we need to differentiate $l(p_{11},\ldots,p_{tc},\lambda)$ with respect to p_{sk} and let it equal to zero, which means

$$\frac{\partial l(p_{11},\ldots,p_{tc},\lambda)}{\partial p_{sk}} = \sum_{i=1}^{n} \frac{I(x_i = s, y_i = k)}{p_{sk}} - \lambda = 0 \quad \Rightarrow \quad \hat{p}_{sk} = \frac{\sum_{i=1}^{n} I(x_i = s, y_i = k)}{\lambda}.$$

According to the restriction $\sum\limits_{s=1}^{t} p_{sk} = 1$, we would get $\lambda = \sum\limits_{i=1}^{n} I(y_i = k)$, it follows that

$$\hat{p}_{sk} = \frac{\sum_{i=1}^{n} I(x_i = s, y_i = k)}{\sum_{i=1}^{n} I(y_i = k)}, \qquad s = 1,\ldots,t, \quad k = 1,\ldots,c.$$

5.

参照《understanding machine learning》p.221, Lemma 19.1