Jonathan Nusantara jan26 Assignment #2

Problem 1 (15 points). In class we said that for a generative model (e.g., Naive Bayes), the optimal estimator will be the maximum aposteriori probability (MAP) estimator that when given a feature $\overline{X}$, outputs the label that satisfies:

$$\arg\max_{y\in\mathcal{Y}} p(y|\overline{X}).$$

The maximum likelihood (ML) estimator outputs the label that maximizes the likelihood (probability) of the observation (which is the feature $\overline{X}$):

$$\arg\max_{y\in\mathcal{Y}} p(\overline{X}|y).$$

In this problem we will see that this is the predictor with the least error probability if the underlying data is generated from the model.

We will simplify the setting by considering a binary classification task, where the labels have two possible values, say $\mathcal{Y} = \{-1, +1\}$. Suppose the model that generates the data is $p(\overline{X}, y)$, which is **known**.

1. What is the distribution of y when we observe a feature $\overline{X}$?
2. Suppose we predict the label -1 upon seeing $\overline{X}$. Show that the probability of error is $p(y = +1|\overline{X})$.
3. Use this to argue that for any $\overline{X}$ the prediction to minimize the error probability is

$$\max_{y\in\{-1,+1\}} p(y|\overline{X}).$$

This shows that the MAP estimator is the optimal estimator for the binary task. This also extends to larger $\mathcal{Y}$.

4. Show that if the distribution over the labels is uniform, namely $p(y = -1) = p(y = +1) = 0.5$, then the MAP estimator and ML estimator are the same.
5. Construct *any* generative model where the MAP and ML estimator are not the same.

① $P(y|\bar{x}) = \frac{P(x|y)P(y)}{P(\bar{x})} = \frac{P(\bar{x},y)}{P(\bar{x})} = \frac{P(\bar{x},y)}{\sum_{y'\in\mathcal{Y}} P(\bar{x},y')} = \frac{P(\bar{x},y)}{P(\bar{x},+1)+P(\bar{x},-1)}$

② Find $P(y=-1|\bar{x}) = 1 - P(y=+1|\bar{x})$

$P(-1|\bar{x}) = \frac{P(\bar{x},-1)}{P(\bar{x})} = \frac{P(\bar{x},-1)}{\sum_{y'\in\mathcal{Y}} P(\bar{x},y')} = \frac{P(\bar{x},-1)}{P(\bar{x},+1)+P(\bar{x},-1)}$

$P(+1|\bar{x}) = \frac{P(\bar{x},+1)}{P(\bar{x},+1)+P(\bar{x},-1)}$

$P(y=-1|\bar{x}) = 1 - P(y=+1|\bar{x})$

$= P(y=+1|\bar{x})$ = probability of error of $P(y=-1|\bar{x})$

$\frac{P(\bar{x},-1)}{P(\bar{x},+1)+P(\bar{x},-1)} = 1 - \frac{P(\bar{x},+1)}{P(\bar{x},+1)+P(\bar{x},-1)}$

③ From Q2, we know that:

$$P(y=-1 \mid \bar{x}) = 1 - P(y=+1 \mid \bar{x})$$

Since $y = \{+1, -1\}$, only 2 values/labels (binary).

We know that by using MAP where $\hat{y} = \max_{y=\{+1,-1\}} P(y \mid \bar{x})$, it would be the optimal estimator.

(chosen y that maximizes probability)

Using MAP, we can find the predicted $\hat{y}$ that would make $P(y' \mid \bar{x})$, (where y' is the other y in $\{+1, -1\}$), to be the smallest. → $\underbrace{P(y' \mid \bar{x}) < P(\hat{y} \mid \bar{x})}_{\text{proven and guaranteed.}}$

$P(y' \mid \bar{x})$ is error probability that is minimized

④ $P(\bar{x}, y) = P(y) \cdot P(\bar{x} \mid y) = P(y) \cdot P(\bar{x}^1 \mid y) \dots P(\bar{x}^d \mid y)$ ($P(y) = 0.5$)

ML $\hat{y} = \arg\max P(\bar{x} \mid y) = \max\left(P(\bar{x} \mid +1), P(\bar{x} \mid -1) \right)$

MAP $\hat{y} = \arg\max P(y \mid \bar{x}) = \max\left(\frac{P(\bar{x}, +1)}{P(\bar{x}, +1) + P(\bar{x}, -1)}, \frac{P(\bar{x}, -1)}{P(\bar{x}, +1) + P(\bar{x}, -1)} \right)$

$$= \max\left(\frac{0.5\, P(\bar{x} \mid +1)}{0.5\, P(\bar{x} \mid +1) + 0.5\, P(\bar{x} \mid -1)}, \frac{0.5\, P(\bar{x} \mid -1)}{0.5\, P(\bar{x} \mid +1) + 0.5\, P(\bar{x} \mid -1)} \right)$$

$$= \max\left(\frac{P(\bar{x} \mid +1)}{P(\bar{x} \mid +1) + P(\bar{x} \mid -1)}, \frac{P(\bar{x} \mid -1)}{P(\bar{x} \mid +1) + P(\bar{x} \mid -1)} \right)$$

From above, we can see how $\arg\max P(\bar{x} \mid y)$ is correlated to $\arg\max P(y \mid \bar{x})$.

The numerators, which determine the max value, is the same.

This proves that if distribution over label is equal, MAP = ML estimator

⑤ Decide on a distribution : $P(+1) = 0.7$, $P(-1) = 0.3$

use counter-example $P(\bar{x}|+1) = 0.4$ $P(\bar{x}|-1) = 0.5$

ML $\hat{y} = \arg\max P(\bar{x}|y) = \arg\max\left(P(\bar{x}|+1), P(\bar{x}|-1)\right)$

$= \arg\max\left(0.4, 0.5\right)$

ML $\hat{y} = \underline{-1}$

MAP $\hat{y} = \arg\max P(y|\bar{x}) = \arg\max\left(\frac{P(\bar{x},+1)}{P(\bar{x},+1)+P(\bar{x},-1)}, \frac{P(\bar{x},-1)}{P(\bar{x},+1)+P(\bar{x},-1)}\right)$

$P(\bar{x},y) = P(y)P(\bar{x}|y)$ $= \arg\max\left(\frac{0.7\ P(\bar{x}|+1)}{0.7P(\bar{x}|+1)+0.3P(\bar{x}|-1)}, \frac{0.3\ P(\bar{x}|-1)}{0.7P(\bar{x}|+1)+0.3P(\bar{x}|-1)}\right)$ ($P(+1)$ ↑ 0.7, $P(-1)$ ← 0.3)

$= \arg\max\left(\frac{0.7\times0.4 = .28}{0.7\times0.4+0.3\times0.5}, \frac{0.3\times0.5 = .15}{0.7\times0.4+0.3\times0.5}\right)$

MAP $\hat{y} = \underline{+1}$

From the above, we can see that the result $\hat{y}$ from ML and MAP is not similar. As a result, unlike during equal distribution as in the previous problem, when distribution is not equal, $\underline{MAP \neq ML}$

The reason is because MAP looks at prior $P(Y)$.

proving using counter example!

Problem 2. (10 points). ML vs MAP and add constant smoothing. Suppose you generate n independent coin tosses using a coin with bias μ (i.e. the probability of getting a head for each toss). Let $A(n_H, n_T)$ denote the event of getting n_H heads and $n_T = n - n_H$ tails. Show the following:

1. According to maximum likelihood principle, show that your estimate for μ should be:

$$\hat{p}_{ML} = \hat{\mu} = \frac{n_H}{n_H + n_T}.$$

Hint: Given that the bias is μ, show that:

$$p(A(n_H, n_T)|\mu) = \binom{n_H + n_T}{n_T} \mu^{n_H} (1-\mu)^{n_T}.$$

2. Consider the following generative process: First generate μ, the bias of the coin, according to the *prior* distribution $p(\mu)$. Then generate n independent coin tosses with bias μ. Show that

$$\arg\max_{\mu} p(\mu|A(n_H, n_T)) = \arg\max_{\mu} p(\mu)p(A(n_H, n_T)|\mu).$$

3. Let the prior of the bias be a *Beta* distribution, which is a distribution over $[0, 1]$

$$p(\mu) = \frac{\mu^{\alpha}(1-\mu)^{\beta}}{\int_0^1 x^{\alpha}(1-x)^{\beta} dx},$$

show that:

$$\arg\max_{\mu} p(\mu|n_H, n_T) = \frac{n_H + \alpha}{n_H + \alpha + n_T + \beta}.$$

Remark: This shows that add constant smoothing is equivalent to inducing a *Beta* distribution prior on the parameter of the generating model.

---

(1) By bernoulli, given probability of μ, if n_H successes is to be found from $n_H + n_T$ trials, the probability of seeing $(A(n_H, n_T))$ is:

$$\binom{n_H+n_T}{n_H} \mu^{n_H} (1-\mu)^{n_T} = \binom{n_H+n_T}{n_T} \mu^{n_H} (1-\mu)^{n_T}$$

Since $n = n_H + n_T$, then $\binom{n_H+n_T}{n_H} = \binom{n_H+n_T}{n_T}$, so

$$f(\mu) = \binom{n_H+n_T}{n_T} \mu^{n_H} (1-\mu)^{n_T}$$

To find max, multiply by ln, since when $f(x) > 0$, maximizing $f(x)$ is equal to maximizing $\log(f(x))$

$$\ln f(\mu) = \ln\binom{n_H+n_T}{n_T} + \ln \mu^{n_H} + \ln (1-\mu)^{n_T}$$

$$= \ln\binom{n_H+n_T}{n_T} + n_H \ln \mu + n_T \ln(1-\mu)$$

$$\ln f'(\mu) = \frac{n_H}{\mu} - \frac{n_T}{1-\mu}$$

Let $\ln f'(\mu) = 0$

$$\frac{n_H}{\mu} - \frac{n_T}{1-\mu} = 0 \rightarrow \frac{n_H}{\mu} = \frac{n_T}{1-\mu} \rightarrow n_H - n_H\mu = n_T\mu$$

$$\hat{\mu} = \frac{n_H}{n_T + n_H} \quad \text{(proven)}$$

② From Bayes rule, we know $P(y|\bar{x}) = \frac{P(\bar{x}, y)}{P(\bar{x})}$

Bayes ← $= \frac{P(\bar{x}|y)\,P(y)}{P(\bar{x})}$

since we are speaking of optimization, we only need the proportionality and can ignore normalizing constant $P(\bar{x})$

$$\frac{P(\bar{x}|y)\,P(y)}{P(\bar{x})} \propto P(\bar{x}|y)\,P(y)$$

We have bias μ and prior distribution $p(\mu)$.

$$\text{MAP } \hat{\mu} = \arg\max_{\mu} P(\mu | A(n_H, n_T)) = \arg\max_{\mu} \left(\frac{P(A(n_H, n_T), \mu)}{P(A(n_H, n_T))} \right)$$

naive bayes theorem ←

$$= \arg\max_{\mu} \left(\frac{P(\mu)\,P(A(n_H, n_T)|\mu)}{P(A(n_H, n_T))} \right)$$

denominator ignored as we find max by comparing numerator, and divisor $P(A(n_H, n_T))$ takes effect to all candidates, ←

$$= \arg\max_{\mu} \left(P(\mu)\,P(A(n_H, n_T)|\mu) \right)$$

(proven)

③ We want to use hint from Q1 $p(A(n_H, n_T)|\mu) = \binom{n_H + n_T}{n_T} \mu^{n_H}(1-\mu)^{n_T}.$

By bernoulli, given probability of μ, if n_H successes is to be found from $n_H + n_T$ trials, the probability of seeing $(A(n_H, n_T))$ is:

$$\binom{n_H+n_T}{n_H} \mu^{n_H}(1-\mu)^{n_T} = \binom{n_H+n_T}{n_T} \mu^{n_H}(1-\mu)^{n_T}$$

Since $n = n_H + n_T$, then $\binom{n_H+n_T}{n_H} = \binom{n_H+n_T}{n_T}$

In Q2, we've proven $\arg\max P(\mu|A(n_H,n_T)) = \arg\max\left(P(\mu)P(A(n_H,n_T)|\mu)\right)$

with $P(\mu) = \frac{\mu^\alpha(1-\mu)^\beta}{\int_0^1 x^\alpha(1-x)^\beta dx}$, and using equation from Q1 and 2

so, $\arg\max(\mu|n_H,n_T) = \arg\max(P(\mu)P(n_H,n_T|\mu)$

$$= \arg\max \left(\frac{\mu^\alpha(1-\mu)^\beta}{\int_0^1 x^\alpha(1-x)^\beta dx}\right)\binom{n_H+n_T}{n_T}\mu^{n_H}(1-\mu)^{n_T}$$

$$= \arg\max \left(\frac{1}{\int_0^1 x^\alpha(1-x)^\beta dx}\right)\binom{n_H+n_T}{n_T}\mu^{n_H+\alpha}(1-\mu)^{n_T+\beta}$$

ln both sides →

$$= \arg\max \ln\left(\frac{\binom{n_H+n_T}{n_T}}{\int_0^1 x^\alpha(1-x)^\beta dx}\right) + \ln\left(\mu^{n_H+\alpha}\right) + \ln\left((1-\mu)^{n_T+\beta}\right)$$

$$= \arg\max \ln\left(\frac{\binom{n_H+n_T}{n_T}}{\int_0^1 x^\alpha(1-x)^\beta dx}\right) + (n_H+\alpha)\ln(\mu) + (n_T+\beta)\ln(1-\mu)$$

Derivative and set = 0 to find max value.

constant $\ln\left(\frac{\binom{n_H+n_T}{n_T}}{\int_0^1 x^{\alpha}(1-x)^{\beta}dx}\right) = 0$ after derivative

$$0 = \frac{n_H+\alpha}{\mu} - \frac{n_T+\beta}{1-\mu}$$

$$\frac{n_T+\beta}{1-\mu} = \frac{n_H+\alpha}{\mu}$$

$$\mu(n_T+\beta) = (1-\mu)(n_H+\alpha)$$

$$\mu(n_T+\beta) = (n_H+\alpha) - \mu(n_H+\alpha)$$

$$\mu(n_H+\alpha+n_T+\beta) = n_H+\alpha$$

$$\underset{\mu}{\arg\max}\left(\mu \mid n_H, n_T\right) = \hat{\mu} = \frac{n_H+\alpha}{n_H+\alpha+n_T+\beta}$$ (proven)

Problem 3. (10 points). Consider the Tennis data set. In this problem, you don't need the smoothing constant when estimating the prior probabilities of the labels.

1. For $\beta = 1$ (smoothing constant), write down the probabilities of all the features conditioned on the labels. The total number of probabilities you need to compute should not be more than twenty.
2. What are the prior probabilities of the labels?
3. For a new label $(Overcast, Hot, High, Strong)$, what does the Naive Bayes classifier predict for $\beta = 0, \beta = 1$, and for $\beta \to \infty$?

Playing Tennis Example

Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

(1) With $\beta = 1$

$P(O=\text{Sunny} \mid \text{Yes}) = \frac{2+1}{9+3} = \frac{3}{12}$

$P(O=\text{Overcast} \mid \text{Yes}) = \frac{4+1}{9+3} = \frac{5}{12}$

$P(O=\text{Rain} \mid \text{Yes}) = \frac{3+1}{9+3} = \frac{4}{12}$

$P(T=\text{Hot} \mid \text{Yes}) = \frac{2+1}{9+3} = \frac{3}{12}$

$P(T=\text{Mild} \mid \text{Yes}) = \frac{4+1}{9+3} = \frac{5}{12}$

$P(T=\text{Cool} \mid \text{Yes}) = \frac{3+1}{9+3} = \frac{4}{12}$

$P(H=\text{High} \mid \text{Yes}) = \frac{3+1}{9+2} = \frac{4}{11}$

$P(H=\text{Normal} \mid \text{Yes}) = \frac{6+1}{9+2} = \frac{7}{11}$

$P(W=\text{Strong} \mid \text{Yes}) = \frac{3+1}{9+2} = \frac{4}{11}$

$P(W=\text{Weak} \mid \text{Yes}) = \frac{6+1}{9+2} = \frac{7}{11}$

$P(O=\text{Sunny} \mid \text{No}) = \frac{3+1}{5+3} = \frac{4}{8}$

$P(O=\text{Overcast} \mid \text{No}) = \frac{0+1}{5+3} = \frac{1}{8}$

$P(O=\text{Rain} \mid \text{No}) = \frac{2+1}{5+3} = \frac{3}{8}$

$P(T=\text{Hot} \mid \text{No}) = \frac{2+1}{5+3} = \frac{3}{8}$

$P(T=\text{Mild} \mid \text{No}) = \frac{2+1}{5+3} = \frac{3}{8}$

$P(T=\text{Cool} \mid \text{No}) = \frac{1+1}{5+3} = \frac{2}{8}$

$P(H=\text{High} \mid \text{No}) = \frac{4+1}{5+2} = \frac{5}{7}$

$P(H=\text{Normal} \mid \text{No}) = \frac{1+1}{5+2} = \frac{2}{7}$

$P(W=\text{Strong} \mid \text{No}) = \frac{3+1}{5+2} = \frac{4}{7}$

$P(W=\text{Weak} \mid \text{No}) = \frac{2+1}{5+2} = \frac{3}{7}$

② Prior probabilities of label = $P(Y)$, where $Y = \{Yes, No\}$

$P(Yes) = \frac{9}{14}$, $P(No) = \frac{5}{14}$

③ Find $P(Overcast, Hot, High, Strong)$ for $\beta = 0, 1, \to\infty$

$P(\bar{x}, y) = P(y) \cdot P(\bar{x}|y) = P(y) \cdot P(\bar{x}^1|y) \dots P(\bar{x}^d|y)$

Let $\bar{x}$ be (Overcast, Hot, High, Strong)

For $\beta = 0$,

$P(\bar{x}, Yes) = \frac{9}{14} \times \frac{4}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{4}{567}$

$P(\bar{x}, No) = \frac{5}{14} \times \frac{0}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} = 0$

For $\beta = 1$,

$P(\bar{x}, Yes) = \frac{9}{14} \times \frac{5}{12} \times \frac{3}{12} \times \frac{4}{11} \times \frac{4}{11} = \frac{15}{1694} \approx 0.0089$

$P(\bar{x}, No) = \frac{5}{14} \times \frac{1}{8} \times \frac{3}{8} \times \frac{5}{7} \times \frac{4}{7} = \frac{75}{10976} \approx 0.0068$

For $\beta = \infty$,

$P(\bar{x}, Yes) = \frac{9}{14} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} = \frac{9}{14}$

$P(\bar{x}, No) = \frac{5}{14} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} \times \frac{\infty}{\infty} = \frac{5}{14}$

So, for $\beta = 0, 1$, and $\to\infty$

$P(\bar{x}, Yes) > P(\bar{x}, No)$

So, $\hat{y} = Yes$

$\hat{y}$ that maximizes probability

Problem 4. (15 points). Recall the Gaussian distribution with mean μ, and variance σ^2. The density is given by:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Given n independent samples $X_1, \ldots, X_n$ from the same Gaussian distribution with unknown mean, and variance, let μ_{ML}, and σ^2_{ML} denote the maximum likelihood estimates of mean and variance.

Hint: (1) What is the joint density $p(x_1, \ldots, x_n)$ for i.i.d Gaussian random variables $X_1, \ldots, X_n$? (2) If $f(x) > 0$, maximizing $f(x)$ is equivalent as maximizing $\log[f(x)]$.

1. Show that
$$\mu_{ML} = \frac{\sum_{i=1}^n X_i}{n}.$$ -> estimator not mean
2. Show that
$$\sigma^2_{ML} = \frac{1}{n}\left(\sum_{i=1}^n (X_i - \mu_{ML})^2\right).$$
3. What is the expectation and variance of μ_{ML}?

① joint density probability -> $p(x_1)\,p(x_2)\cdots p(x_n)$ for $i = 1 \ldots n$

If multiply by ln -> ln(joint prob) -> $\sum_{i=1}^n \ln(p(x_i))$

Since maximizing $f(x)$ is equivalent to maximizing $\log(f(x))$

$$\ln(p_{\mu,\sigma^2}(x)) = \sum_{i=1}^n \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right) = \sum_{i=1}^n \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^n \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^n \left(\ln(1) - \ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^n \left(-\ln(\sigma) - \frac{1}{2}\ln\left(\frac{1}{2}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial\mu}\ln(p_{\mu,\sigma^2}(x)) = \sum_{i=1}^n \left(-\frac{1}{2\sigma^2}\frac{\partial}{\partial\mu}(x_i-\mu)^2\right)$$

$$= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$

To find max ->
$$\sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0$$

$$\sum_{i=1}^n x_i = \sum_{i=1}^n \mu$$

$$\sum_{i=1}^n x_i = n\mu$$

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

② $\sum_{i=1}^{n} \ln\left(p_{\mu,\sigma^2}(x)\right) = \sum_{i=1}^{n} \left(-\ln(\sigma) - \frac{1}{2}\ln\left(\frac{1}{2}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$\frac{\partial}{\partial\sigma} \ln\left(p_{\mu,\sigma^2}(x)\right) = \sum_{i=1}^{n} \left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right)$ -> set to 0

$$\sum_{i=1}^{n} \left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right) = 0$$

$$\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = \sum_{i=1}^{n} \frac{1}{\sigma}$$

$$\sum_{i=1}^{n} (x_i-\mu)^2 = \sum_{i=1}^{n} \sigma^2$$

$$\sum_{i=1}^{n} \sigma^2 = \sum_{i=1}^{n} (x_i-\mu)^2$$

$$n\sigma^2 = \sum_{i=1}^{n} (x_i-\mu)^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} (x_i-\mu)^2\right)$$ (proven)

③ From ①, we've found the $\mu_{mL} = \frac{1}{n}\sum_{i=1}^{n} X_i$

↑ estimator of the mean

$$E[\mu_{mL}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \quad \rightarrow E[aX] = a\,E[X]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \quad \rightarrow E[X_i] = \mu$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \left(\frac{1}{n}\right)(n)(\mu)$$

$$= \underline{\mu}$$

$$V[\mu_{mL}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \quad \rightarrow Var[aX] = a^2\,Var[X]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] \quad \rightarrow V[X_i] = \sigma^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= \frac{1}{n^2}(n)\sigma^2$$

$$= \underline{\frac{\sigma^2}{n}}$$