UNSW SYDNEY

Machine Learning and Data Mining

COMP9417 Final Exam (Question 4)

z5286124 Jinghan Wang

(a)

(i).

$$L(\mu_1, \mu_2, \ldots \mu_n, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{\exp\left(-\frac{(X_i-\mu_i)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \quad \frac{\exp\left(-\frac{(Y_i-\mu_i)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$$

$$= \frac{\prod_{i=1}^{n} \exp\left(\frac{-(X_i-\mu_i)^2-(Y_i-\mu_i)^2}{2\sigma^2}\right)}{\left(\sqrt{2\pi\sigma^2}\right)^{2n}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n} \exp\left(\frac{-\sum_{i=1}^{n}(X_i-\mu_i)^2-\sum_{i=1}^{n}(Y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n} \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\mu_i)^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{i=1}^{n}(Y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_i)^2\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu_i)^2\right)$$

(ii). According to the (i).

$$L(\mu_1,\mu_2,\dots\mu_n,\sigma^2)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_i)^2\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2\right)$$

$$l(\mu_1,\mu_2,\dots\mu_n,\sigma^2)=\log\left(L(\mu_1,\mu_2,\dots\mu_n,\sigma^2)\right)$$

$$=\sum_{i=1}^{n}\left(-\log(2\pi\sigma^2)-\frac{(x_i-\mu_i)^2+(y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$\frac{\partial l}{\partial \mu_i}=\frac{x_i-\mu_i}{\sigma^2}+\frac{y_i-\mu_i}{\sigma^2}=0$$

$$x_i+y_i=2\mu_i \Rightarrow \mu_i=\frac{x_i+y_i}{2}$$

MLE for μ_i is: $\hat{\mu}=\frac{x_i+y_i}{2}$

(iii) According to (ii) prove,

$$\hat{\mu}_i = \frac{x_i + y_i}{2}$$

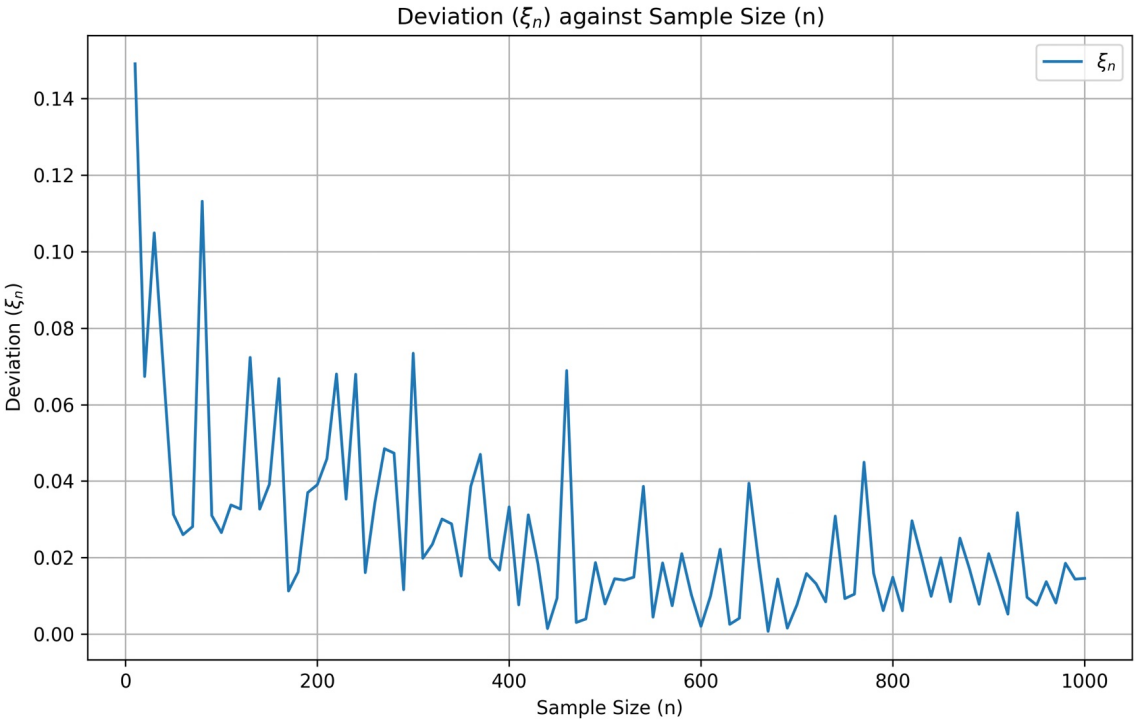
$$g(\sigma^2) = L\left(\frac{x_1 + x_2}{2}, \cdots, \frac{x_n + y_n}{2}, \sigma^2\right)$$

$$g(\sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(\frac{-(x_i - \frac{x_i+y_i}{2})^2 - (y_i - \frac{x_i+y_i}{2})^2}{2\sigma^2}\right)$$

$$\log(g(\sigma^2)) = \sum_{i=1}^{n}\left(-\log(2\pi\sigma^2) - \frac{(x_i - \frac{x_i+y_i}{2})^2 + (y_i - \frac{x_i+y_i}{2})^2}{2\sigma^2}\right)$$

$$\frac{\partial \log(g(\sigma^2))}{\partial \sigma^2} = 0$$

(iv).



From the plot, we observe that as the sample size increases, the deviation tends to decrease. This suggests that the estimator becomes closer to the true value as we use more data, which is a desirable property for an estimator. The estimator is consistent, as the sample size increases, its estimate converges to the true parameter value.

```python
def compute_sigma2_hat(X, Y, mu_i):
    n = len(X)
    total_sum = sum((X[i] - mu_i) ** 2 + (Y[i] - mu_i) ** 2 for i in range(n))
    return total_sum / (2 * n)


def q4_a_iv():
    mu_i = 1
    sigma2_true = 0.5
    sigma_true = np.sqrt(sigma2_true)
    sample_sizes = np.arange( start: 10,  *args: 1001, 10)
    xi_values = []

    for n in sample_sizes:
        X = np.random.normal(mu_i, sigma_true, n)
        Y = np.random.normal(mu_i, sigma_true, n)
        sigma2_hat = compute_sigma2_hat(X, Y, mu_i)
        xi_n = abs(sigma2_hat - sigma2_true)
        xi_values.append(xi_n)

    plt.figure(figsize=(10, 6))
    plt.plot( *args: sample_sizes, xi_values, label=r'$\xi_n$')
    plt.xlabel('Sample Size (n)')
    plt.ylabel(r'Deviation ($\xi_n$)')
    plt.title(r'Deviation ($\xi_n$) against Sample Size (n)')
    plt.legend()
    plt.grid(True)
    plt.savefig( *args: 'q4_a_iv.png', dpi=300)
    plt.show()
```

(v) -

Each observation and comes from a normal distribution with its own mean. This is quite different from the traditional problem where all observations come from a distribution with a common mean. This individual difference introduces a significant amount of variability into the model, making the estimation of more challenging.

The model is in some sense non-identifiable because different combinations of may lead to the same likelihood for. This may render the MLE estimate of unstable at particularly small sample sizes.

(b).

(i). $H(Y) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$

Total = 10

Y = 1 : 5

Y = 2 : 5

$H(Y) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) = 1.0$

(ii).

$$G(A) = H(Y) - \sum_{v \in X} \left(\frac{D_v}{D} \times H(D_v) \right)$$

When $X = 0$, $D_0 = 3$ $H(D_0) = 0$.

When $X = 1$ $D_1 = 4$ $E(D_1) = 1.0$

When $X = 2$ $D_2 = 3$ $E(D_2) = 0$

$$G(A) = 1 - \left[\frac{3}{10} \cdot 0 + \frac{4}{10} \times 1 + \frac{3}{10} \times 0 \right]$$

$$= 1 - 0.4$$

$$= 0.6$$

(iii).

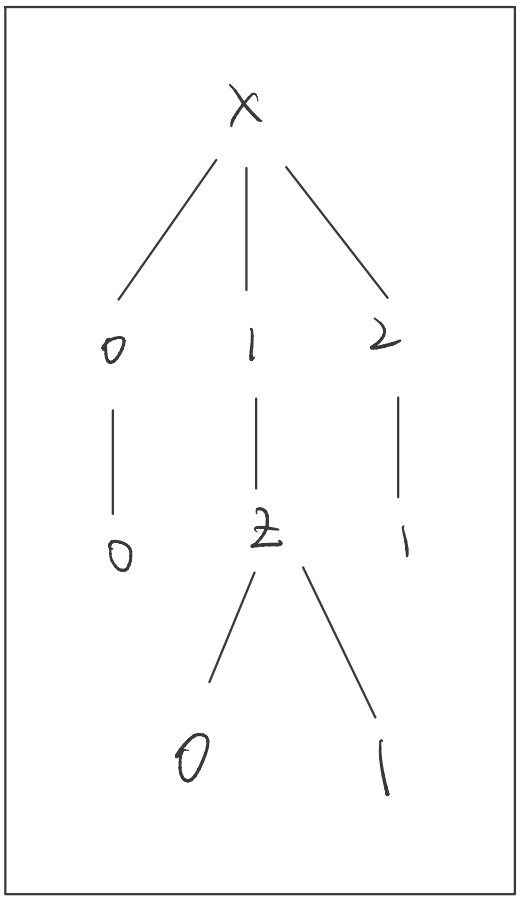
Set root node : X

For X=0 : All instance have Y=0, create leaf node Y=0

For X=2 All instance have Y=1, create leaf node Y=0

For X=1 : Exist Y=0 and Y=1 calcaton W and Z's info gain

$G(W) = 0$ $G(Z) = 1$



For X=0, prediction : Y=0

For X=1 and Z=0

prediction : Y=1

For X=1 and Z=1

prediction : Y=0

For X=2, prediction Y=1

W not provide information gain.

[C).

(i). True. For any dataset of 4 samples and any possible arrangement of labels on these samples, there exists at least one classifier in that can achieve an error of zero on these points.

(ii). False. The VC dimension being 4 for hypothesis class means that can shatter any dataset of size 4. However, it doesn't guarantee that cannot shatter datasets of size larger than 4, just that it can't shatter all datasets of size larger than 4.

(iii). False. Consider the hypothesis class of linear classifiers in a 2D space. This class has a VC dimension of 3, yet there are infinitely many possible linear classifiers in the 2D space.

(d).

I ⇒ P4

II ⇒ P2

III ⇒ P3

IV ⇒ P1