Q1)

Theory-based Questions

A → 200
B → 200
400

Problem 1: Decision Trees (12pts)

Consider a binary dataset with 400 examples, where half of them belong to class A and the rest belong to class B. Next, consider two decision stumps (i.e. trees with depth 1) $\mathcal{T}_1$ and $\mathcal{T}_2$, each with two children. For $\mathcal{T}_1$, the left child has 150 examples in class A and 50 examples in class B. For $\mathcal{T}_2$, the left child has 0 examples in class A and 100 examples in class B. (You can infer the number of examples in the right child using the total number of examples.)

1.1 (6 pts) In class, we discussed entropy and Gini impurity as measures of uncertainty at a leaf. Another possible metric is the classification error at the leaf, assuming that the prediction at the leaf is the majority class among all examples that belong to that leaf. For each leaf of $\mathcal{T}_1$ and $\mathcal{T}_2$, compute the entropy (base e), Gini impurity and classification error. You can either exactly express the final numbers in terms of fractions and logarithms, or round them to two decimal places.

1.2 (6 pts) Compare the quality of $\mathcal{T}_1$ and $\mathcal{T}_2$ (that is, the two different splits of the root) based on conditional entropy (base e), weighted Gini impurity, and total classification error. Intuitively, which of $\mathcal{T}_1$ or $\mathcal{T}_2$ appears to be a better split to you (there may not necessarily be one correct answer to this)? Based on your conditional entropy, Gini impurity, and classification error calculations, which of the metrics appear to be more suitable choices to decide which variable to split on?

Problem 1

1.1. T_1

Left Child: { class A: 150, class B: 50, total: 200 }

Entropy: $H = -\sum_i p_i \cdot \ln(p_i) = -p(A)\cdot\ln(p(A)) - p(B)\cdot\ln(p(B))$
$= -\left(\frac{150}{200}\right)\cdot\ln\left(\frac{150}{200}\right) - \left(\frac{50}{200}\right)\cdot\ln\left(\frac{50}{200}\right) = 0.56$

Gini impurity: $Gini = 1 - \sum_i p_i^2 = 1 - \left(\frac{150}{200}\right)^2 - \left(\frac{50}{200}\right)^2 = 0.375 \approx 0.38$

Classification error $= 1 - \frac{150}{200} = 0.25$

Right Child: { class A: 50, class B: 150, total: 200 }

Entropy: $H = -p(A)\cdot\ln(p(A)) - p(B)\cdot\ln(p(B))$
$= -\left(\frac{50}{200}\right)\cdot\ln\left(\frac{50}{200}\right) - \left(\frac{150}{200}\right)\cdot\ln\left(\frac{150}{200}\right) = 0.56$

Gini impurity: $Gini = 1 - \sum_i p_i^2 = 1 - \left(\frac{50}{200}\right)^2 - \left(\frac{150}{200}\right)^2 = 0.375 \approx 0.38$

Classification error $= 1 - \frac{150}{200} = 0.25$

Side Note:

classification error: $1 - \max(p_A, p_B)$

I_2

Left Child:
- class A: 0
- class B: 100
- total: 100

Entropy: $H = -P(A)\cdot\ln(P(A)) - P(B)\cdot\ln(P(B))$
$= 0 - \left(\frac{100}{100}\right)\cdot\ln\left(\frac{100}{100}\right) = 0$

Gini impurity: $Gini = 1 - \sum_i P_i^2 = 1 - 0 - \left(\frac{100}{100}\right)^2 = 0$

Classification error $= 1 - \frac{100}{100} = 0$

Right Child:
- class A: 200
- class B: 100
- total: 300

Entropy: $H = -P(A)\cdot\ln(P(A)) - P(B)\cdot\ln(P(B))$
$= -\left(\frac{200}{300}\right)\cdot\ln\left(\frac{200}{300}\right) - \left(\frac{100}{300}\right)\cdot\ln\left(\frac{100}{300}\right) = 0.64$

Gini impurity: $1 - \sum_i P_i^2 = 1 - \left(\frac{200}{300}\right)^2 - \left(\frac{100}{300}\right)^2 \approx 0.44$

Classification error $= 1 - \frac{200}{300} = 0.33$

1.2. T_1

Conditional entropy = $\frac{200}{400} \cdot 0.56 + \frac{200}{400} \cdot 0.56 = 0.56$

Weighted Gini impurity = $\frac{200}{400} \cdot 0.38 + \frac{200}{400} \cdot 0.38 = 0.38$

Total classification error = $\frac{200}{400} \cdot 0.25 + \frac{200}{400} \cdot 0.25 = 0.25$

T_2

Conditional entropy = $\frac{100}{400} \cdot 0 + \frac{300}{400} \cdot 0.64 = 0.48$

Weighted Gini impurity = $\frac{100}{400} \cdot 0 + \frac{300}{400} \cdot 0.44 = 0.33$

↳ gini impurity right

Total classification error = $\frac{100}{400} \cdot 0 + \frac{300}{400} \cdot 0.33 = 0.25$

↳ classification erro right

Intuitively, T_2 seems to have a better split. Although the total classification errors for both trees are the same, T_2 has less entropy and less Gini impurity, making it a more suitable choice.

Q2)

Problem 2: Gaussian Mixture Model and EM (10pts + 5 pts bonus)

In class, we applied EM to learn Gaussian Mixture Models (GMMs) and showed the M-Step without proof. Now, it is time that you prove it.

Consider a GMM with the following PDF of $\mathbf{x}_i$:

$$p(\mathbf{x}_i) = \sum_{j=1}^{k} \pi_j N(\mathbf{x}_i \mid \mu_j, \Sigma_j) = \sum_{j=1}^{k} \frac{\pi_j}{(\sqrt{2\pi})^d |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mu_j)\right)$$

where k is the number of Gaussian components, d is dimension of a data point $\mathbf{x}_i$ and N is the usual Gaussian pdf ($|\Sigma|$ in the pdf denotes the determinant of matrix Σ). This GMM has k tuples of model parameters $\{(\mu_j, \Sigma_j, \pi_j)\}_{j=1}^{k}$, where the parameters represent the mean vector, covariance matrix, and component weight of the j-th Gaussian component. For simplicity, we further assume that all components are isotropic Gaussian, i.e., $\Sigma_j = \sigma_j^2 I$. *

2.1 (10 pts) Find the MLE of *the expected complete log-likelihood*. Equivalently, find the optimal solution to the following optimization problem.

$$\begin{aligned} \underset{\pi_j, \mu_j, \Sigma_j}{\text{argmax}} \; & \sum_i \sum_j \gamma_{ij} \ln \pi_j + \sum_i \sum_j \gamma_{ij} \ln N(\mathbf{x}_i \mid \mu_j, \Sigma_j) \\ \text{s.t. } & \pi_j \geq 0 \\ & \sum_{j=1}^{k} \pi_j = 1 \end{aligned}$$

where γ_{ij} is the posterior of latent variables computed from the E-Step.

You can use the following fact: Given $a_1, \ldots, a_k \in \mathbb{R}^+$, the solution to the following optimization problem over $q_1, \ldots, q_k$:

$$\begin{aligned} \underset{q_j}{\text{argmax}} & \sum_{j=1}^{k} a_j \ln q_j, \\ \text{s.t. } & q_j \geq 0, \\ & \sum_{j=1}^{k} q_j = 1. \end{aligned}$$

is given by:

$$q_j^* = \frac{a_j}{\sum_{k'} a_{k'}}. \quad *$$

① To find $\pi_1, \pi_2, \ldots, \pi_k$ we know:

$$\underset{\pi}{\arg\max} \sum_i \sum_j \gamma_{ij} \ln \pi_j$$

s.t $\pi_j \geq 0$ $\quad \sum_{j=1}^{k} \pi_j = 1$

* we can derive $\pi_k^* = \frac{\sum_i \gamma_{ij}}{\sum_i \sum_j \gamma_{ij}} = \frac{\sum_i \gamma_{ij}}{\sum_i 1} = \frac{\sum_i \gamma}{N}$

To find μ_j & σ_j :

$$\underset{\mu_j, \Sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \ln N(x_i | \mu_j, \Sigma_j) =$$

$$\underset{\mu_j, \Sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \ln \Big[\underbrace{\frac{1}{(\sqrt{2\pi})^d |\Sigma_j|^{\frac{1}{2}}}}_{\text{using } \star} \exp\left(-\tfrac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)\right) \Big] =$$

$$\underset{\mu_j, \sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \ln \left[\frac{1}{(\sqrt{2\pi}\sigma_j)^d} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right) \right] =$$

$$\underset{\mu_j, \sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \left(-d \ln \sigma_j - \frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right)$$

get derivative w.r.t to μ_j and set equal to zero

$$\frac{\partial}{\partial \mu_j} = 0 \rightarrow \begin{cases} \frac{\partial}{\partial \mu_j}(-d \ln \sigma_j) = 0 \\ \frac{\partial}{\partial \mu_j}\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right) = \frac{1}{\sigma_j^2}(x_i - \mu_j) \end{cases}$$

Note: $\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{v\left(\frac{du}{dx}\right) - u\frac{dv}{dx}}{v^2}$

$$\frac{d}{d\mu_j}\|x_i - \mu_j\|^2 = \frac{d}{d\mu_j}\left((x_i - \mu_j)^T (x_i - \mu_j)\right)$$

$u = -(\|x_i - \mu_j\|^2)$

$v = 2\sigma_j^2$

$$= -2(x_i - \mu_j)$$

$\gamma_{ij} \rightarrow$ const

$$\frac{1}{\sigma_j^2} \sum_i \gamma_{ij}(x_i - \mu_j) = 0 \Rightarrow \mu_j^* = \frac{\sum_i \gamma_{ij} x_i}{\sum_i \gamma_{ij}}$$

now w.r to σ_j taking partial derivative and set equal to zero.

$$\sum \gamma_{ij} \left(-\frac{d}{\sigma_j} + \frac{\|x_i - \mu_j\|^2}{\sigma_j^3} \right) = 0 \rightsquigarrow (\sigma_j^*)^2 = \frac{\sum_i \gamma_{ij} \|x_i - \mu_j^*\|^2}{d \sum_i \gamma_{ij}}$$

Q2.2) Bonus

$$P(z_i = j \mid x_i) = \frac{P(x_i \mid z_i = j) P(z_i = j)}{\sum_{j=1}^{k} P(x_i \mid z_i = j) P(z_i = j)} =$$

$$\frac{\frac{\pi_j}{(\sqrt{2\pi}\sigma_j)^d} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right)}{\sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right)}$$

setting $\sigma_j = \sigma \to 0$ & $\pi_j = \frac{1}{k}$:

$$P(x_n, z_n = j) \propto \exp -\frac{1}{2\sigma_j^2} \|x_n - \mu_j\|^2$$

const terms ignored

the posterior becomes:

$$P(z_i = j \mid x_i) = \lim_{\sigma \to 0} \frac{\exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}\right)}{\sum_{j=1}^{k} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}\right)}$$

where $\begin{cases} 1, & \text{if } j = \operatorname{argmin}_j \|x_i - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$