# Quiz 2 - Introduction to Unsupervised Learning Techniques

Name *

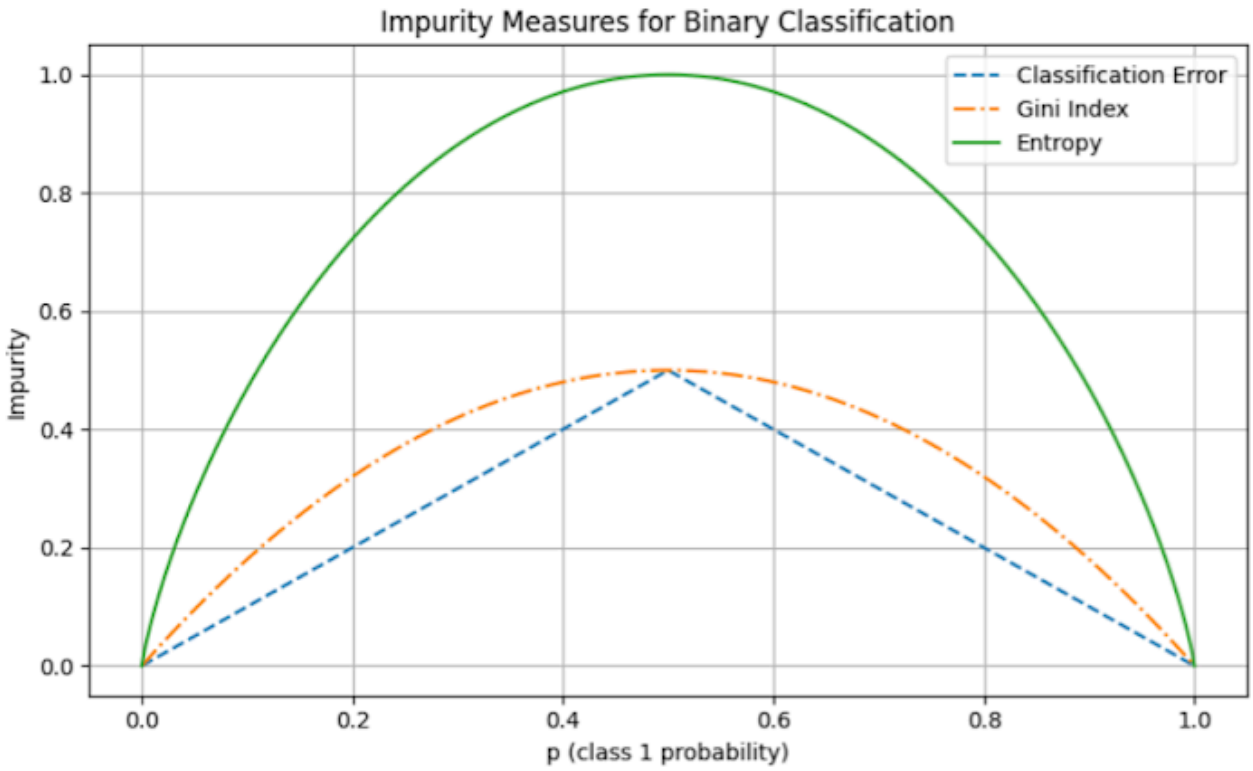
Dawei Chen

Email *

dawei.chen24@imperial.ac.uk

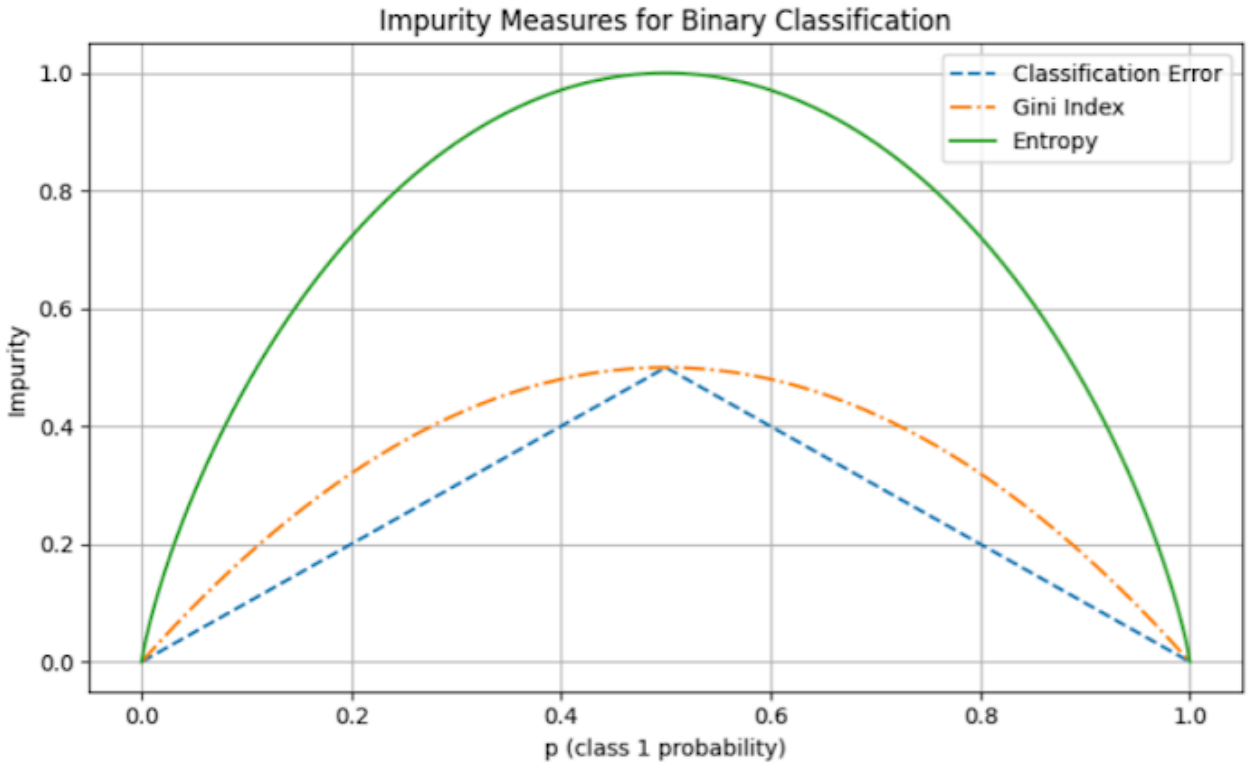
## Decison Trees

What is the entropy of a node where the label distribution is uniform (i.e., each class occurs with equal probability)? 1 point



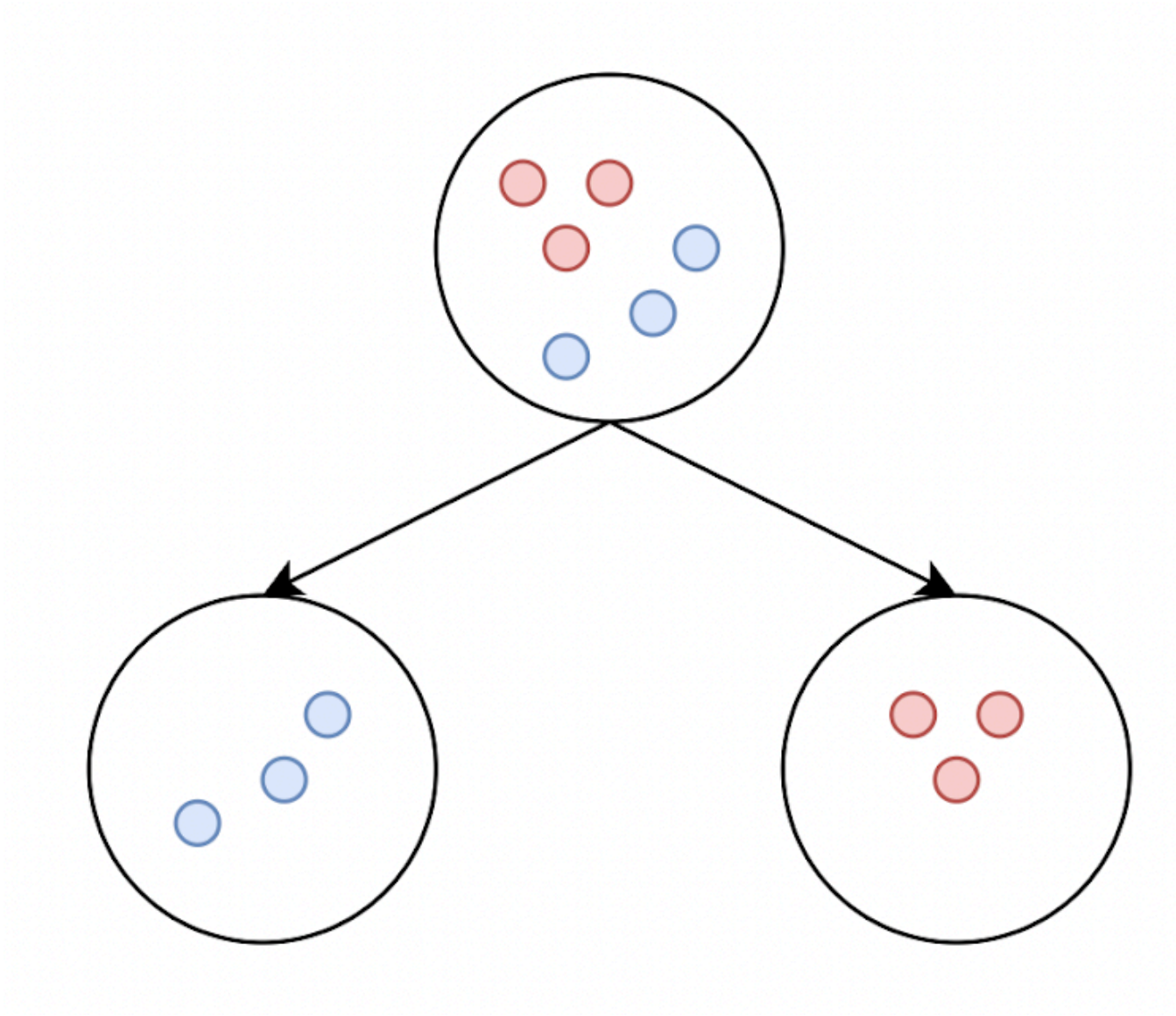

- (●) Entropy is maximum
- ( ) Entropy is minimum
- ( ) Entropy is 0.5

What is the entropy of a node that is pure (i.e., contains only one class)? 1 point




- ○ Entropy is maximum
- ◉ Entropy is minimum
- ○ Entropy is 0.5

Based on the diagram, what is the Information Gain (IG) of this split? 1 point



- ○ 0
- ○ 0.5
- ◉ 1

## Feature Importance Analysis

Which of the following is **NOT** a property of Mean Decrease Impurity (MDI)? 1 point

**Algorithm** Mean Decrease Impurity (MDI) for Random Forests

**Require:** Trained random forest model with $T$ trees
**Ensure:** Feature importance scores $\{\text{MDI}_j\}_{j=1}^{d}$

Initialize importance scores: $\text{IMP}_j = 0$ for all features $j$
**for** each tree $t = 1$ to $T$ in the forest **do**
    **for** each internal node $n$ in tree $t$ **do**
        Identify the feature $F_j$ used for splitting at node $n$
        Let $w_n$ be the proportion of samples reaching node $n$
        Calculate information gain $IG_n(j, \tau_n)$
        Update importance: $\text{IMP}_j = \text{IMP}_j + w_n \cdot IG_n(j, \tau_n)$
    **end for**
**end for**
Compute $\text{MDI}_j = \frac{1}{T} \cdot \frac{\text{IMP}_j}{\sum_{k=1}^{d} \text{IMP}_k}$ for all $j$
**return** $\{\text{MDI}_j\}_{j=1}^{d}$

- (●) It is an out-of-sample measure
- ( ) It is biased toward high-cardinality features
- ( ) It is specific to tree-based models

Which of the following is **NOT** a property of Permutation Feature Importance (PFI)? 1 point

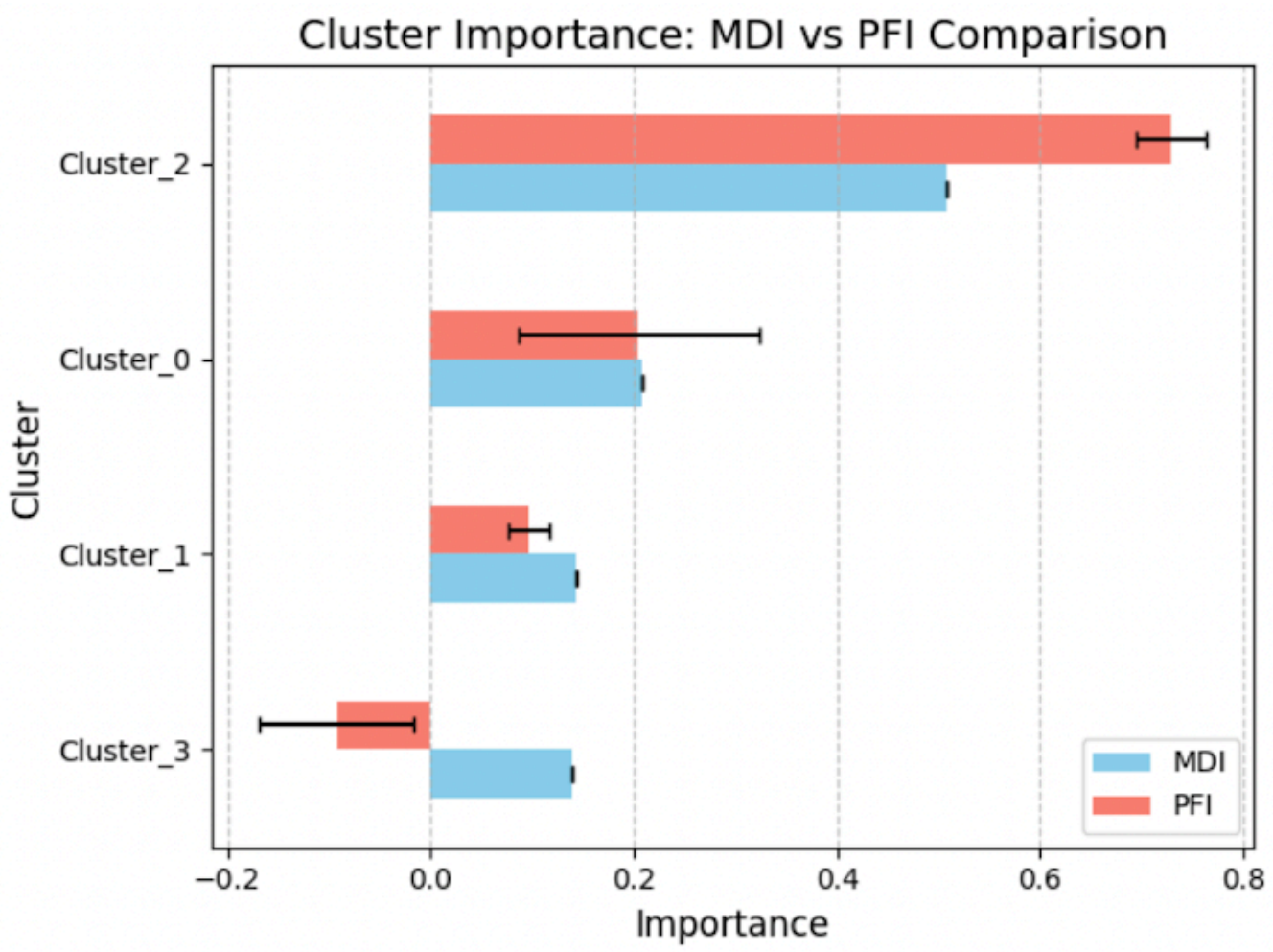
**Algorithm** Permutation Feature Importance (PFI)

**Require:** Fitted model $m$, validation data $D$, repetitions $K$
**Ensure:** Feature importance scores $\{PFI_j\}_{j=1}^{d}$ with stds $\{\sigma_j\}_{j=1}^{d}$

1: Compute reference score $s$ of model $m$ on data $D$
2: **for** each feature $F_j$ (column of $D$) **do**
3: Initialize array $scores_j$ of length $K$
4: **for** each repetition $k$ in $1, \ldots, K$ **do**
5: Randomly shuffle column $j$ of dataset $D$ to generate corrupted version $\tilde{D}_{k,j}$
6: Compute score $s_{k,j}$ of model $m$ on corrupted data $\tilde{D}_{k,j}$
7: Store in array: $scores_j[k] = s - s_{k,j}$
8: **end for**
9: Compute mean importance $PFI_j$ and standard deviation $\sigma_j$ from array $scores_j$
10: **end for**
11: **return** $\{PFI_j\}_{j=1}^{d}$ and $\{\sigma_j\}_{j=1}^{d}$

- ○ It is agnostic to the evaluation metric
- ◉ It doesn't suffer from the substitution effect when features are correlated
- ○ It is an out-of-sample measure

Based on the image from the programming session of feature importance at the cluster level, Cluster 3 has **positive MDI** and **negative PFI**. What can you deduce? 1 point




- (●) Cluster 3 contains noise features that the model mistakenly used during training
- ( ) Cluster 3 is highly informative, but PFI underestimates it due to randomness
- ( ) PFI is incorrect, because feature importance cannot be negative

## K-means and GMMs

What does the K-means algorithm aim to minimize? 1 point

**Algorithm** The K-means Algorithm

**Require:** A data set $X = \{x_1, \ldots, x_n\}$ $(x_i \in \mathbb{R}^p)$

**Ensure:** An assignment function $\Psi^*$ and the associated centroids $c_1^*, \ldots, c_K^*$.

1: Initialization: Choose $c_1, \ldots, c_K$ in $X$ at random
2: **repeat**
3: **for** $i = 1 \ldots n$ **do**
4: $\Psi(x_i) \leftarrow \arg\min_{k \in \{1,\ldots,K\}} \|x_i - c_k\|^2$
5: **end for**
6: **for** $k = 1 \ldots K$ **do**
7: $c_j \leftarrow \frac{1}{\sum_{i=1}^{n} \mathbb{1}(\Psi(x_i)=k)} \sum_{i=1}^{n} \mathbb{1}(\Psi(x_i) = k)\, x_i$
8: **end for**
9: **until** convergence
10: **return** $\Psi^*, c_1^*, \ldots, c_K^*$

- ○ The number of clusters
- ◉ The within-cluster distortion
- ○ The average silhouette score

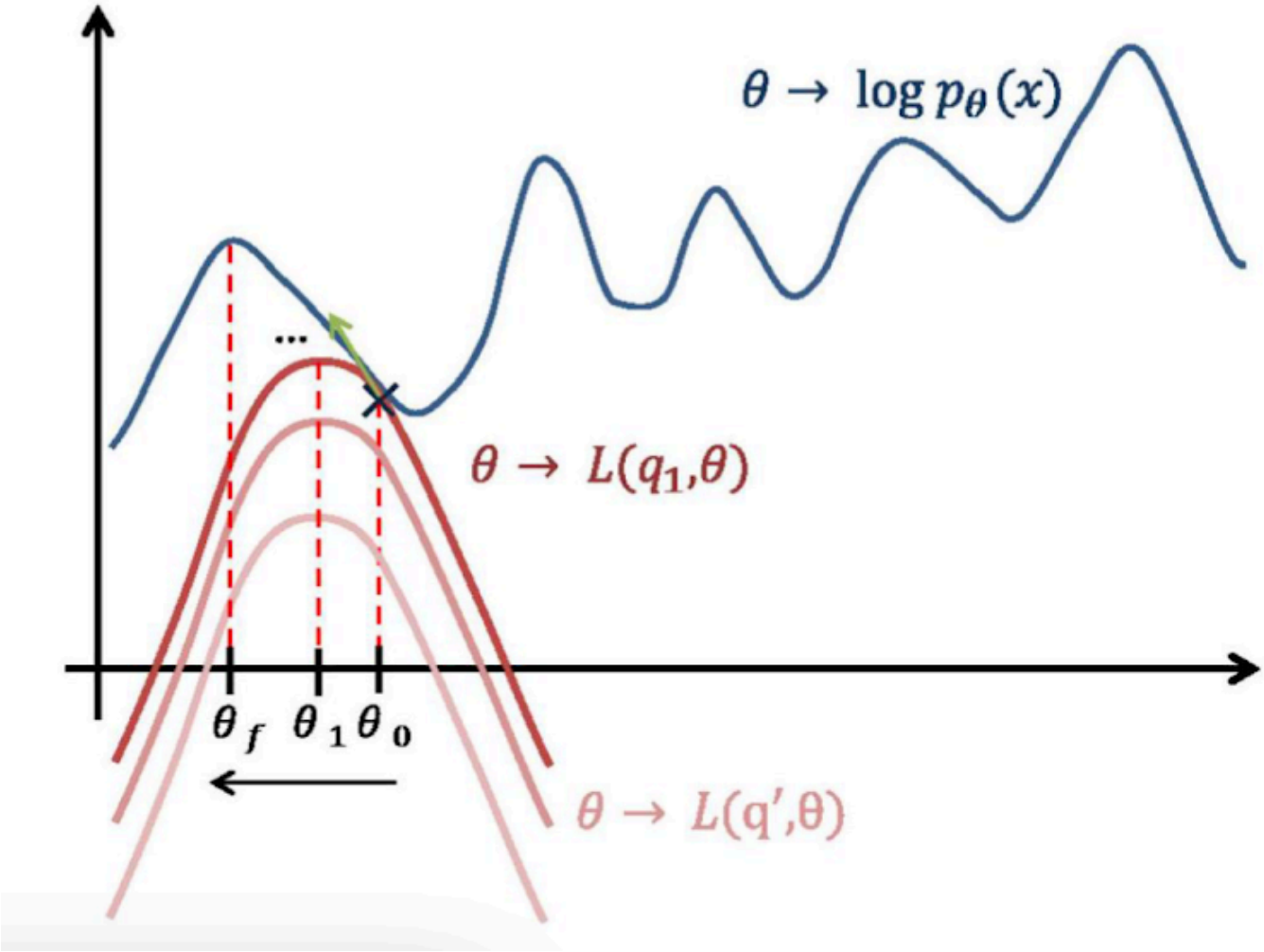
What is the expected silhouette score of a point that lies exactly between two clusters (equidistant to both)? 1 point

For each sample $i$:

- $a(i)$ = average distance to other points in the same cluster
- $b(i)$ = average distance to points in the nearest different cluster
- Silhouette value $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$

- Close to 1, indicating strong cluster assignment
- Close to 0, indicating boundary position
- Close to -1, indicating misclassification

When using the Expectation-Maximization (EM) algorithm for Gaussian Mixture Models, which statement accurately describes its theoretical guarantee? 1 point




- The EM algorithm converges to the global maximum of the marginal likelihood.
- The EM algorithm increases the evidence lower bound (denoted as L(q, θ)) at each iteration.
- It guarantees the optimal number of mixture components

In the programming session, after computing the distance matrix, applying PCA, and then using GMMs, we apply Bayes' rule to compute posterior probabilities. What do these represent? * 1 point

- ○ Hard cluster assignments, like K-means does
- ◉ Soft cluster assignments — probabilities that each point belongs to each cluster
- ○ A transformation of the distance matrix into centroids

## Questions ?

Is the pace of the course appropriate? Would a probability recap during office hours be useful?