# Expectation Maximization, Fano

***Submit a PDF of your answers to Canvas***

**1.** A mixture model assumes that distribution of the data is composed of a mixture of several base distributions:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}).$$

Let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ for $k = 1, \ldots, K$ be the mean and covariance matrix of the $K$ base distributions.

**a)** Find (and simplify) an expression for the mean of $\boldsymbol{x}$.

**b)** Find (and simplify) an expression for the covariance matrix of $\boldsymbol{x}$.

**2.** Download the associated notebook and dataset. Each row of the dataset corresponds to a vectorized MNIST image, projected onto three principle components to create a vector $\boldsymbol{x} \in \mathbb{R}^3$. Three clusters are visible as only data points corresponding to handwritten zeros, threes, and sevens are included.

Previously, using this dataset, we estimated the mean and covariance the three classes. In this problem, we will use an *off-the-shelf* expectation maximization algorithm to cluster the points.

**a)** Ignore the labels, and implement the EM algorithm for Gaussian mixture model on the dataset. Deviating from usual, you may use the `sklearn` library (in particular, `sklearn.mixture.GaussianMixture`. Set the number of mixture components to 3.

**b)** Display a 3-d scatter plot of the points and color the points based on the assigned mixture component.

**c)** Print the mean and covariance matrices produced by the EM algorithm. Do they (approximately) agree with mean and covariance you computed (in a previous activity) when you knew the class label?

**d)** What is the error rate for cluster assignment? Note that you may have to switch labels to find the label assignment with minimum error (since the algorithm doesn't know, for example, what label mixture component 1 corresponds to).

**e)** Implement the K means algorithm on the dataset. What is the error rate when using k-means? How does this compare with EM?
***Hint:*** **You may find `sklearn.cluster.KMeans` helpful.**

**f)** Imagine that you did not know how many clusters/mixture to use. Describe a reasonable way to find a reasonable number of clusters/mixture components.

**3.** This problem continues a problem that was started as an activity. Consider a joint pmf over a feature $x$ and label $y$

| | | $y$ | | |
|---|---|---|---|---|
| | $p(x,y)$ | fish | cat | dog |
| | 1 | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ |
| $x$ | 2 | $\frac{1}{16}$ | 0 | $\frac{1}{4}$ |
| | 3 | 0 | $\frac{1}{8}$ | $\frac{1}{16}$ |
| | 4 | $\frac{1}{16}$ | 0 | 0 |

**a)** Use Fano's inequality to find a lower bound on $\mathbb{P}(f(x) \neq y)$ for any classifier $f(x)$.

**b)** Specify the MAP classifier, and find the error rate. How does this compare with the lower bound?

**c)** In the derivation of Fano's inequality, we used fact that $H(E|\widehat{Y}) \leq H(E) \leq 1$ to bound $H(E|\widehat{Y})$ by 1. Instead of bounding $H(E|\widehat{Y})$ by 1, use the fact that $H(E|\widehat{Y}) \leq H(E)$, and the fact that $H(E) = (1-p_e)\log_2(\frac{1}{1-p_e}) + p_e \log_2(\frac{1}{p_e})$ where $p_e = \mathbb{P}(\widehat{Y} \neq Y)$ to find a tighter expression. Evaluate your expression for values of $p_e$ between 0 and 1 to find the best bound.