AP186 Activity 15

Expectation Maximization

Marc Jerrone R. Castro

2015-07420

Expectation Maximization (EM)

The EM algorithm provides a method for determining the maximum-likelihood estimates for model parameters (or in this case features) when data is either incomplete, has missing data points, or have inherent latent variables. The algorithm also serves as a method for determining heat maps of plausible clustering locations of data points within the feature space using the probability distribution function (pdf) [1].

Reference: [1] https://www.statisticshowto.datasciencecentral.com/em-algorithm-expectation-maximization/

EM Algorithm

The algorithm utilizes a given set of parameters of a certain class to determine the pdf of a certain cluster whereas the pdf of the cluster is given by,

$$p(\boldsymbol{x}|\Theta)=\sum_{l=1}^{M} P_l\, p_l(\boldsymbol{x}|\boldsymbol{\theta}_l)$$

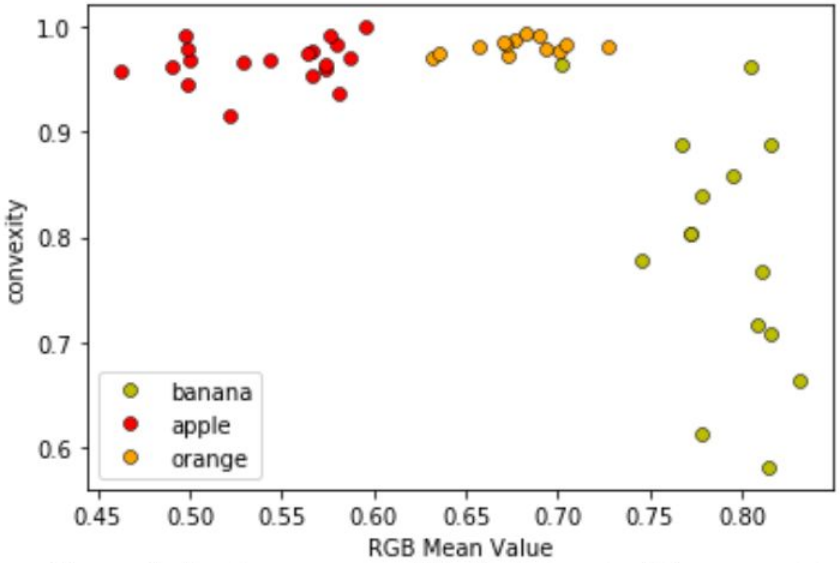


Figure 1. Feature space of fruit dataset with respect to the mean RGB pixel value and convexity of the fruit. The EM algorithm aims to determine the probability distribution of each of these classes.

- P_l is the prior probability of the l_{th} pdf,
- p_l is the probability of observing x given the parameter θ_l.
- M is the number of component pdf's

```
1 def p(x, mu, sigma):
2   return 1/(2*np.pi)**(num_dimensions/2)/la.det(sigma)**(1/2) \
3               * np.exp(-1/2 * (x - mu).T.dot(la.inv(sigma)).dot(x - mu))
```

AP186 Activity 15
Expectation Maximization

EM Algorithm

whereas the equation for the probability of a data point belonging to a specific cluster is given by,

$$P(l|x_i,\Theta)=\frac{P_l\,p_l(x_i|\theta_l)}{p(x_i|\Theta)}=\frac{P_l\,p_l(x_i|\theta_l)}{\sum_{l=1}^{M}P_l\,p_l(x_i|\theta_l)}$$

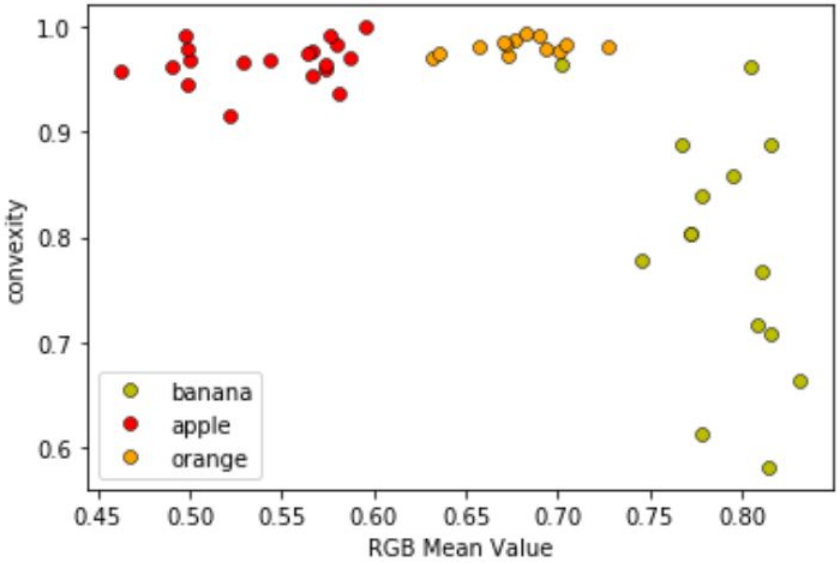


Figure 1. Feature space of fruit dataset with respect to the mean RGB pixel value and convexity of the fruit. The EM algorithm aims to determine the probability distribution of each of these classes.

```
5 def P(l, x):
6         num = P[l] * p(x, theta['mu'][l], theta['cov'][l])
7         den = 0.
8         for m in range(num_classes):
9             den += P[m] * p(x, theta['mu'][m], theta['cov'][l])
10        return num/den
```

AP186 Activity 15
Expectation Maximization

EM Algorithm

Afterwards, we assume a d-dimensional Gaussian distribution , where the component pdf is given by,

$$p_l(\boldsymbol{x}|\boldsymbol{\mu_l},\boldsymbol{\Sigma_l}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma_l}|^{1/2}} \exp\{-1/2(\boldsymbol{x}-\boldsymbol{\mu_l})^T \boldsymbol{\Sigma_l^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_l})\}$$

where μ_l is the l_{th} mean , and Σ_l is the co-variance matrix. The following equation determines or hypothesizes a probability distribution for corresponding data inputs and plots it within the feature space. At each epoch, corresponding adjustment are made via a number of update functions presented in the next slide.

EM Algorithm

The following equations are the update functions which will slowly adjust the probability distribution for each epoch and would halt at a stopping parameter whereas the run would've yielded the best approximation if the difference between the new and the old guessed parameters is either smaller or equal to the stopping parameter.

$$P_l^{new} = \frac{1}{N} \sum_{i=1}^{N} P(l|x_i, \Theta^g) \qquad \mu_l^{new} = \frac{\sum_{i=1}^{N} x_i P(l|x_i, \Theta^g)}{\sum_{i=1}^{N} P(l|x_i, \Theta^g)}$$

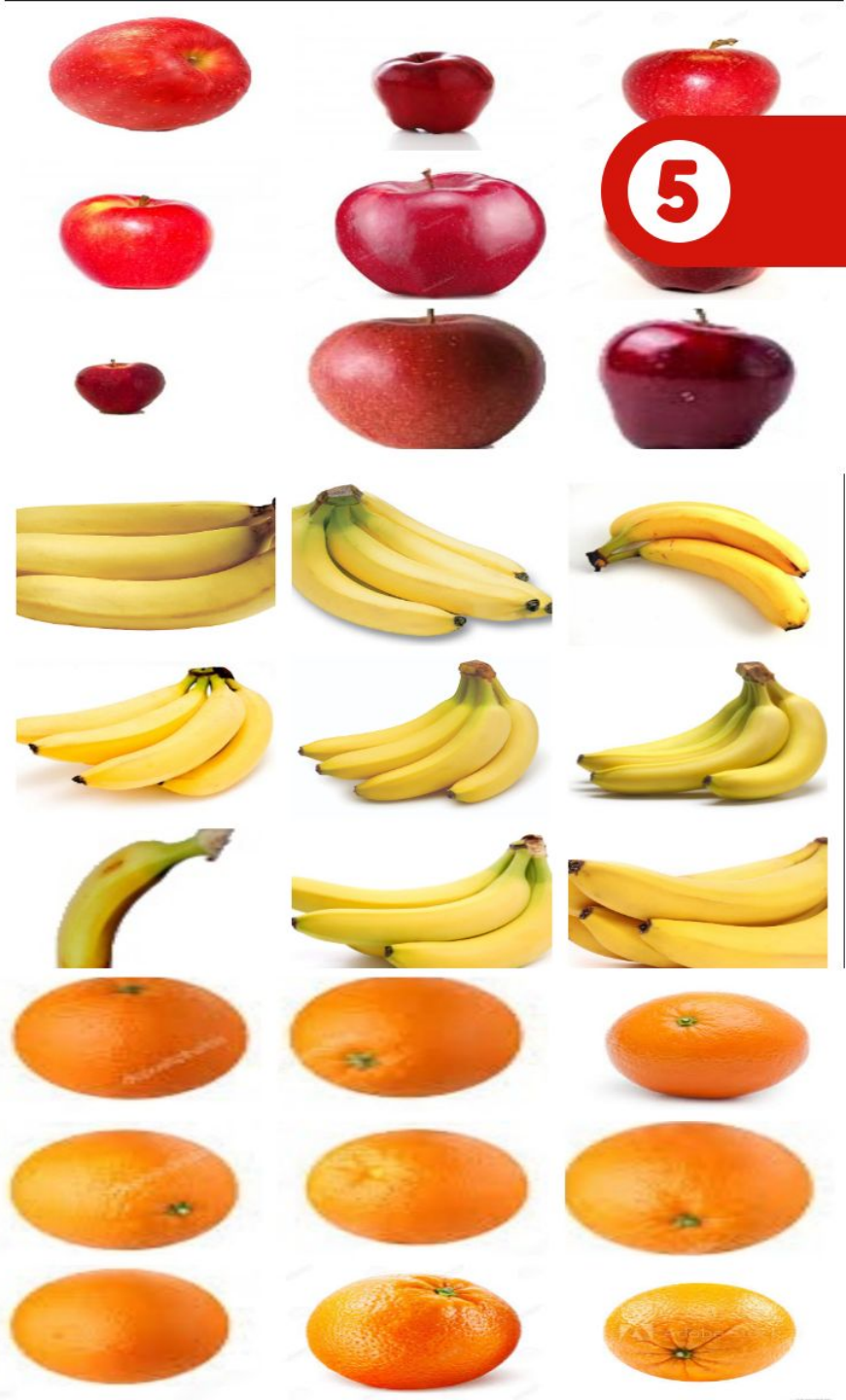
$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} P(l|x_i, \Theta^g)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} P(l|x_i, \Theta^g)}$$

AP186
Activity 15
Expectation Maximization

Data

The data used for this study is a set of images of fruits (apples, oranges, and bananas) taken from Google Images.

Using the same pre-processing steps from Activity 11 - Feature Extraction, image features such as average pixel value, concavity, convexity, and inertia ratio were extracted from each image.



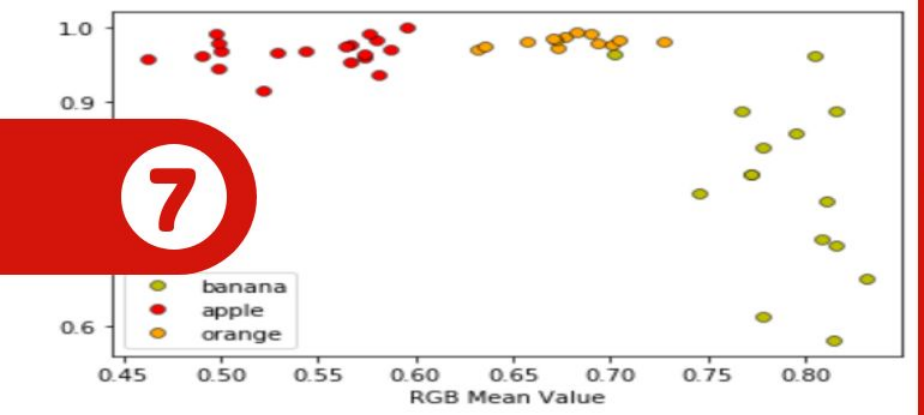
Data

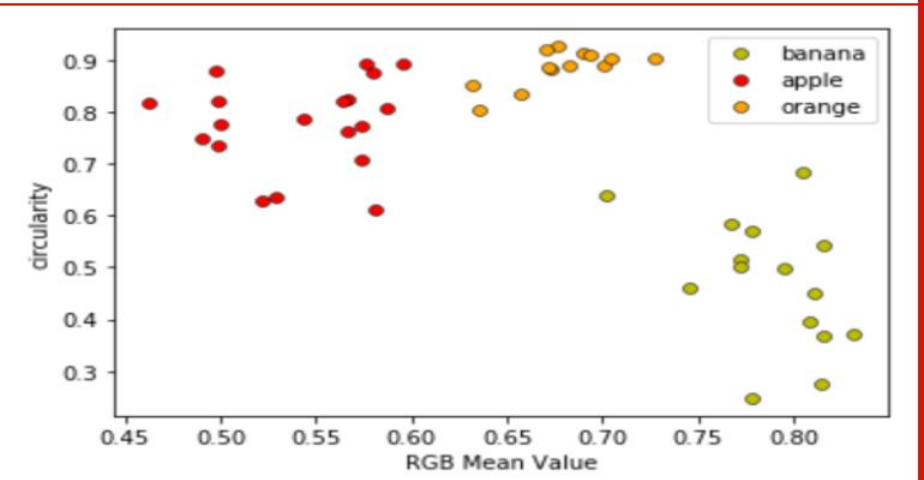
The data used for this study is a set of images of fruits (apples, oranges, and bananas) taken from Google Images.

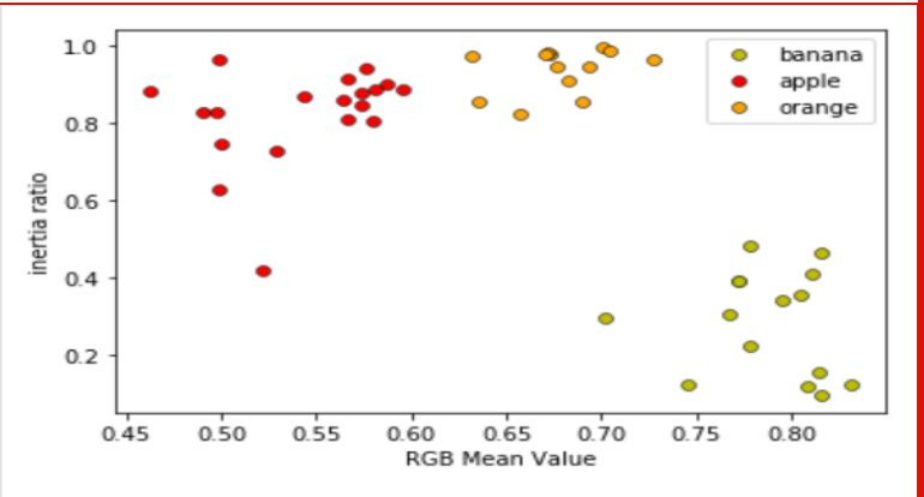
Using the same pre-processing steps from Activity 11 - Feature Extraction, image features such as average pixel value, concavity, convexity, and inertia ratio were extracted from each image.

	file name	fruit type	rgbmean	convexity	circularity	inertia ratio
0	02	red	0.595774	0.999024	0.892545	
1	37	red	0.587019	0.970451	0.807966	
2	19	red	0.500390	0.967210	0.775939	0.746720
3	08	red	0.497882	0.991004	0.878732	0.827307
4	03	red	0.529208	0.965941	0.633915	0.726095
5	25	red	0.566883	0.976699	0.824910	0.808698
6	26	red	0.498262	0.977636	0.821655	0.964853
7	09	red	0.489888	0.962301	0.749591	0.826559
8	34	red	0.564600	0.975133	0.818959	0.860007
9	38	red	0.581233	0.936779	0.611577	0.889765
10	44	red	0.544073	0.968350	0.787053	0.868931
11	49	red	0.579252	0.982881	0.875581	0.805130
12	45	red	0.573520	0.960229	0.707369	0.847519
13	47	red	0.521722	0.914428	0.628469	0.419497
14	43	red	0.566402	0.952610	0.761042	0.915711

7







FEATURE SPACES

From the extracted features, I decided to test the EM algorithm against 3 data pairs - RGB vs Centrality, RGB vs Convexity, and RGB vs Inertia Ratio.

Each data pair show distinct separations between clusters and as such - I expect the PDFs of each class to be able to detect which specific areas of the feature spaces are highly probable for specific classes.

Probability Distribution

Average pixel value vs Circularity

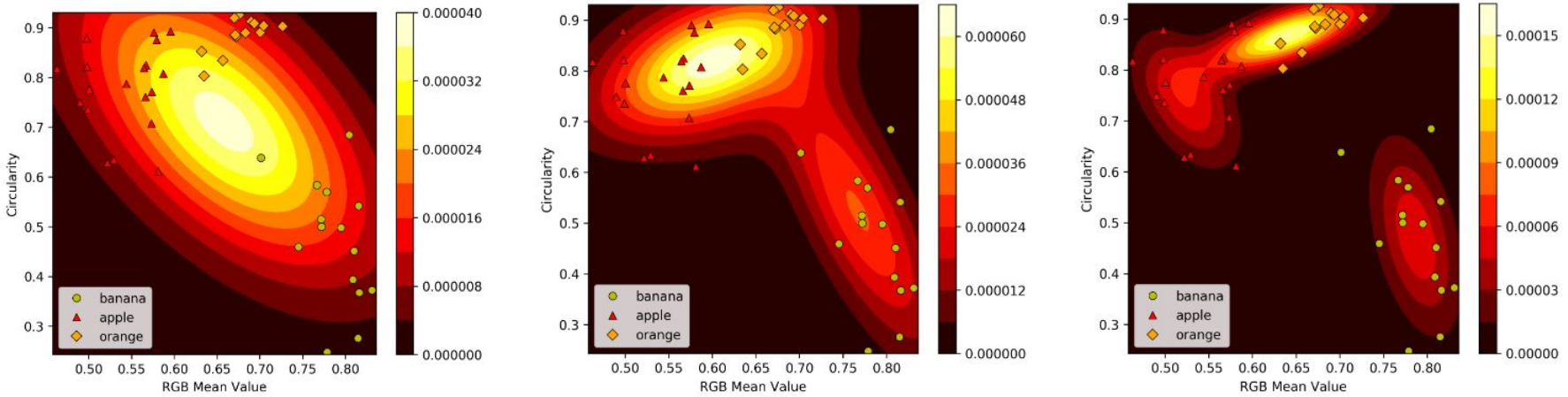


Figure 2. Estimated probability distribution via expectation maximization of the average pixel value and circularity feature space of the fruit dataset.The leftmost image shown is the initial estimate of the pdf while the succeeding images are the 8th and 40th epochs of the EM algorithm. Darker areas suggest lower probability of a particular member of that cluster being located there, while lighter areas depict higher probability. As one may observe, the initial distribution is depicted as a single large contour. However, after a number of itterations/epochs the distribution eventually splits into three distinct clusters - whereas the pdfs of apples and oranges overlap. Using visual inference, we can confirm that data points where clustered along the lighter portions of the contours while some stray data points lie outside the low probability and darker countours.

* A corresponding animation for this feature space was attached in the submission

Probability Distribution

Average pixel value vs Convexity

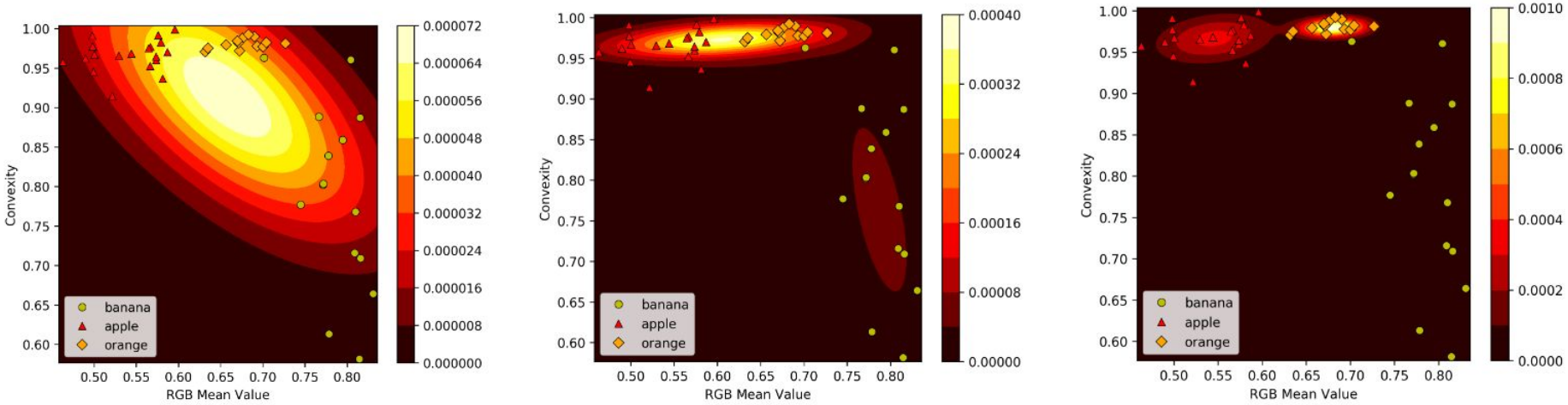


Figure 3. Estimated probability distribution via expectation maximization of the average pixel value and convexity feature space of the fruit dataset.The leftmost image shown is the initial estimate of the pdf while the succeeding images are the 7th and 62nd epochs of the EM algorithm. From observation, the final pdf was unable to depict the probability distribution of the bananas as this particular dataset has a high spread with respect to the convexity of the images. Although a distribution can be observed for the bananas from the 7th epoch up to the 46th epoch.

* A corresponding animation for this feature space was attached in the submission

Probability Distribution

Average pixel value vs Inertia Ratio

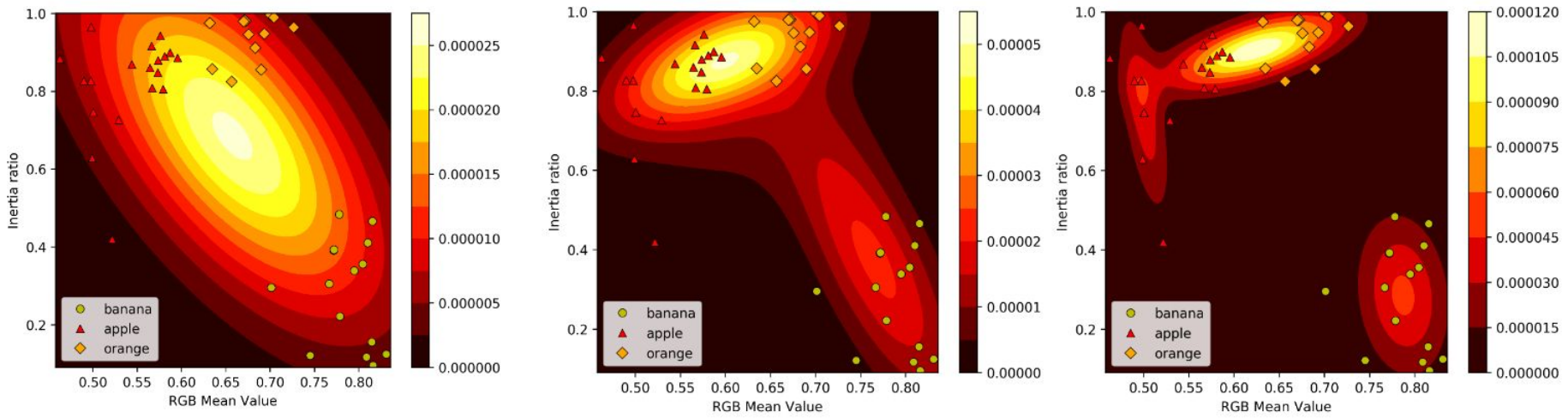


Figure 4. Estimated probability distribution via expectation maximization of the average pixel value and inertia ratio feature space of the fruit dataset.The leftmost image shown is the initial estimate of the pdf while the succeeding images are the 7th and 28th epochs of the EM algorithm. Similar to the preceeding figures, the initial distribution is depicted as a single large contour. Whereas, after several itterations would split into three distinct clusters - although an obvious overlap was observed for apples and oranges.

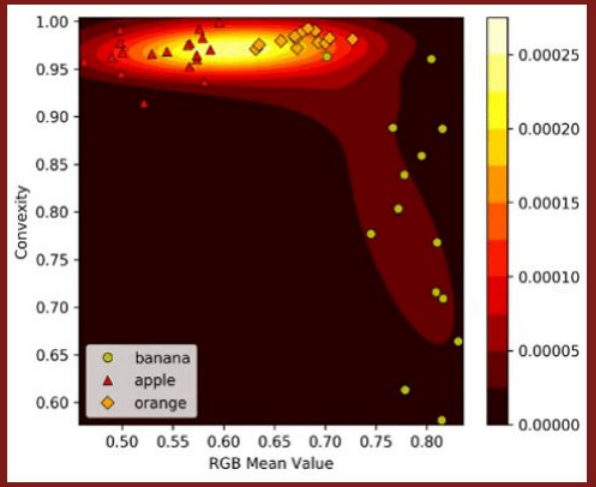
* A corresponding animation for this feature space was attached in the submission

score

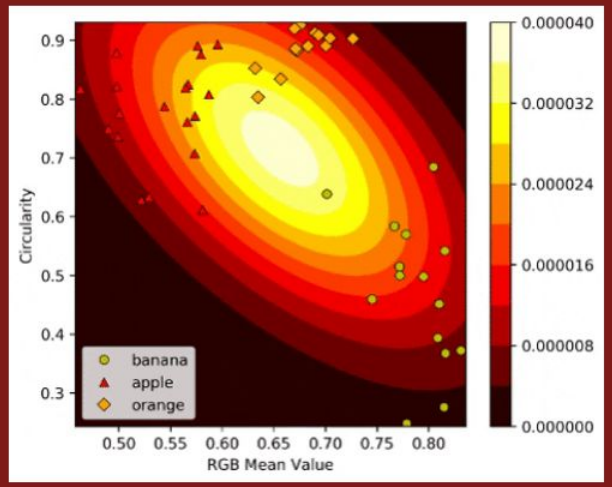
Quality of Presentation	5 out of 5
Technical Correctness	5 out of 5
Initiative	2 out of 2
Overall	12 out of 10

AP186
Activity 15

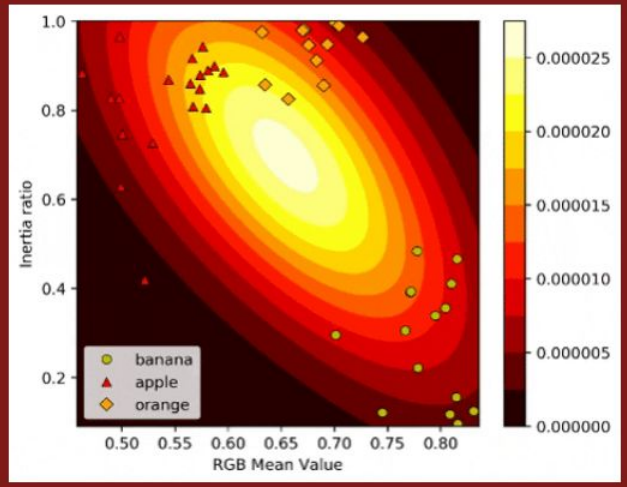
Expectation Maximization GIFs



[Pixel value vs convexity]



[Pixel value vs centrality]



[Pixel value vs inertia ratio]

AP186 Activity 15

Expectation Maximization

Marc Jerrone R. Castro

2015-07420