Problem Set 3: A mixture of five

You're still sorting out the mess that Lestrade left behind when he quit the lab. Apparently some of his last experiments were single-cell RNA-seqs on differentiated sand mouse embryonic stem (ES) cells. You've found records from an experiment in which he was looking at two key early transcription factor genes called *Caraway* and *Kiwi*. Previous work had shown that *Caraway* and *Kiwi* are expressed at intermediate levels in ES cells, but upon differentiation, their mRNA expression patterns break into four different cell types with all four possible combinations of low vs. high expression of these two TFs.

Lestrade's single cell RNA-seq dataset

You've found a data file, `pset3-data.tbl`, where Lestrade had collected mapped read counts for Caraway and Kiwi in 1000 single differentiated ES cells. Because he expected five "cell types" in the data, he used K-means clustering, with K=5, to try to assign each cell to one of the five cell types, and thus estimate the mean expression level (in mapped counts) of the two genes in each cell type, and the relative proportions of the cell types in the population.

However, it's obvious from a figure you found taped into Lestrade's notebook, and various profanities written therein, that the K-means clustering did not go as hoped. His visualization of his data does show five clear clusters, but his K-means clustering failed to identify them (**Figure 1**).

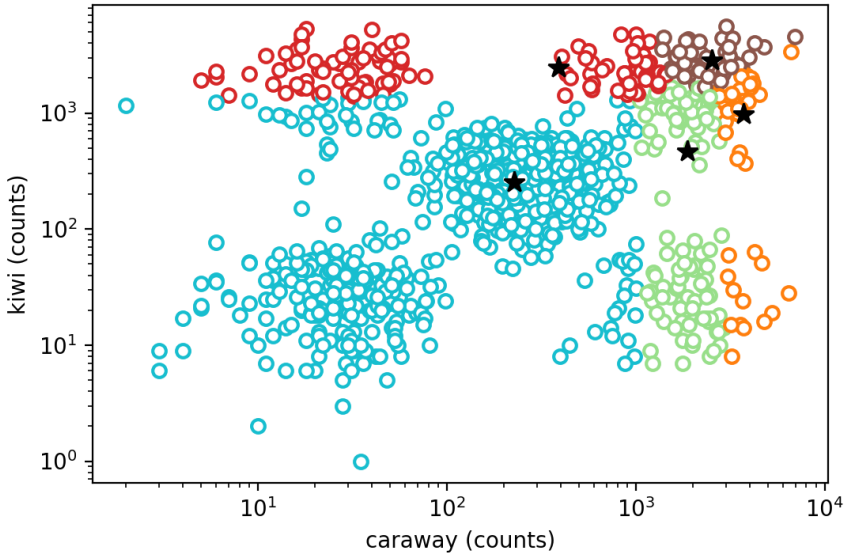

***Figure 1**: Lestrade's figure from his notebook, visualizing his K-means clustering of the 1000 cells in his single cell RNA-seq experiment, for K=5. Black stars indicate the five fitted centroids from his best K-means clustering; colors indicate the assignments of the 1000 cells to the 5 clusters.*

You can see in his notebook that he understood that K-means is prone to local minima, so it's not as if this is a one-off bad solution. His notes indicate that he selected the best of 20 solutions, starting from different random initial conditions. You find the following data table, and a note that this solution had a best "final totdist = 465548.7", of 20 solutions with "totdist" from 465548.7 to 487373.0.

		mean counts	mean counts
cluster	fraction	*Caraway*	*Kiwi*
0	0.054	938.6	3631.0
1	0.106	2052.6	303.2
2	0.111	461.3	1839.1
3	0.650	229.3	227.0
4	0.079	3217.6	1876.1

Interestingly, it looks like all Lestrade was trying to do was to get K-means clustering to work. He must have also had the ES cells marked with reporters that unambiguously labeled each of their cell types, because his data file includes a column for the true cell type (0-4), so the true clustering is known in these data (**Figure 2**).

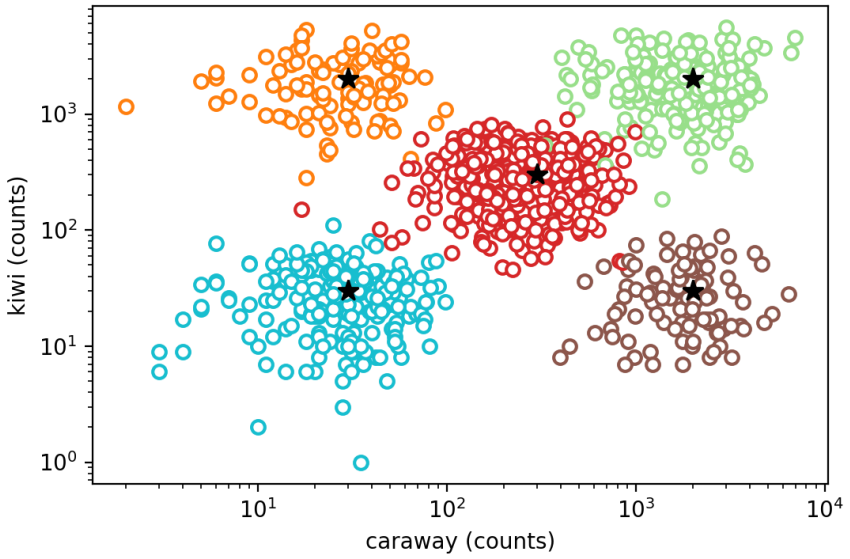

***Figure 2**: Another figure from Lestrade's notebook, showing the true clustering of the 1000 cells.*

1. Reproduce Lestrade's K-means result

Write a standard K-means clustering procedure. Use it to cluster Lestrade's data into K=5 clusters. Plot the results, similar to his figure. You should be able to reproduce a similar result.

You'll want to run the K-means algorithm multiple times and choose the best. What is a good statistic for choosing the "best" solution for K-means? You should be able to reproduce Lestrade's "totdist" measure.

Why is K-means clustering producing this result, when there are clearly five distinct clusters in the data?

2. Mixture negative binomial fitting

Now you're going to use what you've learned about mixture models, and about the negative binomial (NB) distribution for RNA-seq data.

Write an expectation maximization algorithm to fit a mixture negative binomial distribution to Lestrade's data, for $Q = 5$ components in the mixture.

Assume there is a common dispersion $\phi = 0.3$. This means that all you need to re-estimate in the EM algorithm are the means μ for each mixture component.

Like K-means, EM is a local optimizer, so you will want to run your EM algorithm from multiple initial conditions, and take the best one. What is an appropriate statistic for choosing the "best" fit?

What are the estimated mean expression levels of *Caraway* and *Kiwi* in the five cell types, and the relative proportions of each cell type in the 1000 cells?

Visualize your result in a plot similar to Lestrade's.

3. Find a simple fix for K-means

Suggest a simple fix for the problem in applying a standard K-means clustering algorithm to Lestrade's single cell RNA-seq data. Implement the fix, re-run the K-means clustering, pick a "best" solution; report and visualize it.

hints

- The true cell type is noted in the data file, so you can check whether your clustering algorithms are working well.
- K-means (and fitting mixture models by EM) is extraordinarily prone to spurious local optima – as you'll surely see. A lot of the art is in choosing initial conditions wisely. You may want to try some different strategies.
- The mixture modeling EM algorithm will be interestingly parallel to your K-means algorithm. The expectation step, in which you calculate the posterior probability that each data point i belongs to component q, is analogous to the K-means step of assigning a data point i to the current closest centroid q. The maximization step, in which you re-estimate the μ parameters given the posterior-weighted points, corresponds to the K-means step of estimating new centroids from the mean of their assigned points.
- Because the dispersion ϕ is given to you, you only need to estimate μ, the means of each NB component.
- A big hint on part (3): consider how K-means assigns its points to centroids, versus how we're plotting the axes to visualize these data.
- K-means is also prone to artifacts when the variances of the clusters are different, or when the variances aren't uniform in the different directions (dimensions), because it implicitly assumes spherical Gaussian distributions of equal variance.
- You locate the script, `pset3-visualize.py` , Lestrade used to read his data file and produce **Figure 2**. This might help you avoid some of the routine hassles of parsing input and producing output, and focus on the good bit (K-means and EM fitting of a mixture model).