Introduction to Machine Learning
Instructor: Anna Choromanska
Due date: 11/05/2019

1 Problem 1 (10 points): EM Derivation

Consider a random variable x that is categorical with M possible values $1, \ldots, M$. Suppose x is represented as a vector such that $x(j) = 1$ if x takes the j^{th} value, and $\sum_{j=1}^{M} x(j) = 1$. The distribution of x is described by a mixture of K discrete multinomial distributions such that:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$$

where

$$p(x|\mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}$$

where π_k denotes the mixing coefficients for the k^{th} component (aka the prior probability that the hidden variable $z = k$), and μ_k specifies the parameters of the k^{th} component. Specifically, $\mu_k(j)$ represents the probability $p(x(j) = 1|z = k)$ (and, therefore, $\sum_j \mu_k(j) = 1$). Given an observed data set $\{x_n\}$, $n = 1, \cdots, N$, derive the E and M step equations of the EM algorithm for optimizing the mixing coefficients and the component parameters $\mu_k(j)$ for this distribution. For your reference, here is the generic formula for the E and M steps. Note that θ is used to denote all parameters of the mixture model.
E-step. For each n, calculate $\tau_{nj} = p(z_n = j|x_n, \theta)$, i.e., the probability that observation i belongs to each of the K clusters.
M-step. Set

$$\theta := \arg\max_{\theta} \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{nj} \log \frac{p(x_n, z_n = j|\theta)}{\tau_{nj}}.$$

2 Problem 2 (20 points): EM for Bernoulli Mixtures

Part A: Start by downloading the implementation of the Expectation-Maximization algorithm for Gaussian mixture-models for clustering d-dimensional vector data using a mixture of M multivariate Gaussian models. The code is available form the tutorials link as mixmodel.m. You will also need the four .m files below it. This includes randInit.m to initialize the parameters randomly and the functions plotClust.m and plotGauss.m to show the Gaussians overlayed on a plot of the first two dimensions of the data sets after EM converges. Try this code out on datasetA and datasetB by showing a good fit of these two data-sets with 3 Gaussians.

Part B: Now implement a new EM algorithm for clustering Bernoulli models rather than Gaussians for the following game: Alice has K rigged coins, she performs 1000 times the following routine: she picks a coin and tosses it 50 times. The results are reported in the file 'problem2.mat'. Your goal is to determine the most likely K and the Bernoulli coefficients associated with each coin.

Cross-validate to determine the best K ($K \in \{1, 2, 3, 4, 5\}$) by splitting the documents into training and testing. Report the training and testing log-likelihoods as you vary the number of clusters for various random initializations (average and standard deviation).

Report the optimal Bernoulli coefficients for the most likely K (average and standard deviation).

3 Problem 3 (10 points): K-Means for image segmentation

In this question you will implement the K-Means algorithm for image segmentation. Load the built-in image 'trees.tif', crop it and scale it using the following code:

```
raw_im = Tiff('trees.tif','r');
im = raw_im.readRGBAImage();
im = im2double(im(1:200,1:200, :));
```

You can display the image using the following command:

```
imshow(im);
```

Each pixel is represented by a 3-dimensional vector, corresponding to the RGB values. Implement the K-means algorithm as discussed in class on the set of pixels. Show your results for a few values of K (you should display an image where the pixel values are replaced by the nearest centroid mean value, as given in the example).

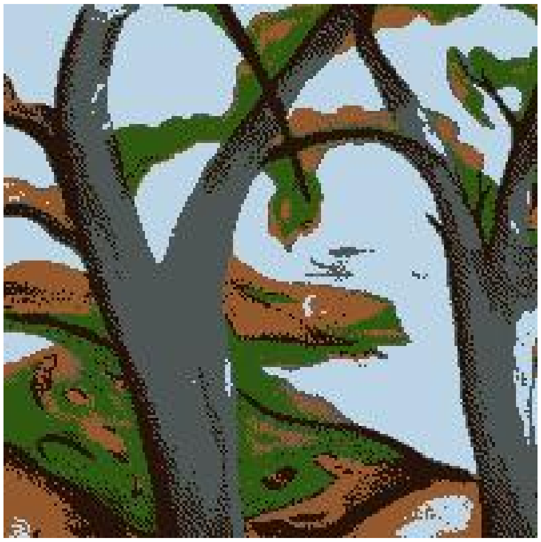

Figure 1: Results for K = 5

As you randomly initialize the K means, you might have some numerical inconsistencies. Explain why you get these inconsistencies and discuss a smartest way of initializing your means. Finally, explain how you would improve this method to get better segmentation.

4 Problem 4 (10 points): Jensen's inequality

Prove the following statements:
a) The arithmetic mean of non-negative numbers is at least their geometric mean.

b) $\sum_{i=1}^{m} \exp(\theta^\top f_i) \geq \exp\left(\theta^\top \sum_{i=1}^{m} \alpha_i f_i - \sum_{i=1}^{m} \alpha_i \log \alpha_i\right)$, where $\alpha_i = \frac{\exp\left(\hat{\theta}^\top f_i\right)}{\sum_{j=1}^{m} \exp(\hat{\theta}^\top f_j)}$.

HINT: Use Jensen's inequality.