STATS 306B Methods for Applied Statistics: Unsupervised Learning
Stanford University, Spring 2014

Problem Set 1

Due: Wednesday, April 16, 2014 (at the start of class)

Instructions:

- Please submit any code written for this assignment along with your derivations and plots.

Practical advice:

- Problem 2 may require a **large** amount of computation time. Please plan accordingly.
- When running Lloyd's algorithm for k-means, you may encounter a situation in which no datapoints are assigned to a particular cluster. In this case, it is common to "reboot" the cluster by selecting a new cluster mean uniformly at random from the datapoints. In fact, any choice for the updated cluster mean is valid in the sense that it cannot upset the monotonic decrease of the k-means objective. However, different reboot strategies will lead to different k-means solutions.

Problem 1 (Choosing the k in k-means)**.**

(a) Generate a training dataset of $n = 20$ datapoints in $\mathbb{R}^3$ by sampling 10 datapoints with independent $\mathcal{N}(2, 1)$ coordinates and 10 with independent $\mathcal{N}(6, 1)$ coordinates. For $k = 1, 2, \ldots, 15$, run Lloyd's algorithm for k-means clustering on the training dataset using ten random restarts (for each random restart, initialize the cluster means to datapoints selected uniformly without replacement from the training data, and for each k only retain the solution that achieves the lowest k-means objective). You are free to use an existing implementation of Lloyd's algorithm or to implement your own version.

Plot the resulting k-means objective on the training dataset as a function of k. Comment on what you see. What value of k would you select based on this plot alone?

(b) Generate a validation dataset of the same size as the training set and in the same fashion. For each k, assign each validation datapoint to the closest learned cluster mean from part (a) and plot the resulting validation k-means objective as a function of k. Comment on what you see. What value of k would be selected if you were to use the minimum validation objective as a selection criterion?

(c) For each $k > 1$, compute and plot the CH statistic of [2] on the training dataset

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)},$$

where $W(k)$ is the within-cluster sum of squares (also known as the k-means objective) on the training data, and $B(k)$ is the between-cluster sum of squares on the training data:

$$B(k) = \sum_{j=1}^{k} n_j \left\| m_j - \bar{m} \right\|_2^2 .$$

Here, n_j is the number of points assigned to cluster j, m_j is the mean of cluster j, and $\bar{m}$ is mean of the entire dataset. Plot CH as a function of k, and comment. What value of k maximizes this criterion?

(d) For each k, compute the gap statistic estimate of [3] on the training dataset. Plot the gap statistic estimate $g(k)$ as a function of k, along with the standard error bars based on the estimate s_k advocated in [3]. Comment on the results. What is the smallest value of k for which $g(k) \geq g(k+1) - s_{k+1}$? (This is the selection criterion recommended in [3].)

Problem 2 (Initializing k-means)**.** In this problem, you will compare three different initialization schemes for k-means clustering:

1. The standard uniform scheme, in which the initial k means are selected uniformly at random (without replacement) from the datapoints in the dataset.
2. The k-means++ initialization scheme [1].
3. An initialization procedure of your own design (your procedure should not be more expensive than a single round of k-means).

For each value of $k \in \{2, 4, 8, 16, 32\}$ and each of the three schemes above, run Lloyd's algorithm on the color image `Colorful-Flowers.jpg` to obtain compressed representations; treat each pixel as a 3-dimensional datapoint. You are free to use an existing implementation of Lloyd's algorithm or to implement your own version. Run each initialization procedure 5 times with different random seeds and, for each procedure, plot the minimum and median k-means objective achieved as a function of k. For each procedure and each k, visualize the compressed image using the best random run (please do not print these images, but do submit your code for displaying the images). What is the smallest value of k for which you are satisfied with the output?

Note: If you use the R package 'jpeg' to load your image, the pixel color values will be represented as real numbers in $[0, 1]$, not integers in $\{0, \ldots, 255\}$, so you will not need to round your codewords to the nearest integer.

Problem 3 (EM for Gaussian Mixture Models)**.**

(a) Derive an EM algorithm for the *isotropic* Gaussian mixture model in which the j-th component distribution has the form $\mathcal{N}(\mu_j, \sigma_j^2 I)$ for unknown $\mu_j \in \mathbb{R}^d$ and $\sigma_j^2 \in \mathbb{R}$.

(b) Implement the EM algorithm of part (a) (do not use a pre-existing implementation) and use it to learn the parameters of a GMM with $k = 2$ and $d = 2$ using the observations in `faithful.csv`. Compute the log likelihood of the training data after each iteration, and plot the likelihood as a function of time. Based on your learned parameters, compute hard assignments for each data point, and plot the data in a way that indicates to which cluster each point belongs; for each cluster, overlay the learned cluster mean and a 95% confidence ellipse for the learned Gaussian component (i.e., an ellipse centered on the cluster mean containing 95% of the probability mass of the learned Gaussian distribution). You may find the R package 'ellipse' helpful.

How would you expect your algorithm behavior and results to change if you instead used a GMM with full unknown covariance matrices Σ_j?

Problem 4 (Admixture / Mixed Membership Modeling)**.** Consider the following admixture model that could be used to infer unobserved ancestral heritage from genetic markers:

- For each individual i in a sample, we observe L genetic markers $g_{il} \in \{0, 1\}$.
- Associated with each individual i is an unknown vector θ_i in the simplex that parameterizes a distribution over k possible ancestral populations.
- Associated with each ancestral population j and each marker location l is an unknown marker frequency p_{jl}.
- For each marker location l and each individual i, there exists a latent ancestry indicator $z_{il} \overset{\text{ind}}{\sim} \text{Mult}(\theta_i, 1)$ representing the ancestral population responsible for g_{il}. All z_{il} are generated independently.
- Each observed marker has the conditional distribution $g_{il} \mid z_{il} = j \overset{\text{ind}}{\sim} \text{Ber}(p_{jl})$. All g_{il} are conditionally independent given $(z_{il})_{i,l}$.

Derive an EM algorithm for estimating the parameters $(p_{jl})_{j,l}, (\theta_i)_i$ of this model.

Problem 5 (Feedback)**.** (This "problem" is not graded.)

(a) How much time did you spend on this problem set?

(b) Which problems did you find valuable?

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[3] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.