Assignment #4

Due: 23:55pm Friday 12 April 2019

Upload at: `https://dropbox.cs.princeton.edu/COS324_S2019/HW4`

Problem 1 (1pt)

Use the instructions above to put your name on the assignment when you compile this document.

Problem 2 (10pts)

In d dimensions, consider a hypersphere of unit radius, centered at zero, which is inscribed in a hypercube, also centered at zero, with edges of length two. What fraction of the hypercube's volume is contained within the hypersphere? Write this as a function of d and make a log-log plot. What happens when d becomes large? [You can just look up the equation for the volume of a hypersphere.]

Problem 3 (15pts)

Consider a d-dimensional Gaussian distribution with zero mean and identity covariance matrix, i.e., $\mathcal{N}(0, \boldsymbol{I}_d)$. Imagine drawing points from this distribution and calculating their squared L^2 distance from the origin. What distribution will these squared distances have for a given d? [Hint: look up the gamma distribution.] Plot the probability density function over the squared distance for $d = 1, 10, 100, 1000, 10000$. Simulate many such Gaussian random variables and create a histogram to verify that your calculations are correct.

Problem 4 (34pts)

Implement K-Means clustering from scratch. That is, don't use a third-party machine learning implementation like `scikit-learn`; math libraries like `numpy` are fine. Go out and grab an image data set like:

- CIFAR-10 or CIFAR-100:
  `http://www.cs.toronto.edu/~kriz/cifar.html`
- MNIST Handwritten Digits:
  `http://yann.lecun.com/exdb/mnist/`
- Small NORB (toys):
  `http://www.cs.nyu.edu/~ylclab/data/norb-v1.0-small/`
- Street View Housing Numbers:
  `http://ufldl.stanford.edu/housenumbers/`
- STL-10:
  `http://cs.stanford.edu/~acoates/stl10/`
- Labeled Faces in the Wild:
  `http://vis-www.cs.umass.edu/lfw/`

Figure out how to load it into your environment and turn it into a set of vectors. Run K-Means on it for a few different K and show some results from the fit. What do the mean images look like? What are some representative images from each of the clusters? Are the results wildly different for different restarts and/or different K? Plot the K-Means objective function (distortion measure) as a function of iteration and verify that it never increases.

Problem 5 (20pts)

Implement K-Means++ and see if it gives you more satisfying initializations for K-Means. Explain your findings.

Problem 6 (20pts)

Download the `cities100.csv` data set from the course website. This file contains the latitude and longitude of the 100 most populous cities in the world. Convert these latitudes and longitudes into pairwise distances using geodesic or great-circle distance (look at `geopy` for this conversion). With this distance matrix in hand, use `scipy.cluster.hierarchy` to explore these data with hierarchical clustering. Produce at least three different dendrograms (with the city names labeled) that use different configurations of linkage, etc. Explain any differences you see arising from different choices of linkage.

Changelog

- 1 April 2019 – Initial version.