Homework 4: Clustering

Writeup due 23:59 on Friday 3 April 2015

You will do this assignment individually and submit your answers as a PDF via the Canvas course website. There is a mathematical component and a programming component to this homework. Do not submit code.

1. The Curse of Dimensionality I [5 pts]

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. What happens when d becomes large?

2. The Curse of Dimensionality II [5 pts]

Consider a d-dimensional Gaussian distribution with zero mean and identity covariance matrix. Imagine drawing points from this distribution and calculating their distance from the origin. What distribution will these distances have, as a function of d? Plot the probability density function. Simulate many such Gaussian random variables and create a histogram to verify that your calculations are correct.

3. Implement K-Means [10 pts]

Implement K-Means clustering from scratch.[1] 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://www.stanford.edu/~acoates//stl10/`
- Labeled Faces in the Wild:
  `http://vis-www.cs.umass.edu/lfw/`

[1] That is, don't use a third-party machine learning implementation like `scikit-learn`; math libraries like `numpy` are fine.

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.

4. Implement K-Means++ [4 pts]

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

5. Calibration [1pt]

Approximately how long did this homework take you to complete?

Changelog

- **v1.0** – 27 March 2015 at 18:00