CS 5785 Homework 2

Applied Machine Learning

Vijay Pillai

Zachary Pardey

November 11, 2017

**Programming Exercises**

2. Clustering for Text Analysis (see Q2.ipynb)
   1. We selected k=4 for the best value of k following empirical review. At K=4, we noticed 4 distinct groupings:
      1. Biochem and Genetics
      2. Earth Science
      3. Biomedicine and Infectious Disease

Physics and Material Science

Top 10 documents of each cluster:

Top 10 in cluster 0

"Requirement of NAD and SIR2 for Life-Span Extension by Calorie Restriction in Saccharomyces Cerevisiae"

"Suppression of Mutations in Mitochondrial DNA by tRNAs Imported from the Cytoplasm"

"Distinct Classes of Yeast Promoters Revealed by Differential TAF Recruitment"

"Efficient Initiation of HCV RNA Replication in Cell Culture"

"Ubiquitination: More Than Two to Tango"

"Negative Regulation of the SHATTERPROOF Genes by FRUITFULL during Arabidopsis Fruit Development"

"T Cell-Independent Rescue of B Lymphocytes from Peripheral Immune Tolerance"

"Reduced Food Intake and Body Weight in Mice Treated with Fatty Acid Synthase Inhibitors"

"Patterning of the Zebrafish Retina by a Wave of Sonic Hedgehog Activity"

"Coupling of Stress in the ER to Activation of JNK Protein Kinases by Transmembrane Protein Kinase IRE1"

Top 10 in cluster 1

"Population Dynamical Consequences of Climate Change for a Small Temperate Songbird"

"The Formation of Chondrules at High Gas Pressures in the Solar Nebula"

"Subducted Seamount Imaged in the Rupture Zone of the 1946 Nankaido Earthquake"

"Tectonic Implications of U-Pb Zircon Ages of the Himalayan Orogenic Belt in Nepal"

"Nitric Acid Trihydrate (NAT) in Polar Stratospheric Clouds"

"Reconstruction of the Amazon Basin Effective Moisture Availability over the past 14,000 Years"

"Greenland Ice Sheet: High-Elevation Balance and Peripheral Thinning"

"Earth's Enigmatic Interface"

"Crossing the Hopf Bifurcation in a Live Predator-Prey System"

"Frozen Methane Escapes from the Sea Floor"

Top 10 in cluster 2

"Algorithmic Gladiators Vie for Digital Glory"

"Reopening the Darkest Chapter in German Science"

"Information Technology Takes a Different Tack"

"National Academy of Sciences Elects New Members"

"Archaeology in the Holy Land"

"Heretical Idea Faces Its Sternest Test"

"Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?"

"Corrections and Clarifications: Charon's First Detailed Spectra Hold Many Surprises"

"Corrections and Clarifications: Unearthing Monuments of the Yarmukians"

"Divining Diet and Disease from DNA"

Top 10 in cluster 3

"Synthesis and Characterization of Helical Multi-Shell Gold Nanowires"

"A Monoclinic Post-Stishovite Polymorph of Silica in the Shergotty Meteorite"

"Ambipolar Pentacene Field-Effect Transistors and Inverters"

"A Stable Bicyclic Compound with Two Si=Si Double Bonds"

"Xenon as a Complex Ligand: The Tetra Xenono Gold(II) Cation in <latex>$AuXe\_4^{2+}(Sb\_2F\_{11}^-)\_2$</latex>"

"Atomic Layer Deposition of Oxide Thin Films with Metal Alkoxides as Oxygen Sources"

"Direct Condensation of Carboxylic Acids with Alcohols Catalyzed by Hafnium(IV) Salts"

"A Cyclic Carbanionic Valence Isomer of a Carbocation: Diphosphino Analogs of Diaminocarbocations"

"Graphical Evolution of the Arnold Web: From Order to Chaos"

"High-Gain Harmonic-Generation Free-Electron Laser"

* 1. We picked k = 10 for this method of clustering following empirical review. Again, at k=10, the terms seemed to be more heavily correlated (terms belonged to vocabularies of specific domains). While there were some thematic repeats across clusters, the clusters certainly appeared to follow some logic.

Top 10 words in each cluster:

Top 10 in cluster 0

blot

incubated

stained

induction

staining

kinase

intracellular

inhibition

assays

promoter

Top 10 in cluster 1

aptamers

trxr

lcts

dnag

proteorhodopsin

doxy

nompc

neas

lg268

rory

Top 10 in cluster 2

dispersion

photon

approximation

momentum

angular

polarization

finite

excited

coherent

energies

Top 10 in cluster 3

figs

intermediate

natl

acad

start

composed

represented

substantially

follows

marked

Top 10 in cluster 4

polymerase

nucleotide

genomic

pcr

conserved

acids

residues

amino

mrna

mutation

Top 10 in cluster 5

interglacial

clim

volcanism

upwelling

interannual

crater

tectonics

plume

decadal

convective

Top 10 in cluster 6

november

Top 10 in cluster 7

concentration

concentrations

Top 10 in cluster 8

vol

p21

cdnas

triton

cyclin

cytosol

eco

mitochondria

methionine

isoforms

Top 10 in cluster 9

recalls

clinton

geneticist

fight

security

prize

spending

campaign

hes

rights

How might such an algorithm be useful?

Clustering by words might be useful in discovering trends in research or in characterizing domain-specific languages (in English). Clustering words gives one insight into the usage of the words, while clustering documents by the corpus of words gives you insight into the themes of the documents.

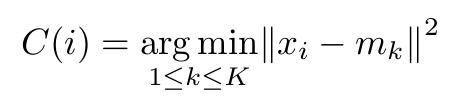
1. EM Algorithm and Implementation (see Q3.ipynb
   1. Show that the alternating algorithm for k-means (in Lec. 11) is a special case of the EM algorithm and show the corresponding objective functions for E-step and M-step.

With help from Lecture 11 Notes:

When you set the covariance matrices to σ2I and drive σ2 to 0, the E-step produces hard assignments of responsibilities (1 for probable classes and 0 for non-probable classes). Following this, the M-step is completed by simply calculating the means.

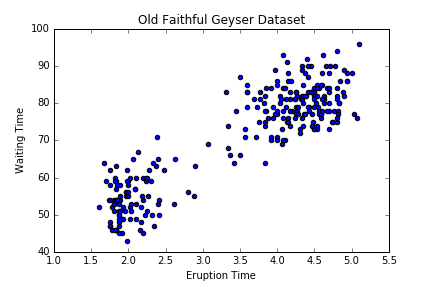
E-step:

γk = IC(i), where

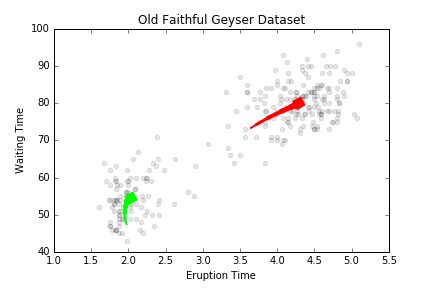
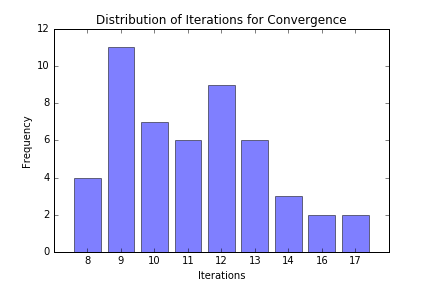
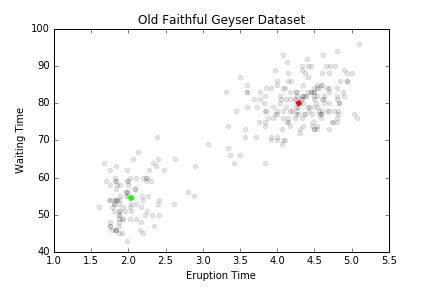
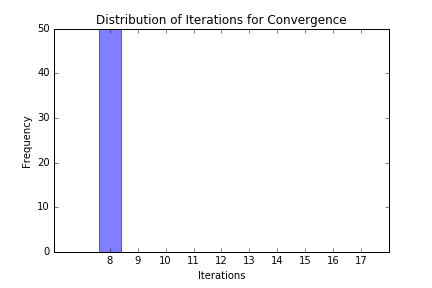


for i = 1… N

M-step:

* 1.   
       
       
       
       
       
       
       
       
       
       
     1. Termination criteria: That the change in gamma between two iterations is sufficiently small.

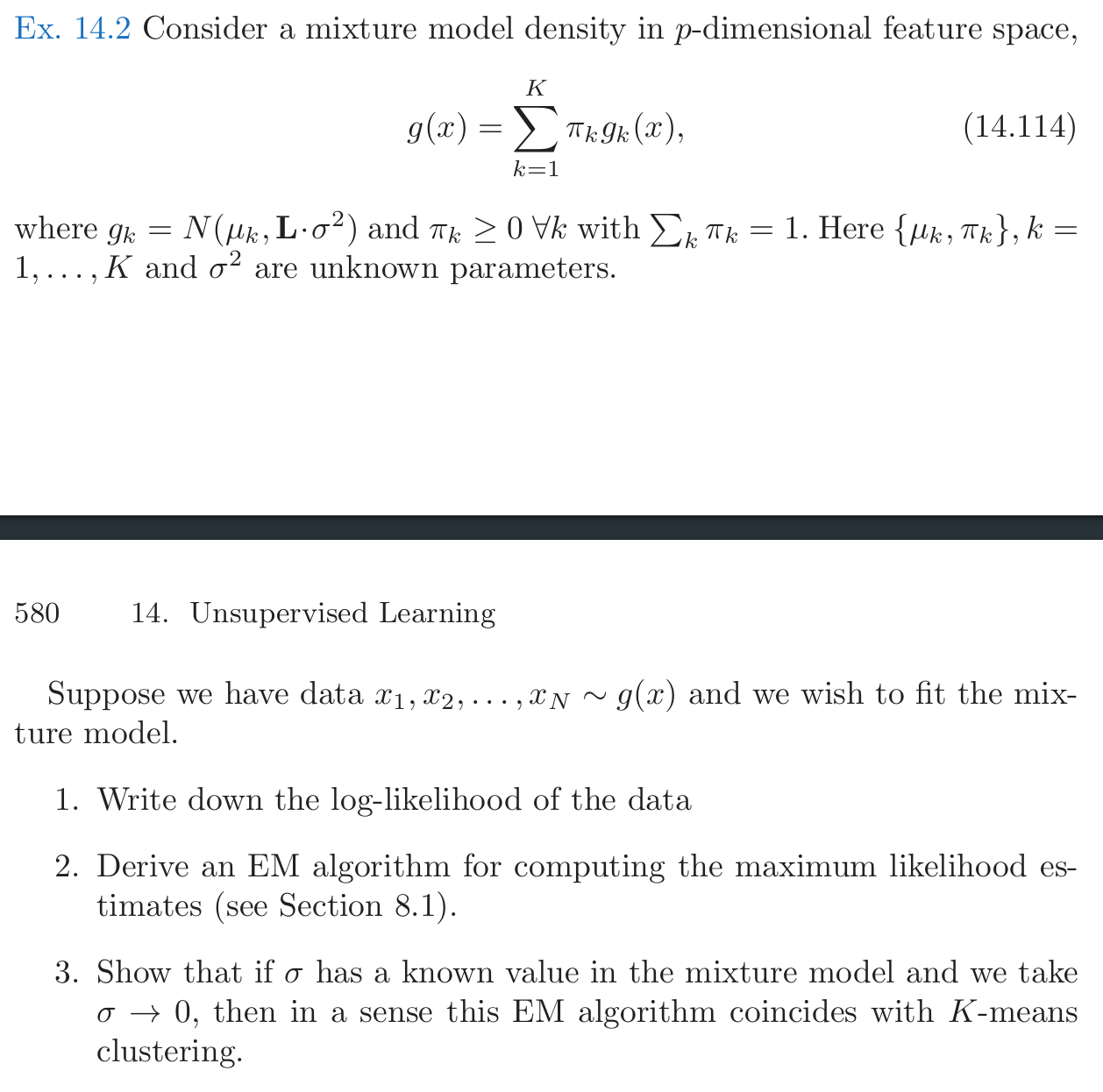
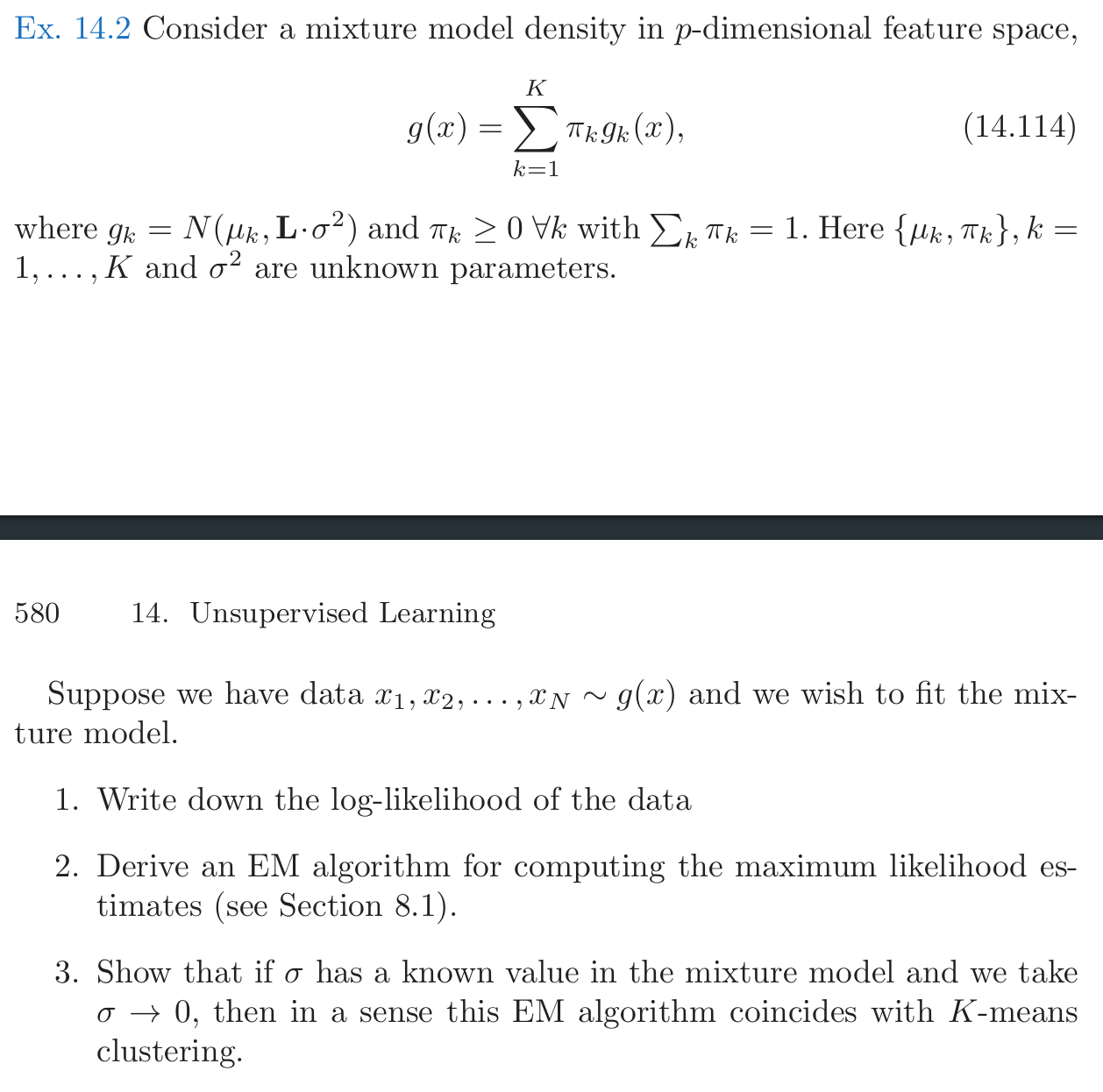
Reasoning: Gamma is the responsibility for both cluster centers. If the change in responsibility is negligible, then the GMM is approaching "equilibrium."

* + 1. 
    2. 
    3. 
    4. 
    5. Compare the algorithm performances of c and d:

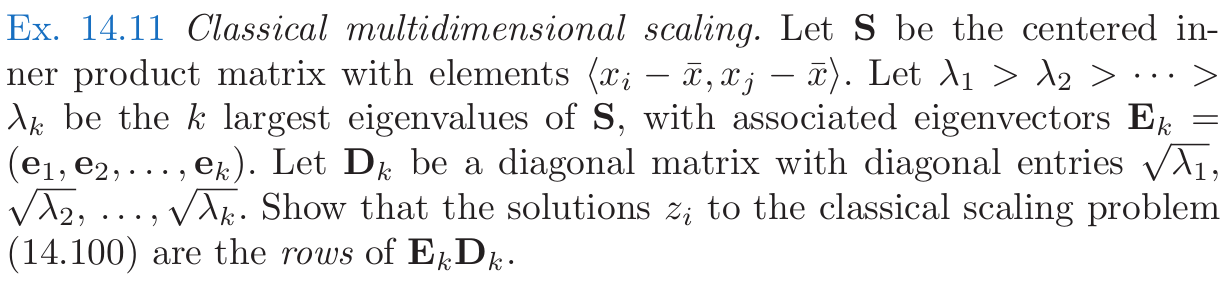
Initializing the parameters with k-means gets the GNN to consistently converge at a minimal number of iterations. Furthermore, while the GNN with random initialization converged at varying intervals (depending on the distance of the initialized centers from the actual centers), k-means causes the GNN to converge at the same number of iterations every time.

**Written Exercises**





* 1. From HTF 8.12, the log likelihood must be:
  2. UNSURE
  3. See: Programming Q3(a).

From Lecture Notes 14 (p4) and Lecture Notes 8 (p3-4):

S = XXT

rank(S) = rank(XXT) = rank(X)

Λ = D2

S = EDDTET = EΛET

X = EΛ1/2=ED

Given that λ1 > … > λkare the k largest eigenvalues, the solutions zi are the rows of EkDk.

1. Decision Trees