Math 189r
Homework 3
November 7, 2016

There are 5 problems in this set. You need to do 3 problems the first week and 2 the second week. Instead of a sixth problem, youare encouraged to work on your final project. Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though. When implementing algorithms you may not use any library (such as `sklearn`) that already implements the algorithms but you may use any other library for data cleaning and numeric purposes (`numpy` or `pandas`). Use common sense. Problems are in no specific order.

1 (Murphy 11.2 - EM for Mixtures of Gaussians) Show that the M step for ML estimation of a mixture of Gaussians is given by

$$\boldsymbol{\mu}_k = \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k}$$
$$\boldsymbol{\Sigma}_k = \frac{1}{r_k}\sum_i r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top = \frac{1}{r_k}\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^\top - r_k\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top.$$

2 (Murphy 11.3 - EM for Mixtures of Bernoullis) Show that the M step for ML estimation of a mixture of Bernoullis is given by

$$\mu_{kj} = \frac{\sum_i r_{ik}x_{ij}}{\sum_i r_{ik}}.$$

Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(\alpha, \beta)$ prior is given by

$$\mu_{kj} = \frac{\left(\sum_i r_{ik}x_{ij}\right) + \alpha - 1}{\left(\sum_i r_{ik}\right) + \alpha + \beta - 2}.$$

3 (MAP Mixture of Gaussians) Consider a mixture of Gaussians with a Dirichlet prior on the mixture weights $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ and a Negative Inverse Walshart prior on the mean and covariance within each class $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \mathrm{NIW}(\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$ with $\kappa_0 = 0$ so only the covariance matrices are regularized. Use $\mathbf{S}_0 = \mathrm{diag}(s_1^2, \ldots, s_D^2)/K^{1/D}$ where $s_j = \sum_i (x_{ij} - \overline{x}_j)^2/N$ is the pooled variance in dimension j. Use $\nu_0 = D + 2$, as that is the weakest prior that is still proper. Use $\boldsymbol{\alpha} = \mathbf{1}$. This is all detailed in Murphy 11.4.2.8. Download the wine quality data at `https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv` and `https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv`. Pool both red and white wine datasets into one dataset and cluster this data using a 2 component MAP Gaussian mixture model with the

EM algorithm. Do the clusters roughly correspond to the color of the wine {white,red} (back this with numbers)? Provide a convergence plot of the MAP objective. If it doesn't monotonically increase there is a bug in your code or math.

4 (MAP Mixture of Bernoullis) Consider a mixture of Bernoullis with a Dirichlet prior on the mixture weights $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ and a Beta prior on the mean parameter $\mu_{kj} \sim \beta(\alpha, \beta)$. Use $\boldsymbol{\alpha} = 1$ and choose an appropriate (α, β) pair for your prior (back this up). Note that the M step for the mean is given in problem 2 (Murphy 11.3). Cluster the MNIST training dataset we used from homework 1 (`http://pjreddie.com/media/files/mnist_train.csv`). Provide a convergence plot of the MAP objective (which must monotonically increase) and plot the mean images for each class. Do the clusters roughly correspond to different digits (back this up with numbers)?

5 (Operations Preserving Kernels) Let $\kappa(\cdot, \cdot)$ and $\lambda(\cdot, \cdot)$ be valid positive semi-definite (Mercer) kernels mapping from a sample space $\mathcal{S}$ to $\mathbb{R}$. Let $\alpha \geq 0$ be a real number and let x and y be elements of $\mathcal{S}$. Prove that

(a) $\alpha\kappa(x, y)$ is a valid kernel.

(b) $\kappa(x, y) + \lambda(x, y)$ is a valid kernel.

(c) $\kappa(x, y)\lambda(x, y)$ is a valid kernel. *Hint:* consider the Cholesky decomposition of the corresponding covariance matrix generated by the product of the kernels. If you're using Bochner's theorem the Convolution Theorem might be of use.

(d) $p(\kappa(x, y))$ is a valid kernel where $p(\cdot)$ is a polynomial with non-negative coefficients.

(e) $\exp(\kappa(x, y))$ is a valid kernel.

(f) $f(x)\kappa(x, y)f(y)$ for all $f : \mathcal{S} \to \mathbb{R}$.

There are multiply ways to prove these. The main methods would be to consider an arbitrary covariance matrix generated by these kernels. If you're versed in functional analysis and Fourier transforms, Bochner's theorem might make some of these easier, but it's not necessary.