

Homework Set Five
ECE 271A
Department of Computer and Electrical Engineering
University of California, San Diego
©Nuno Vasconcelos

This HW set contains several problems. Only the problem labeled **Quiz** must be handed in and will be graded. The remaining problems are for practice. You should not submit them for grade. By submitting your Quiz solution, you agree to comply with the following.

1. The Quiz is treated as a **take-home test** and is an **INDIVIDUAL** effort. **NO collaboration is allowed.** The submitted work must be yours and must be original.
2. The work that you turn-in is your own, using the resources that are available to all students in the class.
3. You can use the help of **GENERAL** resources on programming, such as MATLAB tutorials, or related activities.
4. You are not allowed to consult or use resources provided by tutors, previous students in the class, or any websites that provide solutions or help in solving assignments and exams.
5. You will not upload your solutions or any other course materials to any web-sites or in some other way distribute them outside the class.
6. 0 points will be assigned if your work seems to violate these rules and, if recurrent, the incident(s) will be reported to the Academic Integrity Office.

1. BDR and nearest neighbors Consider a classification problem with c classes and uniform class probabilities, i.e. $P_Y(i) = 1/c, \forall i$. Assume that the goal is to classify an iid sequence of observations $\mathcal{X} = \{x_1, \dots, x_n\}$ as a whole (i.e. the samples are not classified one at a time).

a) Compute the BDR for this problem and show that it converges (in probability) to a nearest neighbor rule based on the class-conditional distributions and the distribution of the observations. Show that the distance function is the Kullback-Leibler divergence

$$\mathcal{D}[p(\mathbf{x})||q(\mathbf{x})] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}.$$

This proves that the BDR for the classification of sequences is really just a nearest neighbor rule.

b) Assuming that all densities are Gaussian with equal covariance Σ , the class conditional densities have mean μ_i and the observation density has mean μ write down an expression for the decision rule as a function of the Gaussian parameters. Provide an interpretation for this new decision rule, by stating what are the items being compared and what is the distance function.

2. Kernel methods Problem 4.3.3 in DHS

3. Multinomial EM In this problem we consider an example where there is a closed-form solution to ML estimation from incomplete data. The goal is to compare with the EM solution and get some insight on how the steps of the latter can be substantially easier to derive than the former.

Consider our bridge example and let U be the type of vehicle that crosses the bridge. U that can take 4 values, (*compact*, *sedan*, *station wagon*, and *pick-up truck*) that we denote by $U \in \{1, 2, 3, 4\}$. On a given day, an operator collects an iid sample of size n from U and the number of vehicles of each type is counted and stored in a vector $\mathcal{D} = (x_1, x_2, x_3, x_4)$. The resulting random variable X (the histogram of vehicle classes) has a multinomial distribution

$$P_{X_1, X_2, X_3, X_4}(x_1, x_2, x_3, x_4; \Psi) = \frac{n!}{x_1!x_2!x_3!x_4!} \left(\frac{1}{2} + \frac{1}{4}\Psi\right)^{x_1} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{x_2} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{x_3} \left(\frac{1}{4}\Psi\right)^{x_4}.$$

However, it is later realized that the operator included *motorcycles* in the *compact class*. It is established that bikes have probability $\frac{1}{4}\Psi$, which leads to a new model

$$\begin{aligned} P_{X_{11}, X_{12}, X_2, X_3, X_4}(x_{11}, x_{12}, x_2, x_3, x_4; \Psi) &= \\ &= \frac{n!}{x_{11}!x_{12}!x_2!x_3!x_4!} \left(\frac{1}{2}\right)^{x_{11}} \left(\frac{1}{4}\Psi\right)^{x_{12}} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{x_2} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{x_3} \left(\frac{1}{4}\Psi\right)^{x_4}. \end{aligned}$$

Determining the parameter Ψ from the available data is as a problem of ML estimation with *missing data*, since we only have measurements for

$$x_1 = x_{11} + x_{12}$$

but not for x_{11} and x_{12} independently.

a) Determine the value of Ψ that maximizes the likelihood of \mathcal{D} , i.e.

$$\Psi_i^* = \arg \max_{\Psi} P_{X_1, X_2, X_3, X_4}(\mathcal{D}; \Psi)$$

by using standard ML estimation procedures.

b) Assume that we have the complete data, i.e. $\mathcal{D}_c = (x_{11}, x_{12}, x_2, x_3, x_4)$. Determine the value of Ψ that maximizes its likelihood, i.e.

$$\Psi_c^* = \arg \max_{\Psi} P_{X_{11}, X_{12}, X_2, X_3, X_4}(\mathcal{D}_c; \Psi),$$

by using standard ML estimation procedures. Compare the difficulty of obtaining this solution vs. that of obtaining the solution in **a)**. Does this look like a problem where EM might be helpful?

c) Derive the E and M-steps of the EM algorithm for this problem.

d) Using the equations for the EM steps, determine the fixed point of the algorithm (i.e. the solution) by making

$$\Psi^{k+1} = \Psi^k$$

where k is the iteration number. Compare to the solution obtained in **a)**.

4. Mixtures of Gaussians The goal of this problem is to give you some “hands-on” experience on the very important case of EM as a tool for the estimation of the parameters of a mixture. Consider a mixture of two Gaussians

$$P_X(x) = \sum_{c=1}^2 \pi_c \mathcal{G}(x, \mu_c, \Sigma_c)$$

where the covariance matrices are diagonal, i.e. $\Sigma_c = \text{diag}(\sigma_{c,1}^2, \sigma_{c,2}^2)$, and a training sample of five points

$$\mathcal{D} = \{(-2.5, -1), (-2, 0.5), (-1, 0), (2.5, -1), (2, 1)\}.$$

a) Assume that the following hold

$$\begin{aligned} \mu_1 &= -\mu_2 \\ \Sigma_1 &= \Sigma_2 = \sigma^2 \mathbf{I}, \\ \pi_1 &= \pi_2 = \frac{1}{2}. \end{aligned}$$

Plot the log-likelihood surface $\log P_X(\mathcal{D})$ as a function of the mean parameters (entries of μ_1) for $\sigma^2 \in \{0.1, 1, 2\}$. Let the coordinate axis cover the range $([-5, 5])$. What can you say about the local maxima of the likelihood surface, and how it changes with σ^2 ? How does the convergence to the optimal depend on the location of the initial parameter guess?

b) Starting from the initial parameter estimate

$$\begin{aligned} \pi_1^{(0)} &= \pi_2^{(0)} = \frac{1}{2} \\ \mu_1^{(0)} &= -\mu_2^{(0)} = (-0.1, 0) \\ \Sigma_1^{(0)} &= \Sigma_2^{(0)} = \mathbf{I}, \end{aligned}$$

compute all the quantities involved in the first 3 iterations of the EM algorithm. For each iteration produce

- plot 1: the posterior surface $P_{Z|\mathbf{x}}(1|\mathbf{x})$ for the first class as a function of \mathbf{x} ,
- plot 2: the mean of each Gaussian, the contour where the Mahalanobis distance associated with it becomes 1, the points in \mathcal{D} , and the means of the solutions obtained the previous steps.

Let EM run until convergence, storing the mean estimates at each iteration. Produce the two plots above for the final solution. In plot 2, plot the values of the means as they progress from the initial to the final estimate.

5. EM and MAP estimates In this problem we use EM for the maximization of the posterior probability

$$\Psi^* = \arg \max_{\Psi} P_{\Psi|X}(\Psi|x).$$

Consider the binomial distribution of problem **3.** and a Gamma prior

$$P_{\Psi}(\Psi) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \Psi^{\nu_1-1} (1 - \Psi)^{\nu_2-1}.$$

Derive the equations of the EM algorithm for MAP estimation of the parameter Ψ .

6. (Quiz) This week we use the *cheetah* image to evaluate the performance of a classifier based on mixture models estimated with EM. Once again we use the decomposition into 8×8 image blocks, compute the DCT of each block, and zig-zag scan. For this (using the data in `TrainingSamplesDCT_new_8.mat`) we fit a mixture of Gaussians of diagonal covariance to each class, i.e.

$$P_{X|Y}(x|i) = \sum_{c=1}^C \pi_c \mathcal{G}(\mathbf{x}, \mu_c, \Sigma_c)$$

where all Σ_c are diagonal matrices. We then apply the BDR based on these density estimates to the *cheetah* image and measure the probability of error as a function of the number of dimensions of the space (as before, use $\{1, 2, 4, 8, 16, 24, 32, \dots, 64\}$ dimensions).

a) For each class, learn 5 mixtures of $C = 8$ components, using a random initialization (recall that the mixture weights must add up to one). Plot the probability of error vs. dimension for each of the 25 classifiers obtained with all possible mixture pairs. Comment the dependence of the probability of error on the initialization.

b) For each class, learn mixtures with $C \in \{1, 2, 4, 8, 16, 32\}$. Plot the probability of error vs. dimension for each number of mixture components. What is the effect of the number of mixture components on the probability of error?