ECE 271A: Homework #5

Due on December 8, 2023 at 11:59pm

Professor Vasconcelos

Ray Tsai

A16848188

Problem 1

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

Part 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 sequence is really just a nearest neighbor rule.

Solution

$$\begin{aligned} i^*(\mathbf{x}) &= \arg\max_i \log P_{\mathbf{X}|Y}(\mathbf{x}|i) + \log P_Y(i) \\ &= \arg\max_i \log P_{\mathbf{X}|Y}(\mathbf{x}|i) \\ &= \arg\max_i \sum_{k=1}^{n} \log P_{\mathbf{X}|Y}(\mathbf{x_k}|i) \\ &= \arg\max_i \frac{1}{n} \sum_{k=1}^{n} \log P_{\mathbf{X}|Y}(\mathbf{x_k}|i). \end{aligned}$$

Thus, when $n \to \infty$, $i^*(\mathbf{x}) \to \arg\max_i E_{\mathbf{X}}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i)]$, by the Law of large numbers. This immediately follows that

$$\begin{aligned} i^*(\mathbf{x}) &= \arg\min_i \; -E_{\mathbf{X}}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i)] \\ &= \arg\min_i E_{\mathbf{X}}[\log Q_X(\mathbf{x})] - E_{\mathbf{X}}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i)] \\ &= \arg\min_i E_{\mathbf{X}}\left[\log \frac{Q_X(\mathbf{x})}{P_{\mathbf{X}|Y}(\mathbf{x}|i)}\right] \\ &= \arg\min_i \int Q_{\mathbf{X}}(\mathbf{x}) \log \frac{Q_{\mathbf{X}}(\mathbf{x})}{P_{\mathbf{X}|Y}(\mathbf{x}|i)} \, dx \\ &= \arg\min_i \mathcal{D}[Q_{\mathbf{X}}(\mathbf{x}) \| P_{\mathbf{X}|Y}(\mathbf{x}|i)], \end{aligned}$$

where $Q_{\mathbf{X}}(\mathbf{x})$ is the density function from which X was sampled. Therefore, the BDR is equivalent to the nearest neighbor search for $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ that's closest to $Q_{\mathbf{X}}(\mathbf{x})$, with the KL-divergence as the distance metric.

Part B

Assuming that all densities are Gaussian with equal covariance $\mathbf{\Sigma}$, 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.

Solution

In this case,

$$\begin{aligned}
i^*(\mathbf{x}) &= \arg\min_i E_{\mathbf{X}}\left[\log\frac{Q_X(\mathbf{x})}{P_{\mathbf{X}|Y}(\mathbf{x}|i)}\right] \\
&= \arg\min_i E_{\mathbf{X}}\left[-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)+\frac{1}{2}(x-\mu_i)^T\mathbf{\Sigma}^{-1}(x-\mu_i)\right] \\
&= \arg\min_i \, -\frac{1}{2}E_{\mathbf{X}}\left[\mu^T\mathbf{\Sigma}^{-1}\mu-\mu_i^T\mathbf{\Sigma}^{-1}\mu_i-2\mu^T\mathbf{\Sigma}^{-1}x+2\mu_i^T\mathbf{\Sigma}^{-1}x\right] \\
&= \arg\min_i \frac{1}{2}\mu_i^T\mathbf{\Sigma}^{-1}\mu_i-\mu_i^T\mathbf{\Sigma}^{-1}\mu+\frac{1}{2}\mu^T\mathbf{\Sigma}^{-1}\mu \\
&= \arg\min_i \frac{1}{2}(\mu_i-\mu)^T\mathbf{\Sigma}^{-1}(\mu_i-\mu).
\end{aligned}$$

Therefore, the BDR is simply looking for the class conditional densities mean μ_i that has the smallest Mahalanobis distance to the observation density μ.

Problem 2

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

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

Determining the parameters 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.

Part 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.

Solution

$$\begin{aligned} \Psi_i^* &= \arg\max_{\Psi} P_{X_1,X_2,X_3,X_4}(\mathcal{D}; \Psi) \\ &= \arg\max_{\Psi} \log P_{X_1,X_2,X_3,X_4}(\mathcal{D}; \Psi) \\ &= \arg\max_{\Psi} x_1 \log\left(\frac{1}{2} + \frac{1}{4}\Psi\right) + x_2 \log\left(\frac{1}{4} - \frac{1}{4}\Psi\right) + x_3 \log\left(\frac{1}{4} - \frac{1}{4}\Psi\right) + x_4 \log\left(\frac{1}{4}\Psi\right). \end{aligned}$$

Let $L = \log P_{X_1,X_2,X_3,X_4}(\mathcal{D};\Psi)$. Since

$$\begin{aligned}
\frac{\partial L}{\partial \Psi} &= \frac{1}{4}\left(\frac{x_1}{\frac{1}{2}+\frac{1}{4}\Psi} - \frac{x_2}{\frac{1}{4}-\frac{1}{4}\Psi} - \frac{x_3}{\frac{1}{4}-\frac{1}{4}\Psi} + \frac{x_4}{\frac{1}{4}\Psi}\right) \\
&= \frac{x_1}{2+\Psi} - \frac{x_2}{1-\Psi} - \frac{x_3}{1-\Psi} + \frac{x_4}{\Psi} \\
&= \frac{\Psi(1-\Psi)^2 x_1 - \Psi(1-\Psi)(2+\Psi)x_2 - \Psi(1-\Psi)(2+\Psi)x_3 + (2+\Psi)(1-\Psi)^2 x_4}{\Psi(2+\Psi)(1-\Psi)^2} \\
&= \frac{\Psi(1-\Psi)x_1 - \Psi(2+\Psi)x_2 - \Psi(2+\Psi)x_3 + (2+\Psi)(1-\Psi)x_4}{\Psi(2+\Psi)(1-\Psi)} \\
&= \frac{-n\Psi^2 + (x_1 - 2x_2 - 2x_3 - x_4)\Psi + 2x_4}{\Psi(2+\Psi)(1-\Psi)} = 0,
\end{aligned}$$

we get the solution $\Psi = \frac{-(x_1-2x_2-2x_3-x_4)\pm\sqrt{(x_1-2x_2-2x_3-x_4)^2+8nx_4}}{2n}$.

Part 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 difficuly of obtaining this solution vs. that of obtaining the solution in part A. Does this look like a problem where EM might be helpful?

Solution

$$\begin{aligned}
\Psi_i^* &= \arg\max_{\Psi} P_{X_{11},X_{12},X_2,X_3,X_4}(\mathcal{D};\Psi) \\
&= \arg\max_{\Psi} \log P_{X_{11},X_{12},X_3,X_4}(\mathcal{D};\Psi) \\
&= \arg\max_{\Psi} x_{12}\log\left(\frac{1}{4}\Psi\right) + x_2\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_3\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_4\log\left(\frac{1}{4}\Psi\right).
\end{aligned}$$

Let $L = \log P_{X_{11},X_{12},X_3,X_4}(\mathcal{D};\Psi)$. Since

$$\begin{aligned}
\frac{\partial L}{\partial \Psi} &= \frac{1}{4}\left(\frac{x_{12}}{\frac{1}{4}\Psi} - \frac{x_2}{\frac{1}{4}-\frac{1}{4}\Psi} - \frac{x_3}{\frac{1}{4}-\frac{1}{4}\Psi} + \frac{x_4}{\frac{1}{4}\Psi}\right) \\
&= \frac{x_{12}}{\Psi} - \frac{x_2}{1-\Psi} - \frac{x_3}{1-\Psi} + \frac{x_4}{\Psi} \\
&= \frac{(1-\Psi)x_{12} - \Psi x_2 - \Psi x_3 + (1-\Psi)x_4}{\Psi(1-\Psi)} \\
&= \frac{-(x_{12}+x_2+x_3+x_4)\Psi + x_{12} + x_4}{\Psi(1-\Psi)} = 0,
\end{aligned}$$

we get the solution $\Psi = \frac{x_{12}+x_4}{x_{12}+x_2+x_3+x_4}$. This solution is a lot simplier than the previous one, and the EM algorithm would be helpful here.

Part C

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

Solution

The observed variables are $X = \{X_1, X_2, X_3, X_4\}$, and the hidden variables are $Z = \{X_{11}, X_{12}\}$. Hence, the Q function is

$$\begin{aligned} Q(\Psi; \Psi^{(n)}) &= E_{Z|X;\Psi^{(n)}}\left[\log P_{X,Z}(\mathcal{D}, Z; \Psi)|\mathcal{D}\right] \\ &= E_{Z|X;\Psi^{(n)}}\left[\log L(\mathcal{D}; \Psi)|\mathcal{D}\right]. \end{aligned}$$

Thus, by the linearity of expectation, the only unknown part of $Q(\Psi; \Psi^{(n)})$ is $E_{X_{12}|X;\Psi^{(n)}}[X_{12}|\mathcal{D}]$. Since X_1, X_2, X_3, X_4 are independent, we are essentially counting the subset of a set of size X_1. We know the probability of event for X_{12} is

$$p = \frac{\frac{\Psi}{4}}{\frac{1}{2} + \frac{\Psi}{4}} = \frac{\Psi}{2 + \Psi},$$

and so

$$P_{X_{12}|X_1}(x_{12}|x_1) = \binom{x_1}{x_{12}} p^{x_{12}}(1-p)^{x_1 - x_{12}}.$$

Thus, the E step consists of computing

$$\hat{x}_{12} = E_{X_{12}|X;\Psi^{(n)}}[X_{12}|\mathcal{D}] = px_1 = \frac{\Psi^{(n)}x_1}{2 + \Psi^{(n)}}.$$

Since the M step is to calculate

$$\begin{aligned} \Psi^{(n+1)} &= \arg\max_{\Psi} Q(\Psi; \Psi^{(n)}) \\ &= \arg\max_{\Psi} E_{Z|X;\Psi^{(n)}}\left[\log L(\mathcal{D}; \Psi)|\mathcal{D}\right] \\ &= \arg\max_{\Psi} \hat{x}_{12}\log\left(\frac{1}{4}\Psi\right) + x_2\log\left(\frac{1}{4} - \frac{1}{4}\Psi\right) + x_3\log\left(\frac{1}{4} - \frac{1}{4}\Psi\right) + x_4\log\left(\frac{1}{4}\Psi\right), \end{aligned}$$

we know $\Psi^{(n+1)} = \frac{\hat{x}_{12}+x_4}{\hat{x}_{12}+x_2+x_3+x_4}$ from part B.

To sum it up:

$$\begin{aligned} \text{E-step} : \hat{x}_{12} &= \frac{\Psi^{(n)}x_1}{2 + \Psi^{(n)}}, \\ \text{M-step} : \Psi^{(n+1)} &= \frac{\hat{x}_{12} + x_4}{\hat{x}_{12} + x_2 + x_3 + x_4}. \end{aligned}$$

Part 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 part A.

Solution

Suppose that $\Psi^{k+1} = \Psi^k$. Then,

$$\begin{aligned}
\frac{\frac{\Psi^{(k)}x_1}{2+\Psi^{(k)}} + x_4}{\frac{\Psi^{(k)}x_1}{2+\Psi^{(k)}} + x_2 + x_3 + x_4} &= \Psi^{(k)} \\
\frac{\Psi^{(k)}x_1 + (2+\Psi^{(k)})x_4}{\Psi^{(k)}x_1 + (2+\Psi^{(k)})(x_2+x_3+x_4)} &= \Psi^{(k)} \\
(x_1+x_2+x_3+x_4)\left(\Psi^{(k)}\right)^2 + 2(x_2+x_3+x_4)\Psi^{(k)} &= (x_1+x_4)\Psi^{(k)} + 2x_4 \\
(x_1+x_2+x_3+x_4)\left(\Psi^{(k)}\right)^2 + (-x_1+2x_2+2x_3+x_4)\Psi^{(k)} - 2x_4 &= 0.
\end{aligned}$$

Notice that this equation coincides with the equation we obtained in part A, indicating that the EM algorithm would eventually converge to the ML solution.

Problem 3

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 2. 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.

Solution

In E-step, we first write out the expectation equation we are to maximize:

$$\begin{aligned}
&E_{Z|X;\Psi}[\log P_{\Psi|X,Z}(\Psi|\mathcal{D},z)|\mathcal{D},\Psi^{(n)}] \\
&= E_{Z|X;\Psi}[\log P_{X,Z|\Psi}(\mathcal{D},z|\Psi)|\mathcal{D},\Psi^{(n)}] + E_{Z|X;\Psi}[\log P_\Psi(\Psi)|\mathcal{D},\Psi^{(n)}] - E_{Z|X;\Psi}[\log P_{X,Z}(\mathcal{D},z)|\mathcal{D},\Psi^{(n)}] \\
&= Q(\Psi;\Psi^{(n)}) + \log P_\Psi(\Psi) - E_{Z|X;\Psi}[\log P_{X,Z}(\mathcal{D},z)|\mathcal{D},\Psi^{(n)}].
\end{aligned}$$

Since the last term does not depend on Ψ, we may ignore it. Thus, in the case of problem 2, the equation we are to maximize becomes

$$\begin{aligned}
Q(\Psi;\Psi^{(n)}) + \log P_\Psi(\Psi) &= E_{Z|X;\Psi^{(n)}}[\log L(\mathcal{D},x_{12};\Psi)|\mathcal{D}] + \log P_\Psi(\Psi) \\
&= \hat{x}_{12}\log\left(\frac{1}{4}\Psi\right) + x_2\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_3\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_4\log\left(\frac{1}{4}\Psi\right) \\
&\quad + (\nu_1-1)\log\Psi + (\nu_2-1)\log(1-\Psi),
\end{aligned}$$

where $\hat{x}_{12} = \frac{\Psi^{(n)}x_1}{2+\Psi^{(n)}}$.

In the M-step, we update the parameter Ψ by calculating

$$\begin{aligned}
\Psi^{(n+1)} = \arg\max_{\Psi} &\hat{x}_{12}\log\left(\frac{1}{4}\Psi\right) + x_2\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_3\log\left(\frac{1}{4}-\frac{1}{4}\Psi\right) + x_4\log\left(\frac{1}{4}\Psi\right) \\
&+ (\nu_1-1)\log\Psi + (\nu_2-1)\log(1-\Psi).
\end{aligned}$$

We do so by taking the partial derivatives

$$\begin{aligned}
\frac{\partial}{\partial\Psi}\left[Q(\Psi;\Psi^{(n)}) + \log P_\Psi(\Psi)\right] &= \frac{-(\hat{x}_{12}+x_2+x_3+x_4)\Psi + \hat{x}_{12}+x_4}{\Psi(1-\Psi)} + \frac{(\nu_1-1)(1-\Psi)-(\nu_2-1)\Psi}{\Psi(1-\Psi)} \\
&= \frac{-(\hat{x}_{12}+x_2+x_3+x_4+\nu_1+\nu_2-2)\Psi + \hat{x}_{12}+x_4+\nu_1-1}{\Psi(1-\Psi)} = 0,
\end{aligned}$$

and we get the solution $\Psi^{(n+1)} = \frac{\hat{x}_{12}+x_4+\nu_1-1}{\hat{x}_{12}+x_2+x_3+x_4+\nu_1+\nu_2-2}$.