Q1 Assume that you are given a dataset $\mathcal{X}\times\mathcal{L} = \{(\mathbf{x}_n, \ell_n); n = 1, \ldots, N\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and ℓ_n is the class label of $\mathbf{x}_n$. Also assume that the dataset is divided into K classes such the set $\mathcal{C}_k$ comprises the vector indexes for which the vectors belong to the k-th class. Linear discriminant analysis (LDA) can project $\mathbf{x}_n$ onto an M-dimensional space to give vector $\mathbf{y}_n$:

$$\mathbf{y}_n = \mathbf{W}^\mathsf{T}(\mathbf{x}_n - \boldsymbol{\mu}), \quad \mathbf{y}_n \in \mathbb{R}^M$$

where $\mathbf{W} = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_M]$ is a $D \times M$ projection matrix comprising M eigenvectors with the largest eigenvalues and $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ is the global mean of the vectors in $\mathcal{X}$.

(a) What is the maximum value of M? Briefly explain your answer.

(5 marks)

(b) The projection matrix $\mathbf{W}$ can be obtained by maximizing the following objective function:

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\left(\mathbf{W}^\mathsf{T}\mathbf{S}_B\mathbf{W}\right)\left(\mathbf{W}^\mathsf{T}\mathbf{S}_W\mathbf{W}\right)^{-1}\right\},$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and the within-class scatter matrices, respectively, and Tr stands for matrix trace. Given that the total scatter matrix $\mathbf{S}_T$ is the sum of $\mathbf{S}_B$ and $\mathbf{S}_W$, i.e.,

$$\mathbf{S}_T = \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T} = \mathbf{S}_B + \mathbf{S}_W,$$

show that

$$\mathbf{S}_B = \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\mathsf{T} \ \text{ and } \ \mathbf{S}_W = \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T},$$

where N_k is the number of samples in class k, i.e., $N_k = |\mathcal{C}_k|$.

(10 marks)

(c) The maximization of $J(\mathbf{W})$ with respect to $\mathbf{W}$ can be achieved by the following constrained optimization:

$$\begin{aligned} \max_{\mathbf{W}} \quad & \mathrm{Tr}\{\mathbf{W}^\mathsf{T}\mathbf{S}_B\mathbf{W}\} \\ \text{subject to} \quad & \mathbf{W}^\mathsf{T}\mathbf{S}_W\mathbf{W} = \mathbf{I} \end{aligned}$$

where $\mathbf{I}$ is an $M\times M$ identity matrix. Show that $\mathbf{W}$ comprises the M eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$. *Hints*: The derivative of matrix trace is

$$\frac{\partial}{\partial\mathbf{X}}\mathrm{tr}\{\mathbf{X}^\mathsf{T}\mathbf{B}\mathbf{X}\mathbf{C}\} = \mathbf{B}\mathbf{X}\mathbf{C} + \mathbf{B}^\mathsf{T}\mathbf{X}\mathbf{C}^\mathsf{T}.$$

(10 marks)

Q2 (a) The output of a support vector machine (SVM) is given by

$$f(\mathbf{x}) = \sum_{i \in \mathcal{S}} a_i K(\mathbf{x}, \mathbf{x}_i) + b,$$

where $\mathbf{x}$ is an input vector on $\mathbb{R}^D$, $\mathcal{S}$ comprises the indexes to a set of support vectors, b is a bias term, a_i's are SVM parameters and $K(\cdot, \cdot)$ is a kernel. For a 2nd-degree polynomial kernel, $K(\cdot, \cdot)$ is given by

$$K(\mathbf{x}, \mathbf{y}) = \left(1 + \mathbf{x}^\mathsf{T}\mathbf{y}\right)^2,$$

where $\mathbf{x}$ and $\mathbf{y}$ are any input vectors on the D-dimensional input space.

(i) Assume that $D = 2$, show that the kernel function maps the input vectors to a 6-dimensional space, where the decision boundary becomes linear.

(10 marks)

(ii) Explain why $f(\mathbf{x})$ is a nonlinear function of $\mathbf{x}$.

(3 marks)

(b) The stochastic gradient-descent algorithm of a single-output deep neural network (DNN) aims to minimize the following instantaneous cross-entropy error:

$$E_{\text{ce}}(\mathbf{W}) = -t \log y - (1 - t) \log(1 - y),$$

where $\mathbf{W}$ comprises the weights of the network, y is the actual output subject to an input vector $\mathbf{x}$, and t is the target output. Assume that the DNN has L layers, show that the error gradient with respect to the weights in the output layer is

$$\frac{\partial E_{\text{ce}}}{\partial w_j^{(L)}} = (y - t) o_j^{(L-1)},$$

where $w_j^{(L)}$ is the weight connecting the j-th hidden neuron in the $(L-1)$-th layer to the output node (at layer L) and $o_j^{(L-1)}$ is the output of this hidden neuron.

(6 marks)

(c) Denote $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as a set of R-dimensional vectors. In factor analysis, $\mathbf{x}_i$'s are assumed to follow a linear model:

$$\mathbf{x}_i = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N$$

where $\mathbf{m}$ is the global mean of vectors in $\mathcal{X}$, $\mathbf{V}$ is a low-rank $R \times D$ matrix, $\mathbf{z}_i$ is a D-dimensional latent factor with prior density $\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\epsilon}_i$ is the residual noise following a Gaussian density with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Based on the Bayes theorem, the posterior density of the latent factor $\mathbf{z}_i$ is given by

$$\begin{aligned} p(\mathbf{z}_i|\mathbf{x}_i) &\propto p(\mathbf{x}_i|\mathbf{z}_i) p(\mathbf{z}_i) \\ &= \mathcal{N}(\mathbf{x}_i|\mathbf{m} + \mathbf{V}\mathbf{z}_i, \boldsymbol{\Sigma}) \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}), \end{aligned}$$

where $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ represents a Gaussian distribution with mean $\boldsymbol{\mu}_z$ and covariance matrix $\boldsymbol{\Sigma}_z$. Show that the posterior mean and posterior moment of $\mathbf{z}_i$ is given by

$$\begin{aligned}\langle \mathbf{z}_i|\mathbf{x}_i\rangle &= \mathbf{L}^{-1}\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m}) \\ \langle \mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathbf{x}_i\rangle &= \mathbf{L}^{-1} + \langle \mathbf{z}_i|\mathcal{X}\rangle\langle \mathbf{z}_i^{\mathsf{T}}|\mathbf{x}_i\rangle,\end{aligned}$$

respectively, where $\mathbf{L}^{-1} = (\mathbf{I} + \mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V})^{-1}$ is the posterior covariance matrix of $\mathbf{z}_i$.

Hints: The Gaussian distribution of random vectors $\mathbf{z}$'s can be expressed as

$$\begin{aligned}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \mathbf{C}_z) &\propto \exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^{\mathsf{T}}\mathbf{C}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)\right\} \\ &\propto \exp\left\{\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\boldsymbol{\mu}_z - \frac{1}{2}\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\mathbf{z}\right\},\end{aligned}$$

where $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are the mean vector and covariance matrix, respectively.

(6 marks)

– END –