Q4 (a) Denote a dataset as $\mathcal{X} \times \mathcal{L} = \{(\mathbf{x}_n, \ell_n); n = 1, \ldots, N\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and ℓ_n is the class label of $\mathbf{x}_n$. Assume that the dataset is divided into two classes such that the sets $\mathcal{C}_1$ and $\mathcal{C}_2$ comprise the vector indexes for which the vectors belong to Class 1 and Class 2, respectively. In Fisher discriminant analysis (FDA), $\mathbf{x}_n$ is projected onto a line to obtain a score $y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n$, where $\mathbf{w}$ is a weight vector defining the orientation of the line. Given that the objective function of FDA is

$$J(\mathbf{w}) = \frac{(\mu_1^y - \mu_2^y)^2}{(\sigma_1^y)^2 + (\sigma_2^y)^2},$$

where μ_k^y and $(\sigma_k^y)^2$ are the mean and variance of the FDA-projected scores for Class k, respectively. Also given is the mean of Class k:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n,$$

where N_k is the number of samples in Class k.

(i) Show that the mean of the projected scores for Class k is

$$\mu_k^y = \mathbf{w}^\mathsf{T}\boldsymbol{\mu}_k.$$

(2 marks)

(ii) Show that the variance of the projected scores for Class k is

$$(\sigma_k^y)^2 = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{w}^\mathsf{T}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}.$$

(5 marks)

(iii) Show that the optimal projection vector $\mathbf{w}^*$ is given by

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} = \frac{\mathbf{w}^\mathsf{T}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{S}_W\mathbf{w}},$$

where

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}$$

and

$$\mathbf{S}_W = \sum_{k=1}^{2} \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}.$$

(6 marks)

(b) In factor analysis, an observed vector $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}$ is the global mean of all possible $\mathbf{x}$'s, $\mathbf{V}$ is a low-rank matrix, $\mathbf{z}$ is the latent factor, and $\boldsymbol{\epsilon}$ is a residue term. Assume that the prior of $\mathbf{z}$ follows a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Show that the covariance matrix of $\mathbf{x}$'s is $\mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}$.

(6 marks)

(c) The kernel K-means algorithm aims to divide a set of training data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ into K disjoint sets $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$ by minimizing the sum of squared error:

$$E_{\boldsymbol{\phi}} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_k} \left\| \boldsymbol{\phi}(\mathbf{x}) - \frac{1}{N_k} \sum_{\mathbf{z} \in \mathcal{X}_k} \boldsymbol{\phi}(\mathbf{z}) \right\|^2, \qquad \text{(Q4-a)}$$

where $\phi(\mathbf{x})$ is a function of $\mathbf{x}$. It can be shown that Eq. Q4-a can be implemented by

$$E'_{\boldsymbol{\phi}} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_k} \left[\frac{1}{N_k^2} \sum_{\mathbf{z} \in \mathcal{X}_k} \sum_{\mathbf{z}' \in \mathcal{X}_k} K(\mathbf{z}, \mathbf{z}') - \frac{2}{N_k} \sum_{\mathbf{z} \in \mathcal{X}_k} K(\mathbf{z}, \mathbf{x}) \right], \qquad \text{(Q4-b)}$$

where $K(\mathbf{z}, \mathbf{z}') = \phi(\mathbf{z})^{\mathsf{T}} \phi(\mathbf{z}')$ is a non-linear kernel.

(i) What is the purpose of the function $\boldsymbol{\phi}(\mathbf{x})$?

(2 marks)

(ii) State an advantage of computing $\phi(\mathbf{z})^{\mathsf{T}} \phi(\mathbf{z}')$ using the non-linear kernel $K(\mathbf{z}, \mathbf{z}')$.

(2 marks)

(iii) Give a function $\phi(\mathbf{x})$ so that $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathsf{T}} \mathbf{y}$.

(2 marks)

Q5 Fig. Q5(a) shows a binary classification problem.

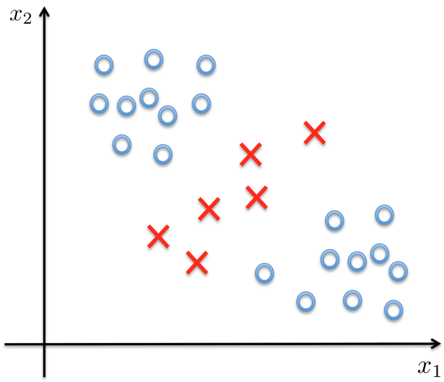

Fig. Q5(a)

(a) (i) Explain why a perceptron (a network with only one neuron) will fail to solve this classification problem.

(3 marks)

(ii) Explain why the network in Fig. Q5(b) can solve this problem perfectly as long as the activation function in the hidden layer is non-linear.

(4 marks)

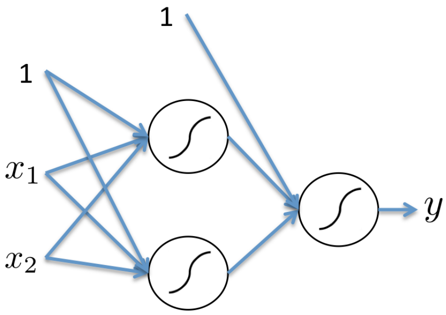

Fig. Q5(b)

(b) The problem in Fig. Q5(a) can also be solved by the network shown in Fig. Q5(c).

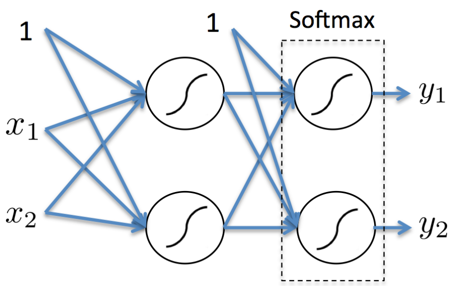

Fig. Q5(c)

The network in Fig. Q5(c) is trained by minimizing the multi-class cross-entropy loss function:

$$E_{\text{mce}} = -\sum_{\mathbf{x}\in\mathcal{X}}\sum_{k=1}^{2} t_k \log y_k, \qquad k = 1, 2$$

where $t_k \in \{0, 1\}$ are the target outputs for the training sample $\mathbf{x} = [x_1 \;\; x_2]^\mathsf{T}$ in the input space and $\mathcal{X}$ is a mini-batch. To use this cross-entropy function, the output nodes should use the softmax function, i.e.,

$$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{2} \exp(a_j)},$$

where a_k is the activation of the k-th node in the output layer.

(i) Show that $0 \leq y_k \leq 1$.

(4 marks)

(ii) Show that E_{mce} can be reduced to the binary cross-entropy:

$$E_{\text{bce}} = \sum_{\mathbf{x} \in \mathcal{X}} [-t_k \log y_k - (1 - t_k) \log(1 - y_k)], \qquad k = 1 \text{ or } 2$$

(4 marks)

(iii) Discuss the implication of the result in Q5(b)(ii).

(3 marks)

(c) Find the extrema (both maxima and minima) of the function $f(x, y) = x + y$ subject to the constraint $x^2 + y^2 \geq 2$. State the values of x and y at which the extrema occur. Give the steps for finding your answers.

(7 marks)

Q6 (a) Given a biased estimator $\hat{\theta}$ with the bias being a function of the true parameter θ, i.e.,

$$\mathbb{E}\{\hat{\theta}\} = \theta + b(\theta),$$

show that the mean square error is

$$\mathrm{mse}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + b^2(\theta),$$

where $\mathrm{var}(\hat{\theta})$ is the variance of $\hat{\theta}$.

(7 marks)

(b) Consider the observation $x[n]$ in Gaussian noise $w[n]$:

$$x[n] = A + w[n], \qquad n = 0, 1, \ldots, N-1,$$

where the noise variance is σ^2, i.e., $w[n] \sim \mathcal{N}(0, \sigma^2)$.

(i) Show that the log-likelihood function of the unknown parameter A is

$$\log p(\mathbf{x}; A) = -\log\left[(2\pi\sigma^2)^{\frac{N}{2}}\right] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2,$$

where $\mathbf{x}[n] = [x[0] \;\; x[1] \;\; \cdots \;\; x[N-1]]^\mathsf{T}$.

(4 marks)

(ii) Show that the Cramer-Rao lower bound (CRLB) of the best unbiased estimator of A is

$$\mathrm{CRLB}(\hat{A}) = \frac{\sigma^2}{N}.$$

(5 marks)

(c) A Kalman filter is used to estimate the position of a train relative to a pole shown in Fig. Q6(a). An RF signal emitting from the pole at regular time intervals is used to estimate the time-of-flight of light z_t (in seconds) from the pole to the train. Denote $\hat{x}_{t|t-1}$ and $\hat{x}_{t|t}$ as the estimates of the train position (in meters) *before* and *after* taking the RF signal at time t into account, respectively. Also denote $\sigma^2_{t|t-1}$ and $\sigma^2_{t|t}$ as the variance of these estimates. The update formulae of the Kalman filter are as follows:

$$\begin{aligned}
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t\left(z_t - \frac{\hat{x}_{t|t-1}}{c}\right) \\
\sigma^2_{t|t} &= \sigma^2_{t|t-1} - \frac{K_t\sigma^2_{t|t-1}}{c} \\
K_t &= \frac{c\sigma^2_{t|t-1}}{\sigma^2_{t|t-1} + c^2\tau^2},
\end{aligned}$$

where τ^2 is the variance of z_t and c is the speed of light (in meter/second).

(i) Show that if the time-of-flight measures $\{z_t\}$ are perfect, the estimate $\hat{x}_{t|t}$

will also be perfect, i.e., having zero variance.

(3 marks)

(ii) Explain why the estimated position of the train becomes more reliable after taking the time-of-flight measurement z_t into account.

(3 marks)

(iii) Show that the Kalman filter will automatically ignore the time-of-flight measure if the measurement becomes very unreliable, i.e., having a large variance.

(3 marks)

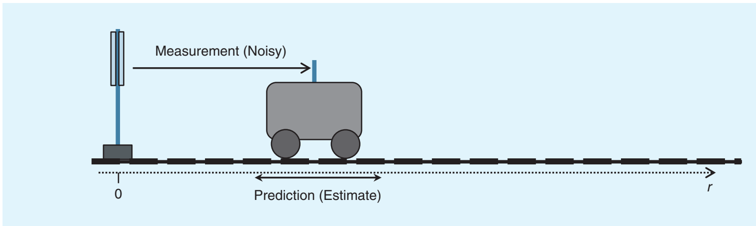

Fig. Q6(a)

— END OF PAPER —