Instructions:

- This final exam is designed to be finished **within 180 minutes**. The 24 hours period are designed to accommodate time zone difference. It includes time for scanning and uploading your submission and any potential technical difficulty. **No late submission is accepted.**
- Follow directions and answer questions with requested supporting work. Be careful not to jump steps.
- Clearly indicate your answer in the allotted space or by putting a box around it.
- The final exam will be posted on June 8th 3:00pm PST. You will have 24 hours to finish and upload your solution to CCLE by June 9th 2:59pm PST. You can use the textbook, any course material posted on CCLE, and your hand-written notes; you are not allowed to use calculators nor the Internet, and you cannot work with anyone else (classmate, family member, private tutor, etc.). You can scan or take high-resolution photos of your hand-written solutions, but the uploaded submission must be a single PDF file.

Problem 1

Given a data set $\mathcal{D} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\} \subseteq \mathbb{R}^D$, it contains $D \times N$ real number entries.

(a) (10 points) Consider an integer $M < D$. PCA takes an M-dimensional subspace $\mathcal{P}_M \subseteq \mathbb{R}^D$ and decomposes each data point

$$\mathbf{x}^{(i)} \approx \pi_{\mathcal{P}_M}(\mathbf{x}^{(i)} - \bar{\mathbf{x}}) + \bar{\mathbf{x}}.$$

Here $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}$ denotes the sample mean, and $\pi_{\mathcal{P}_M}$ denotes the orthogonal projection onto the subspace $\mathcal{P}_M$. What is the choice of $\mathcal{P}_M$ for performing PCA? What is the reason behind this choice? (Hint: This is partially discussed in TP#4 Problem 1 (c) except the reasoning part.)

(b) (10 points) Explain how to use PCA to compress the data. Write the total count of real number entries in the compressed result in terms of D, N and M.

Problem 2

Let $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N} \subset \mathbb{R}^D$ be a set of N data points drawn independently from a unknown probability distribution p.

(a) (5 points) Write down the expressions for the sample mean $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ of the data set $\mathcal{D}$.

(b) (10 points) Find a linear transformation

$$\mathbf{z}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} + \mathbf{b}$$

such that the resulting data set $\mathcal{D}' = \{\mathbf{z}^{(i)}\}_{i=1}^{N}$ has zero mean and identity covariance matrix. Write down the sample mean $\bar{\mathbf{z}}$ and sample covariance matrix $\mathbf{S}'$ of the data set $\mathcal{D}'$ in terms of $\bar{\mathbf{x}}, \mathbf{S}, \mathbf{A}$, and $\mathbf{b}$. Setting $\bar{\mathbf{z}}$ to zero and $\mathbf{S}' = \mathbf{I}$, find the condition on $\mathbf{A}$ and $\mathbf{b}$ for data whitening.

(c) (5 points) Give an example of how data whitening can help a machine learning algorithm.

Problem 3

Consider the following two-layer feed-forward neural network

$$y(\mathbf{x}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}) = \sigma \circ \psi^{(2)} \circ \sigma \circ \psi^{(1)}(\mathbf{x}),$$

where the logistic sigmoid function $\sigma(b) = (1 + \exp(-b))^{-1}$ is chosen as the activation, and the affine map for layer $l = 1, 2$ is given by

$$\psi^{(l)}(\mathbf{z}) = \mathbf{v}^{(l)} + \mathbf{W}^{(l)}\mathbf{z}.$$

Denote the hidden units

$$\begin{aligned}
\mathbf{z}^{(0)} &= \mathbf{x} \in \mathbb{R}^D \\
\mathbf{a}^{(1)} &= \mathbf{v}^{(1)} + \mathbf{W}^{(1)}\mathbf{z}^{(0)} \in \mathbb{R}^M \\
\mathbf{z}^{(1)} &= \sigma(\mathbf{a}^{(1)}) \in \mathbb{R}^M \\
\mathbf{a}^{(2)} &= \mathbf{v}^{(2)} + \mathbf{W}^{(2)}\mathbf{z}^{(1)} \in \mathbb{R}^1 \\
\mathbf{z}^{(2)} &= \sigma(\mathbf{a}^{(2)}) \in \mathbb{R}^1.
\end{aligned}$$

Specifically, the input $\mathbf{x} \in \mathbb{R}^D$ is a D-dimensional vector, and the output $y(\mathbf{x}) \in [0, 1]$ is a scalar. Note that the vector-valued sigmoid function $\sigma(\mathbf{a}) = \sigma(a_1, \cdots, a_M) = (\sigma(a_1), \cdots, \sigma(a_M))$ is defined entry-wise.

(a) (10 points) Using chain rule, compute the following derivatives.

$$\begin{aligned}
&\frac{\partial y}{\partial \mathbf{v}^{(1)}}(\mathbf{x}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}), \\
&\frac{\partial y}{\partial \mathbf{W}^{(1)}}(\mathbf{x}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}).
\end{aligned}$$

(b) (10 points) Consider using this network for a binary classification problem on a data set $\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)}\}_{i=1}^N$ where $t^{(i)} = 1$ denotes class $C_{[1]}$ and $t^{(i)} = 0$ denotes class $C_{[2]}$. The likelihood function is given by

$$\begin{aligned}
p(\mathcal{D}|\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}) &= \prod_{i \in C_{[1]}} y(\mathbf{x}^{(i)}) \prod_{i \in C_{[2]}} (1 - y(\mathbf{x}^{(i)})) \\
&= \prod_{i=1}^N y(\mathbf{x}^{(i)})^{t^{(i)}} (1 - y(\mathbf{x}^{(i)}))^{1 - t^{(i)}}.
\end{aligned}$$

Compute the derivative of the log-likelihood function with respect to $\mathbf{v}^{(1)}$ and $\mathbf{W}^{(1)}$. Is there a closed form for the optimal values that maximize the log-likelihood function?

Problem 4

Given a data set $\mathcal{D} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\} \subseteq \mathbb{R}^D$ sampled from a Gaussian mixture

$$p\left(\mathbf{x}|\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K\right) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \cdots, K\}$ are parameters to be determined. Let $\mathbf{z}^{(i)} \in \{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$ denote the latent variable such that $\mathbf{z}^{(i)} = \mathbf{e}_k$ if $\mathbf{x}^{(i)}$ is sampled from the k-th Gaussian. The EM (*expectation-maximization*) algorithm consists of two steps, E step and M step.

(a) (5 points) The E step treats the variables $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \cdots, K\}$ as constant and updates the posterior probability for cluster membership

$$\gamma_k^{(i)} \approx p(\mathbf{z}^{(i)} = \mathbf{e}_k|\mathbf{x}^{(i)}).$$

Using the Gaussian mixture model and Bayes theorem, give the formula for updating $\gamma_k^{(i)}$ in terms of $\mathbf{x}^{(i)}$ and $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \cdots, K$.

The M step uses the $\gamma_k^{(i)}$ values computed in part (a) and updates the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \cdots, K\}$. The following problems will take you through M step for updating $\boldsymbol{\mu}_k$.

(b) (5 points) The likelihood function is given by

$$\begin{aligned} p\left(\mathcal{D}|\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K\right) &= \prod_{i=1}^N p\left(\mathbf{x}^{(i)}|\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K\right) \\ &= \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right). \end{aligned}$$

Compute the derivative of the log-likelihood function $\log p\left(\mathcal{D}|\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K\right)$ with respect to $\boldsymbol{\mu}_k$ and express the result in terms of $\mathbf{x}^{(i)}, i = 1, \cdots, N, \pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, j = 1, \cdots, K$.

(c) (5 points) Indicate the appropriate term in the result in part (b) that can be approximated by $\gamma_k^{(i)}$ computed in part (a).

(d) (5 points) Set the approximated derivative of the log-likelihood function to zero and find the approximated optimal $\boldsymbol{\mu}_k$ for maximizing the log-likelihood function.

Problem 5

Let $V \subseteq \mathbb{R}^D$ be an M-dimensional subspace with $M < D$. Let p be the probability distribution of a random variable $\mathbf{x} \in \mathbb{R}^D$ such that

$$p(\mathbf{x}) = \begin{cases} (2\pi\sigma^2)^{-\frac{M}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right) & \text{if } \mathbf{x} \in V \\ 0 & \text{otherwise.} \end{cases}$$

(a) (5 points) Let $\mathbf{q}_1, \cdots, \mathbf{q}_M \in \mathbb{R}^D$ be an orthonormal basis of V. Show that the random variable $\mathbf{x}$ defined above satisfies that $\mathbf{x} = \mathbf{Qz}$ where

$$\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_M] \in \mathbb{R}^{D \times M},$$
$$\mathbf{z} \in \mathbb{R}^M, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_M).$$

You can show this by considering a vector valued random variable $\mathbf{y} = \mathbf{Qz} \in \mathbb{R}^D$ with $\mathbf{Q}$ and $\mathbf{z}$ defined above. Write down the probability distribution of $\mathbf{y}$ (in terms of the p.d.f. of $\mathbf{z}$ and $\mathbf{Q}$) and verify that it's indeed the same as $p(\mathbf{x})$.

(b) (5 points) Let $\mathcal{D} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\}$ be a set of N data points drawn independently from the probability distribution p. Suppose that you do not know the value of M. Explain how you could use PCA on $\mathcal{D}$ to estimate M.

(c) (10 points) Now consider a noisy data set $\mathcal{D}_{\text{noisy}} = \{\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(N)}\}$. Each $\mathbf{y}^{(i)}$ is given by

$$\mathbf{y}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\epsilon}^{(i)} \in \mathbb{R}^D,$$

where $\mathbf{x}^{(i)}$ is an independent sample from p, and $\boldsymbol{\epsilon}^{(i)}$ is an independent sample from $\mathcal{N}(\mathbf{0}, \eta^2 \mathbf{I}_D)$. Can you still estimate the value of M using PCA on $\mathcal{D}_{noisy}$? (HINT: your answer may depend on σ and η.)

(d) **(Bonus, up to 10 points)** Suppose the random variable $\mathbf{g} \in \mathbb{R}^D$ is given by

$$\mathbf{g} = \mathbf{Az} + \boldsymbol{\epsilon},$$

where $\mathbf{A} \in \mathbb{R}^{D \times M}$ is a full rank matrix, $\mathbf{z} \in \mathbb{R}^M$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_M)$, and $\boldsymbol{\epsilon} \in \mathbb{R}^D$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \eta^2 \mathbf{I}_D)$. Let $\mathcal{D}'_{\text{noisy}} = \{\mathbf{g}^{(1)}, \cdots, \mathbf{g}^{(N)}\}$ be a data set individually drawn from the probability distribution of $\mathbf{g}$. What is the necessary condition so we can use PCA on $\mathcal{D}'_{\text{noisy}}$ to estimate M?