Total possible marks: 100. Explain all your answers concisely. This set covers chapters 4–9 of the textbook *Introduction to Machine Learning*, 3rd. ed., by E. Alpaydin.

Exercise 1: Euclidean distance classifier (10 points). A Euclidean distance classifier represents each class $k = 1, \dots, K$ by a prototype vector $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and classifies a pattern $\mathbf{x} \in \mathbb{R}^D$ as the class of its closest prototype: $k^* = \arg\min_{k=1,\dots,K} \|\mathbf{x} - \boldsymbol{\mu}_k\|$. Prove that a Gaussian classifier with shared isotropic covariances (i.e., of the form $\boldsymbol{\Sigma}_k = \sigma^2\mathbf{I}$ for $k = 1, \dots, K$, where $\sigma > 0$) and equal class priors (i.e., $p(C_1) = \dots = p(C_K) = \frac{1}{K}$) is equivalent to a Euclidean distance classifier. Prove the class discriminant functions $g_1(\mathbf{x}), \dots, g_K(\mathbf{x})$ are linear and give the expression that defines them.

Exercise 2: bias and variance of an estimator (20 points). Assume we have a sample $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathbb{R}$ of N iid (independent identically distributed) scalar random variables, each of which is drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$. We want to estimate the mean μ of this Gaussian by computing a statistic of the sample $\mathcal{X}$. Consider the following four different statistics of the sample:

1. $\phi_1(\mathcal{X}) = 7$.
2. $\phi_2(\mathcal{X}) = x_1$.
3. $\phi_3(\mathcal{X}) = \frac{1}{N}\sum_{n=1}^N x_n$.
4. $\phi_4(\mathcal{X}) = x_1 x_2$.

For each statistic ϕ, compute:

- (2 points) Its bias $b_\mu(\phi) = \mathrm{E}_{\mathcal{X}}\{\phi(\mathcal{X})\} - \mu$.
- (2 points) Its variance $\mathrm{var}\{\phi\} = \mathrm{E}_{\mathcal{X}}\{(\phi(\mathcal{X}) - \mathrm{E}_{\mathcal{X}}\{\phi(\mathcal{X})\})^2\}$.
- (1 point) Its mean square error $e(\phi, \mu) = \mathrm{E}_{\mathcal{X}}\{(\phi(\mathcal{X}) - \mu)^2\}$.

Based on that, answer the following for each estimator (statistic): is it unbiased? is it consistent?

Hint: expectations wrt the distribution of the N-point sample $\mathcal{X}$ are like this one:

$$\mathrm{E}_{\mathcal{X}}\{\phi(\mathcal{X})\} = \int \phi(\mathbf{x}_1, \dots, \mathbf{x}_N)\, p(\mathbf{x}_1, \dots, \mathbf{x}_N)\, d\mathbf{x}_1 \dots d\mathbf{x}_N \overset{\text{iid}}{=} \int \phi(\mathbf{x}_1, \dots, \mathbf{x}_N)\, p(\mathbf{x}_1) \dots p(\mathbf{x}_N)\, d\mathbf{x}_1 \dots d\mathbf{x}_N.$$

Exercise 3: PCA and LDA (30 points). Consider 2D data points coming from a mixture of two Gaussians with equal proportions, different means, and equal, diagonal covariances (where $\mu, \sigma_1, \sigma_2 > 0$):

$$\mathbf{x} \in \mathbb{R}^2\text{: } p(\mathbf{x}) = \pi_1\, p(\mathbf{x}|1) + \pi_2\, p(\mathbf{x}|2) \qquad p(\mathbf{x}|1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad p(\mathbf{x}|2) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),$$
$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \boldsymbol{\mu}_1 = \mathbf{0},\ \boldsymbol{\mu}_2 = \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

1. (5 points) Compute the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the mixture distribution $p(\mathbf{x})$.

   Hint: let $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k\, p(\mathbf{x}|k)$ for $\mathbf{x} \in \mathbb{R}^D$ be a mixture of K densities, where $\pi_1, \dots, \pi_K \in [0,1]$ and $\sum_{k=1}^{K} \pi_k = 1$ are the component proportions (prior probabilities) and $p(\mathbf{x}|k)$, for $k = 1, \dots, K$, the component densities (e.g. Gaussian, but not necessarily). Let $\boldsymbol{\mu}_k = \mathrm{E}_{p(\mathbf{x}|k)}\{\mathbf{x}\}$ and $\boldsymbol{\Sigma}_k = \mathrm{E}_{p(\mathbf{x}|k)}\{(\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T\}$ be the mean and covariance of component density k, for $k = 1, \dots, K$. Then, the mean and covariance of the mixture are (you should be able to prove this statement):

   $$\boldsymbol{\mu} = \mathrm{E}_{p(\mathbf{x})}\{\mathbf{x}\} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \qquad \boldsymbol{\Sigma} = \mathrm{E}_{p(\mathbf{x})}\left\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right\} = \sum_{k=1}^{K} \pi_k \left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

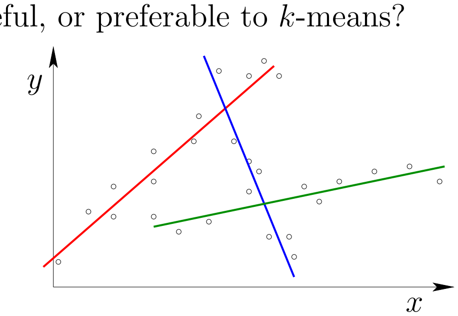
2. (5 points) Compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$ and corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$ of $\boldsymbol{\Sigma}$. Can we have $\lambda_2 > 0$?
3. (2 points) Find the PCA projection to dimension 1.
4. (5 points) Compute the within-class and between-class scatter matrices $\mathbf{S}_W$, $\mathbf{S}_B$ of p.
5. (6 points) Compute the eigenvalues $\nu_1 \geq \nu_2 \geq 0$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$ of $\mathbf{S}_W^{-1}\mathbf{S}_B$. Can we have $\nu_2 > 0$?
6. (2 points) Compute the LDA projection.
7. (5 points) When does PCA find the same projection as LDA? Give a condition and explain it.

Exercise 4: variations of k-means clustering (30 points). Consider the k-means error function:

$$E(\{\boldsymbol{\mu}_k\}_{k=1}^{K}, \mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \quad \text{s.t.} \quad \mathbf{Z} \in \{0,1\}^{NK},\ \mathbf{Z}\,\mathbf{1} = \mathbf{1}$$

over the centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ and cluster assignments $\mathbf{Z}_{N \times K}$, given training points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$.

- **Variation 1:** in k-means, the centroids can take any value in $\mathbb{R}^D$: $\boldsymbol{\mu}_k \in \mathbb{R}^D\ \forall k = 1, \dots, K$. Now we want the centroids to take values from among the training points only: $\boldsymbol{\mu}_k \in \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ $\forall k = 1, \dots, K$.
  1. (8 points) Design a clustering algorithm that minimizes the k-means error function but respecting the above constraint. Your algorithm should converge to a local optimum of the error function. Give the steps of the algorithm explicitly.
  2. (2 points) Can you imagine when this algorithm would be useful, or preferable to k-means?
- **Variation 2:** in k-means, we seek K clusters, each characterized by a centroid $\boldsymbol{\mu}_k$. Imagine we seek instead K lines (or hyperplanes, in general), each characterized by a weight vector $\mathbf{w}_k \in \mathbb{R}^D$ and bias $w_{k0} \in \mathbb{R}$, given a supervised dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ (see figure). Data points assigned to line k should have minimum least-squares error $\sum_{n \in \text{line } k} (y_n - \mathbf{w}_k^T \mathbf{x}_n - w_{k0})^2$.

  1. (8 points) Give an error function that allows us to learn the lines' parameters $\{\mathbf{w}_k, w_{k0}\}_{k=1}^{K}$.
  2. (12 points) Give an iterative algorithm that minimizes that error function.

Exercise 5: mean-shift algorithm (10 points). Consider a Gaussian kernel density estimate

$$p(\mathbf{x}) = \sum_{n=1}^N p(\mathbf{x}|n)p(n) = \frac{1}{N(2\pi\sigma^2)^{D/2}} \sum_{n=1}^N e^{-\frac{1}{2}\left\|\frac{\mathbf{x}-\mathbf{x}_n}{\sigma}\right\|^2} \qquad \mathbf{x} \in \mathbb{R}^D.$$

Derive the mean-shift algorithm, which iterates the following expression:

$$\mathbf{x} \leftarrow \sum_{n=1}^N p(n|\mathbf{x})\mathbf{x}_n \qquad \text{where} \qquad p(n|\mathbf{x}) = \frac{p(\mathbf{x}|n)p(n)}{p(\mathbf{x})} = \frac{\exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_n)/\sigma\|^2\right)}{\sum_{n'=1}^N \exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_{n'})/\sigma\|^2\right)}$$

until convergence to a maximum of p (or, in general, a stationary point of p, satisfying $\nabla p(\mathbf{x}) = \mathbf{0}$). *Hint*: take the gradient of p wrt $\mathbf{x}$, equate it to zero and rearrange the resulting expression.

Bonus exercise: nonparametric regression (20 points). Consider the Gaussian kernel smoother

$$\mathbf{g}(\mathbf{x}) = \sum_{n=1}^N \frac{K\left(\|(\mathbf{x}-\mathbf{x}_n)/h\|\right)}{\sum_{n'=1}^N K\left(\|(\mathbf{x}-\mathbf{x}_{n'})/h\|\right)}\mathbf{y}_n \quad \text{where} \quad K\left(\left\|\frac{\mathbf{x}-\mathbf{x}_n}{\sigma}\right\|\right) \propto \exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_n)/\sigma\|^2\right)$$

estimated on a training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N \subset \mathbb{R}^{D_x} \times \mathbb{R}^{D_y}$.

1. (7 points) What is $\mathbf{g}(\mathbf{x})$ if the training set has only one point ($N = 1$)? Explain.
   Sketch the solution in 1D (i.e., when both $\mathbf{x}_n, \mathbf{y}_n \in \mathbb{R}$).
   Compare with using a least-squares linear regression.
2. (13 points) Prove that, with $N = 2$ points, we can write $\mathbf{g}(\mathbf{x}) = \alpha(\mathbf{x})\,\mathbf{y}_1 + (1 - \alpha(\mathbf{x}))\,\mathbf{y}_2$ where $\alpha(\mathbf{x})$ can be written using the logistic function. Give the detailed expression for $\alpha(\mathbf{x})$.
   Sketch the solution in 1D.
   Compare with using a least-squares linear regression.