Problem 2 (Maximum likelihood in classification, 15pts)

Consider now a generative K-class model. We adopt class prior $p(\mathbf{y} = C_k; \boldsymbol{\pi}) = \pi_k$ for all $k \in \{1, \dots, K\}$ (where π_k is a parameter of the prior). Let $p(\mathbf{x}|\mathbf{y} = C_k)$ denote the class-conditional density of features $\mathbf{x}$ (in this case for class C_k). Consider the data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where as above $\mathbf{y}_i \in \{C_k\}_{k=1}^K$ is encoded as a one-hot target vector and the data are independent.

1. Write out the negative log-likelihood of the data set, $-\ln p(D; \boldsymbol{\pi})$.
2. Since the prior forms a distribution, it has the constraint that $\sum_k \pi_k - 1 = 0$. Using the hint on Lagrange multipliers below, give the expression for the maximum-likelihood estimator for the prior class-membership probabilities, i.e. $\hat{\pi}_k$. Make sure to write out the intermediary equation you need to solve to obtain this estimator. Briefly state why your final answer is intuitive.

For the remaining questions, let the class-conditional probabilities be Gaussian distributions with the same covariance matrix

$$p(\mathbf{x}|\mathbf{y} = C_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \text{ for } k \in \{1, \dots, K\}$$

and different means $\boldsymbol{\mu}_k$ for each class.

3. Derive the gradient of the negative log-likelihood with respect to vector $\boldsymbol{\mu}_k$. Write the expression in matrix form as a function of the variables defined throughout this exercise. Simplify as much as possible for full credit.
4. Derive the maximum-likelihood estimator $\hat{\mu}_k$ for vector $\boldsymbol{\mu}_k$. Briefly state why your final answer is intuitive.
5. Derive the gradient for the negative log-likelihood with respect to the covariance matrix $\boldsymbol{\Sigma}$ (i.e., looking to find an MLE for the covariance). Since you are differentiating with respect to a *matrix*, the resulting expression should be a matrix!
6. Derive the maximum likelihood estimator $\hat{\Sigma}$ of the covariance matrix.

Hint: Lagrange Multipliers. Lagrange Multipliers are a method for optimizing a function f with respect to an equality constraint, i.e.

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } g(\mathbf{x}) = 0.$$

This can be turned into an unconstrained problem by introducing a Lagrange multiplier λ and constructing the Lagrangian function,

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}).$$

It can be shown that it is a necessary condition that the optimum is a critical point of this new function. We can find this point by solving two equations:

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = 0 \text{ and } \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = 0$$

Cookbook formulas. Here are some formulas you might want to consider using to compute difficult gradients. You can use them in the homework without proof. If you are looking to hone your matrix calculus skills, try to find different ways to prove these formulas yourself (will not be part of the evaluation of this homework). In general, you can use any formula from the matrix cookbook, as long as you cite it. We opt for the following common notation: $\mathbf{X}^{-\top} := (\mathbf{X}^\top)^{-1}$

$$\frac{\partial \mathbf{a}^\top \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-\top} \mathbf{a} \mathbf{b}^\top \mathbf{X}^{-\top}$$

$$\frac{\partial \ln |\det(\mathbf{X})|}{\partial \mathbf{X}} = \mathbf{X}^{-\top}$$

Solution

1. Define y_{ik} to be 1 when the ith observation (where $i \in \{1, \ldots, n\}$) is of class k and 0 otherwise. Recall from Lecture 5 that the likelihood of the data set $p(D; \boldsymbol{\pi})$ is equal to the product of the class prior and the class-conditional, that is:

$$p(D; \boldsymbol{\pi}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p(\mathbf{x}_i | \mathbf{y}_i = C_k) p(\mathbf{y}_i = C_k; \boldsymbol{\pi}) \right]^{y_{ik}} = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k p(\mathbf{x}_i | \mathbf{y}_i = C_k) \right]^{y_{ik}}$$

Taking the negative log yields:

$$\boxed{-\ln p(D; \boldsymbol{\pi}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} \big(y_{ik} \ln(\pi_k) + y_{ik} \ln(p(\mathbf{x}_i | \mathbf{y}_i = C_k)) \big)}$$

2. The hint of Lagrange Multipliers says that for an optimization problem of the following form:

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } g(\mathbf{x}) = 0$$

we can equivalently rewrite this as a Lagrangian function:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

For this problem, we have the negative log-likelihood that we've written down from part 2.1:

$$-\ln p(D; \boldsymbol{\pi}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} \big(y_{ik} \ln(\pi_k) + y_{ik} \ln(p(\mathbf{x}_i | \mathbf{y}_i = C_k)) \big)$$

which is subject to the constraint given in the problem:

$$\sum_{k} \pi_k - 1 = 0$$

So we write our Lagrangian function as (note that our above expressions are functions of $\boldsymbol{\pi}$):

$$L(\boldsymbol{\pi}, \lambda) = -\sum_{i=1}^{n} \sum_{k=1}^{K} \big(y_{ik} \ln(\pi_k) + y_{ik} \ln(p(\mathbf{x}_i | \mathbf{y}_i = C_k)) \big) + \lambda \big(\sum_{k=1}^{K} \pi_k - 1 \big)$$

To find the maximum-likelihood estimator for $\hat{\pi}_k$ (the optimum) we need to solve the two equations given in the hints for the Lagrangian Multipliers:

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \boldsymbol{\pi}} = 0 \text{ and } \frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \lambda} = 0$$

Solving for the partial on the left:

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \boldsymbol{\pi}} = \frac{\partial f(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} + \lambda \frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$$

$$\frac{\partial f(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{\pi})}{\partial \pi_1} \\ \frac{\partial f(\boldsymbol{\pi})}{\partial \pi_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{\pi})}{\partial \pi_K} \end{bmatrix} \qquad \frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = \begin{bmatrix} \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_1} \\ \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_2} \\ \vdots \\ \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_K} \end{bmatrix}$$

$$\frac{\partial f(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = \begin{bmatrix} -\sum_{i=1}^{n} \frac{y_{i1}}{\pi_1} \\ -\sum_{i=1}^{n} \frac{y_{i2}}{\pi_2} \\ \vdots \\ -\sum_{i=1}^{n} \frac{y_{iK}}{\pi_K} \end{bmatrix} \qquad \frac{\partial g(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \boldsymbol{\pi}} = \begin{bmatrix} -\sum_{i=1}^{n} \frac{y_{i1}}{\pi_1} \\ -\sum_{i=1}^{n} \frac{y_{i2}}{\pi_2} \\ \vdots \\ -\sum_{i=1}^{n} \frac{y_{iK}}{\pi_K} \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda - \sum_{i=1}^{n} \frac{y_{i1}}{\pi_1} \\ \lambda - \sum_{i=1}^{n} \frac{y_{i2}}{\pi_2} \\ \vdots \\ \lambda - \sum_{i=1}^{n} \frac{y_{iK}}{\pi_K} \end{bmatrix}$$

$$\begin{bmatrix} \lambda - \sum_{i=1}^{n} \frac{y_{i1}}{\hat{\pi}_1} \\ \lambda - \sum_{i=1}^{n} \frac{y_{i2}}{\hat{\pi}_2} \\ \vdots \\ \lambda - \sum_{i=1}^{n} \frac{y_{iK}}{\hat{\pi}_K} \end{bmatrix} = \mathbf{0} \rightarrow \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_K \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda} \sum_{i=1}^{n} y_{i1} \\ \frac{1}{\lambda} \sum_{i=1}^{n} y_{i2} \\ \vdots \\ \frac{1}{\lambda} \sum_{i=1}^{n} y_{iK} \end{bmatrix}$$

The above gives that for some arbitrary k, we have:

$$\hat{\pi}_k = \frac{1}{\lambda} \sum_{i=1}^{n} y_{ik}$$

Therefore, to find our MLE $\hat{\pi}_k$ we simply need to solve for λ.
Solving for the partial on the right:

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1$$

$$\sum_{k=1}^{K} \hat{\pi}_k - 1 = 0$$

$$\sum_{k=1}^{K} \hat{\pi}_k = 1$$

We can substitute our expression above for $\hat{\pi}_k$ to solve for λ:

$$\sum_{k=1}^{K} \frac{1}{\lambda} \sum_{i=1}^{n} y_{ik} = 1$$

$$\sum_{k=1}^{K} \sum_{i=1}^{n} y_{ik} = \lambda$$

Because $\mathbf{y}$ is encoded as a one-hot target vector, we know that $\sum_{k=1}^{K} \sum_{i=1}^{n} y_{ik} = n$ so therefore $n = \lambda$, thus the MLE for the prior class-membership probabilities $\hat{\pi}_k$ is:

$$\boxed{\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} y_{ik}}$$

This makes intuitive sense because it's the sample mean for class k.

3. We have that the class conditional is distributed according to the Gaussian:

$$\mathbf{x}|\mathbf{y} = C_k \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \text{ for } k \in \{1, \dots, K\}$$

Therefore we can write the class-conditional probabilities as:

$$p(\mathbf{x}|\mathbf{y} = C_k) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\Big)$$

Recall the negative log-likelihood from part 2.1 was:

$$-\ln p(D; \boldsymbol{\pi}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \big(y_{ik}\ln(\pi_k) + y_{ik}\ln(p(\mathbf{x}_i|\mathbf{y}_i = C_k))\big)$$

Substituting the class-conditional probabilities, we may rewrite:

$$-\ln p(D; \boldsymbol{\pi}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \Bigg(y_{ik}\ln(\pi_k) + y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\Big(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\Big)\Big)\Bigg)$$

$$-\ln p(D; \boldsymbol{\pi}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \Bigg(y_{ik}\ln(\pi_k) + y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\Big) + y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\Big)\Big)\Bigg)$$

We are seeking an analytical solution to:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D; \boldsymbol{\pi}))$$

Observe that in our expression for the negative log-likelihood that the first and second term are independent of $\boldsymbol{\mu}_k$. Moreover, because we are taking the gradient with respect to a particular class k, we should not longer sum over the K classes because the terms that are not of class k are independent of class k and thus the gradient with respect to k, so we can express the above as:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D; \boldsymbol{\pi})) = -\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\mu}_k}\big(y_{ik}\big(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big)\big)$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D; \boldsymbol{\pi})) = \sum_{i=1}^{n} \frac{y_{ik}}{2}\frac{\partial}{\partial \boldsymbol{\mu}_k}\big((\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big)$$

Solving for the quantity:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\big((\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big)$$

Equation (86) from The Matrix Cookbook gives the following for the Second Order Partial Derivative of a product of matrices (where $\mathbf{W}$ is symmetric):

$$\frac{\partial}{\partial \mathbf{s}}(\mathbf{x} - \mathbf{s})^T \mathbf{W}(\mathbf{x} - \mathbf{s}) = -2\mathbf{W}(\mathbf{x} - \mathbf{s})$$

We can substitute $\boldsymbol{\Sigma}^{-1}$ for $\mathbf{W}$ because $\boldsymbol{\Sigma}^{-1}$ is the symmetric covariance matrix. We can also substitute $\mathbf{x}_i$ for $\mathbf{x}$ and $\boldsymbol{\mu}_k$ for $\mathbf{s}$ to apply the above to solve our partial:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\big((\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big) = -2\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

Substituting the above into our expression for the gradient gives:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D;\boldsymbol{\pi})) = \sum_{i=1}^{n} \frac{y_{ik}}{2} \frac{\partial}{\partial \boldsymbol{\mu}_k}\big((\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big)$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D;\boldsymbol{\pi})) = \sum_{i=1}^{n} \frac{y_{ik}}{2}\big(-2\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\big)$$

$$\boxed{\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n} y_{ik}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)}$$

4. From part 2.4 we found the gradient of the negative log-likelihood with respect to $\boldsymbol{\mu}_k$ to be

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n} y_{ik}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

To find the MLE $\hat{\boldsymbol{\mu}_k}$ requires optimizing the negative log-likelihood by setting the above First Order Condition equal to 0 and then solving for the MLE:

$$-\sum_{i=1}^{n} y_{ik}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n} y_{ik}(\hat{\boldsymbol{\mu}_k} - \mathbf{x}_i) = 0$$

$$\sum_{i=1}^{n} y_{ik}(\hat{\boldsymbol{\mu}_k} - \mathbf{x}_i) = 0$$

$$\sum_{i=1}^{n}(y_{ik}\hat{\boldsymbol{\mu}_k} - y_{ik}\mathbf{x}_i) = 0$$

$$\sum_{i=1}^{n} y_{ik}\hat{\boldsymbol{\mu}_k} - \sum_{i=1}^{n} y_{ik}\mathbf{x}_i = 0$$

$$\sum_{i=1}^{n} y_{ik}\hat{\boldsymbol{\mu}_k} = \sum_{i=1}^{n} y_{ik}\mathbf{x}_i$$

Because $\hat{\boldsymbol{\mu}_k}$ is independent of each observations, rearranging will give our MLE:

$$\hat{\boldsymbol{\mu}_k}\sum_{i=1}^{n} y_{ik} = \sum_{i=1}^{n} y_{ik}\mathbf{x}_i$$

$$\boxed{\hat{\boldsymbol{\mu}_k} = \frac{\sum_{i=1}^{n} y_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} y_{ik}}}$$

This makes intuitive sense because the MLE of the mean $\boldsymbol{\mu}_k$ for class k is the empirical average of the covariates for all the training data points in class k.

5. Recall the negative log-likelihood from part 2.3:

$$-\ln p(D;\boldsymbol{\pi}) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\Bigg(y_{ik}\ln(\pi_k) + y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) + y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Bigg)$$

We are seeking an analytical solution to:

$$\frac{\partial}{\partial\mathbf{\Sigma}}(-\ln p(D;\boldsymbol{\pi}))$$

Observe that in our expression for the negative log-likelihood that the first term is independent of $\boldsymbol{\sigma}$, so we can express the above as:

$$\frac{\partial}{\partial\mathbf{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}\Bigg(y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) + y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Bigg)$$

$$\frac{\partial}{\partial\mathbf{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\Bigg(\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) + \frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Bigg)$$

$$\frac{\partial}{\partial\mathbf{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = -\Bigg(\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) + \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Bigg)$$

$$\frac{\partial}{\partial\mathbf{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)$$

To find an analytical solution, we can solve for the first term and second term separately:
First, solve for:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) = \sum_{i=1}^{n}\sum_{k=1}^{K}y_{ik}\frac{\partial}{\partial\mathbf{\Sigma}}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial\mathbf{\Sigma}}y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}}\Big) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}y_{ik}\frac{\partial}{\partial\mathbf{\Sigma}}\ln\big(\det(2\pi\mathbf{\Sigma})\big)$$

Solving for the partial in the above expression:

$$\frac{\partial}{\partial\mathbf{\Sigma}}\ln\big(\det(2\pi\mathbf{\Sigma})\big)$$

We know that $\mathbf{\Sigma}$ must be $d \times d$ dimensions where d is the number of features in $\mathbf{x}$, therefore:

$$\frac{\partial}{\partial\mathbf{\Sigma}}\ln\big(\det(2\pi\mathbf{\Sigma})\big) = \frac{\partial}{\partial\mathbf{\Sigma}}\ln\big((2\pi)^d\det(\mathbf{\Sigma})\big)$$

$$\frac{\partial}{\partial\mathbf{\Sigma}}\ln\big(\det(2\pi\mathbf{\Sigma})\big) = \frac{\partial}{\partial\mathbf{\Sigma}}\Big(\ln((2\pi)^d) + \ln(\det(\mathbf{\Sigma}))\Big)$$

Observe that the first term is a constant and therefore will evaluate to 0 in our partial, so we have:

$$\frac{\partial}{\partial\mathbf{\Sigma}}\ln\big(\det(2\pi\mathbf{\Sigma})\big) = \frac{\partial}{\partial\mathbf{\Sigma}}\ln(\det(\mathbf{\Sigma}))$$

Recall The Matrix Cookbook formula given in the problem:

$$\frac{\partial\ln|\det(\mathbf{X})|}{\partial\mathbf{X}} = \mathbf{X}^{-\top}$$

Letting $\mathbf{X}$ be $\boldsymbol{\Sigma}$, we can apply the above to rewrite:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln\big(\det(2\pi\boldsymbol{\Sigma})\big) = \boldsymbol{\Sigma}^{-\top}$$

Substituting, we have solved for the first term:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\Big) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\boldsymbol{\Sigma}^{-\top}$$

Next, solve for the second term:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big((\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big)$$

Solving for the partial in the above expression:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big((\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big)$$

Recall The Matrix Cookbook formula given in the problem:

$$\frac{\partial \mathbf{a}^\top\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-\top}\mathbf{a}\mathbf{b}^\top\mathbf{X}^{-\top}$$

Letting $\mathbf{X}$ be $\boldsymbol{\Sigma}$, and both $\mathbf{a}$ and $\mathbf{b}$ be $\mathbf{x}_i$ and $\boldsymbol{\mu}_k$, we can apply the above to rewrite:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big((\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big) = -\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}$$

Substituting, we have solved for the second term:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(-\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}\big)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big) = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}\big)$$

Substituting our solutions for the first and second term, we our desired gradient for the negative log-likelihood with respect to $\boldsymbol{\Sigma}$:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\ln\Big(\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}} y_{ik}\Big(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\Big)\Big)$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\boldsymbol{\Sigma}^{-\top} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}\big)$$

$$\boxed{\frac{\partial}{\partial \boldsymbol{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = \frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\boldsymbol{\Sigma}^{-\top} - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}\big)\Big)}$$

6. From part 2.5 we found the gradient of the negative log-likelihood with respect to $\boldsymbol{\Sigma}$ to be

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}(-\ln p(D;\boldsymbol{\pi})) = \frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\boldsymbol{\Sigma}^{-\top} - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\boldsymbol{\Sigma}^{-\top}(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-\top}\big)$$

To find the MLE $\hat{\mathbf{\Sigma}}$ requires optimizing the negative log-likelihood by setting the above First Order Condition equal to $\mathbf{0}$ and then solving for the MLE:

$$\frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top} - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\big)\Big) = \mathbf{0}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top} - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\big) = \mathbf{0}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\big)$$

Suppose we right multiply both sides by $\hat{\mathbf{\Sigma}}^{\top}$:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top}\hat{\mathbf{\Sigma}}^{\top} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\big)\hat{\mathbf{\Sigma}}^{\top}$$

We know that (letting $\mathbf{I}$ be the identity matrix):

$$\hat{\mathbf{\Sigma}}^{-\top}\hat{\mathbf{\Sigma}}^{\top} = (\hat{\mathbf{\Sigma}}^{-1})^{\top}\hat{\mathbf{\Sigma}}^{\top} = (\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}})^{\top} = \mathbf{I}^{\top} = \mathbf{I}$$

So we have that:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{I} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big(\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\big)\hat{\mathbf{\Sigma}}^{\top}$$

By the Associative property of matrix multiplication, we can rewrite the right-hand side to be:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{I} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\hat{\mathbf{\Sigma}}^{-\top}\hat{\mathbf{\Sigma}}^{\top}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{I} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}(\hat{\mathbf{\Sigma}}^{-\top}\hat{\mathbf{\Sigma}}^{\top})$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{I} = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\hat{\mathbf{\Sigma}}^{-\top}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\mathbf{I}$$

Because $\mathbf{\Sigma}$ is identical for all classes and for all observations, it is akin to a multiplicative constant in the sum, so therefore we can rewrite the right-hand side as:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\mathbf{I} = \hat{\mathbf{\Sigma}}^{-\top}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\mathbf{I}$$

Observe that because our target vector $\mathbf{y}$ is one-hot encoded, we know that $\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} = n$, therefore we may rewrite:

$$n\mathbf{I} = \hat{\mathbf{\Sigma}}^{-\top}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\mathbf{I}$$

$$\mathbf{I} = \frac{1}{n}\hat{\mathbf{\Sigma}}^{-\top}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\top}\mathbf{I}$$

Suppose we left multiply both sides by $\hat{\mathbf{\Sigma}}^\top$

$$\hat{\mathbf{\Sigma}}^\top \mathbf{I} = \frac{1}{n}\hat{\mathbf{\Sigma}}^\top \hat{\mathbf{\Sigma}}^{-\top} \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \mathbf{I}$$

$$\hat{\mathbf{\Sigma}}^\top = \frac{1}{n}\mathbf{I}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \mathbf{I}$$

$$\hat{\mathbf{\Sigma}}^\top = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$$

Taking the transpose of both sides gives our desired MLE for the covariance matrix:

$$(\hat{\mathbf{\Sigma}}^\top)^\top = \frac{1}{n}\big(\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top\big)^\top$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\big(\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top\big)^\top$$

Because the transpose of a sum is equivalent to a sum of transposes, we can simplify the right-hand side to be

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\big((\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top\big)^\top$$

Moreover, the transpose of a product is the product of the transposes in reverse order, so we can further simplify the right-hand side to be:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}((\mathbf{x}_i - \boldsymbol{\mu}_k)^\top)^\top(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$$

$$\boxed{\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top}$$