Problem 1: Consider a generative classification model for C classes defined by class probabilities $\mathrm{p}(y = c) = \pi_c$ and general class-conditional densities $\mathrm{p}(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}_c)$ where $\boldsymbol{x} \in \mathbb{R}^D$ is the input feature vector and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c=1}^C$ are further model parameters. Suppose we are given a training set $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1}^N$ where $\boldsymbol{y}^{(n)}$ is a binary target vector of length C that uses the 1-of-C (one-hot) encoding scheme, so that it has components $y_c^{(n)} = \delta_{ck}$ if pattern n is from class $y = k$. Assuming that the data points are i.i.d., show that the maximum-likelihood solution for the class probabilities $\boldsymbol{\pi}$ is given by

$$\pi_c = \frac{N_c}{N}$$

where N_c is the number of data points assigned to class c.

The data likelihood given the parameters $\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C$ is

$$\mathrm{p}(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C) = \prod_{n=1}^N \prod_{c=1}^C (\mathrm{p}(\boldsymbol{x}^{(n)}|\boldsymbol{\theta}_c)\pi_c)^{y_c^{(n)}}$$

and so the data log-likelihood is given by

$$\log \mathrm{p}(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C) = \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \pi_c + \cancel{\text{const w.r.t. } \pi_c}.$$

In order to maximize the log likelihood with respect to π_c we need to preserve the constraint $\sum_c \pi_c = 1$. For this we use the method of Lagrange multipliers where we introduce λ as an unconstrained additional parameter and find a local extremum of the unconstrained function

$$\sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \pi_c - \lambda \left(\sum_{c=1}^C \pi_c - 1 \right).$$

instead. See wikipedia article on Lagrange multipliers for an intuition of why this works. This function is a sum of concave terms in π_c as well as λ and is therefore itself concave in these variables. We can find the extremum by finding the root of the derivatives. Setting the derivative with respect to π_c equal to zero, we obtain

$$\pi_c = \frac{1}{\lambda} \sum_{n=1}^N y_c^{(n)} = \frac{N_c}{\lambda}.$$

Setting the derivative with respect to λ equal to zero, we obtain the original constraint

$$\sum_{c=1}^C \pi_c = 1$$

where we can now plug in the previous result $\pi_c = \frac{N_c}{\lambda}$ and obtain $\lambda = \sum_c N_c = N$. Plugging this in turn into the expression for π_c we obtain

$$\pi_c = \frac{N_c}{N}$$

which we wanted to show.

1-hot

$C = 5 \rightarrow$ $1 \rightarrow 00001$

$2 \rightarrow 00010$

$3 \rightarrow 00100$

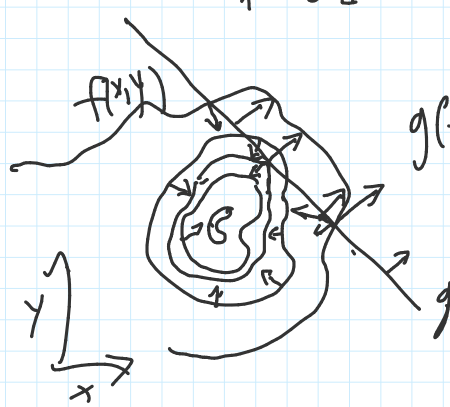
$\vdots$

$$\delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise} \end{cases} \qquad x^0 = 1$$

$$p(D|\pi,\theta) = \prod_n p(x^n, y^n)$$

$$= \prod_n p(x^n | y_n, \theta) \cdot p(y_n | \pi)$$

$$= \prod_n \prod_c \left[p(x_n | y=c, \theta) \cdot p(y=c|\pi) \right]^{y_c^{(n)}}$$



$g(x,y) = 0$

$\nabla f = \lambda \nabla g$

Problem 2: Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a *shared* covariance matrix, so that

$$\mathrm{p}(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}).$$

Show that the maximum likelihood estimate for the mean of the Gaussian distribution for class c is given by

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^N \boldsymbol{x}^{(n)}$$

which represents the mean of the observations assigned to class c.

Similarly, show that the maximum likelihood estimate for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{c=1}^C \frac{N_c}{N} \boldsymbol{S}_c \quad \text{where} \quad \boldsymbol{S}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^N (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}.$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the sample covariances of the data associated with each class, in which the weighting coefficients N_c/N are the prior probabilities of the classes.

We begin by writing out the data log-likelihood.

$$\log \mathrm{p}(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C)$$
$$= \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \pi_c \cdot \mathrm{p}(\boldsymbol{x}^{(n)} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$$

Then we plug in the definition of the multivariate Gaussian

$$= \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \left((2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c) \right) \right) + y^{(n)} \log \pi_c$$

and simplify.

$\in \mathbb{R}$

$$= -\frac{1}{2} \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \left(D \log 2\pi + \log\det(\boldsymbol{\Sigma}) + (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c) - 2 \log \pi_c \right)$$

$$-\log x = \log x^{-1}$$

This expression is concave in $\boldsymbol{\mu}_c$, so we can obtain the maximizer by finding the root of the derivative. With the help of the matrix cookbook, we identify the derivative with respect to $\boldsymbol{\mu}_c$ as

$$\frac{1}{2}\sum_{n=1}^{N} y_c^{(n)}\Sigma^{-1}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)$$

which we can set to 0 and solve for $\boldsymbol{\mu}_c$ to obtain

$$\boldsymbol{\mu}_c = \frac{1}{\sum_{n=1}^{N} y_c^{(n)}}\sum_{n=1}^{N} y_c^{(n)}\boldsymbol{x}^{(n)} = \frac{1}{N_c}\sum_{\substack{n=1\\ y^{(n)}=c}}^{N} \boldsymbol{x}^{(n)}.$$

To find the optimal $\boldsymbol{\Sigma}$, we need the trace trick

$$a = \mathrm{Tr}(a) \text{ for all } a \in \mathbb{R} \quad \text{and} \quad \mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{BCA}).$$

With this we can rewrite

$$(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c) = \mathrm{Tr}\left(\overset{A}{\Sigma^{-1}}\overbrace{(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}}^{B}\right)$$

and use the matrix-trace derivative rule $\frac{\partial}{\partial A}\mathrm{Tr}(\boldsymbol{AB}) = \boldsymbol{B}^{\mathrm{T}}$ to find the derivative of the data log-likelihood with respect to $\boldsymbol{\Sigma}$. Because the log-likelihood contains both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$, we convert one into the other with $\log\det \boldsymbol{A} = -\log\det \boldsymbol{A}^{-1}$ to obtain

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)}\left(-\log\det\boldsymbol{\Sigma}^{-1} + \mathrm{Tr}\left(\Sigma^{-1}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}\right)\right) + \text{const w.r.t. } \boldsymbol{\Sigma}.$$

Finally, we use rule (57) from the matrix cookbook $\frac{\partial \log|\det \boldsymbol{X}|}{\partial \boldsymbol{X}} = (\boldsymbol{X}^{-1})^{\mathrm{T}}$ and compute the derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}^{-1}$ as

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)}\left(-\boldsymbol{\Sigma}^{\mathrm{T}} + (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}\right).$$

$\frac{N_c}{N_c}$

We find the root with respect to $\boldsymbol{\Sigma}$ and find

$$\boldsymbol{\Sigma} = \frac{1}{\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)}}\left(\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}\right)^{\mathrm{T}} = \frac{1}{N}\sum_{c=1}^{C}\sum_{\substack{n=1\\ y^{(n)}=c}}^{N} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}$$

which we can immediately break apart into the representation in the instructions.

$$\mathrm{Tr}\begin{pmatrix} a_{11} & & a_{13} \\ & a_{22} & \vdots \\ c & \cdots & \end{pmatrix} = \text{sum of diagonal} = \sum_i a_{ii}$$

(a)