STAT 543/641 – WINTER 2019 – HOMEWORK #2

DUE MARCH ??, 2019

Let $\sigma(x) = (1 + e^{-x})^{-1}$ be the sigmoid function. The negative log-likelihood function associated to fitting a univariate logistic regression model to a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$\ell(a,b) = -\sum_{i=1}^{n} \Big(y_i \log \sigma(a + bx_i) + (1 - y_i) \log \big(1 - \sigma(a + bx_i)\big) \Big).$$

In this problem we will identify a condition under which $\ell(a, b)$ does not have a global minimum and a codition under which $\ell(a, b)$ has a unique local minimum.

(1) Let σ be the sigmoid function. Prove:

$$(*) \qquad \sigma'(x) = \sigma(x)\big(1 - \sigma(x)\big)$$

Conclude that $\sigma'(x) > 0$ for all x.

(2) Let x and y be constants and let

$$g(a,b) = y \log \sigma(a + bx) + (1 - y) \log \big(1 - \sigma(a + bx)\big).$$

Prove that

$$\nabla g(a,b) = (y - \sigma(a + bx)) \begin{bmatrix} 1 \\ x \end{bmatrix} \qquad \text{(Hint: Use } (*).)$$

and that

$$\nabla^2 g(a,b) = -\sigma'(a + bx) \begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}.$$

Show that $\nabla^2 g(a, b)$ is positive-semidefinite but not positive-definite, making $g(a, b)$ convex but not strictly convex.

(3) By (2) and the linearity of the Hessian operator ∇^2,

$$\nabla^2 \ell(a,b) = \sum_{i=1}^{n} Q_i(a,b), \quad \text{where} \qquad Q_i(a,b) = \sigma'(a + bx_i) \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}$$

(4) Find a basis of the nullspace $N(Q_i(a, b))$ of $Q_i(a, b)$ whose elements do not depend on a and b.

(5) Suppose that there are indices i and j such that $x_i \neq x_j$. Prove that

$$\bigcap_{i=1}^{n} N(Q_i(a,b)) = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

(Hint: Use (4).)

(6) Suppose that there are indices i and j such that $x_i \neq x_j$. Show that $\nabla^2 \ell(a, b)$ is positive-definite and, hence, that $\ell(a, b)$ is strictly convex.

(7) Conclude that if that there are indices i and j such that $x_i \neq x_j$, then maximum likelihood estimates for $\widehat{a}$ and $\widehat{b}$ are unique if they exist.

$$\nabla^2 \ell(a, b) = \sum_{i=1}^{n} Q_i(a, b), \quad \text{where} \qquad Q_i(a, b) = \sigma'(a + bx_i) \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}$$

where

$$Q_i(a, b) = \sigma'(a + bx_i) \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

$$\begin{aligned} \frac{\partial g}{\partial a} &= y \frac{\sigma'(a + bx)}{\sigma(a + bx)} + (1 - y) \frac{(-\sigma'(a + bx))}{1 - \sigma(a + bx)} \\ &= y(1 - \sigma(a + bx)) - (1 - y)\sigma(a + bx) \\ &= y - y\sigma(a + bx) - \sigma(a + bx) + y\sigma(a + bx) \\ &= y - \sigma(a + bx) \end{aligned}$$

$$\begin{aligned} \frac{\partial g}{\partial b} &= y \frac{\sigma'(a + bx)x}{\sigma(a + bx)} + (1 - y) \frac{(-\sigma'(a + bx)x)}{1 - \sigma(a + bx)} \\ &= xy(1 - \sigma(a + bx)) - x(1 - y)\sigma(a + bx) \\ &= xy - xy\sigma(a + bx) - x\sigma(a + bx) + xy\sigma(a + bx) \\ &= x(y - \sigma(a + bx)) \end{aligned}$$

$$\begin{aligned} g''(b) &= -x^2 \sigma'(a + bx) \\ &= -x^2 \sigma(a + bx)(1 - \sigma(a + bx)) \end{aligned}$$

$$\ell(a, b) = -\log L(a, b) = -\sum_{i=1}^{n} \Big(y_i \log \sigma(a + bx_i) + (1 - y_i) \log \big(1 - \sigma(a + bx_i)\big) \Big)$$

$$\ell''(a, b) = \sum_{i=1}^{n} x_i^2 \sigma(a + bx_i) \big(1 - \sigma(a + bx_i)\big)$$

For a vector $v \in \mathbb{R}^2$, write H_v for the open half plane

$$H_v = \{w \in \mathbb{R}^2 : v \cdot w > 0\}.$$

H_v is the connected component of $\mathbb{R}^2 - v^\perp$ containing v itself, $v^\perp$ being the orthogonal complement of v:

$$v^\perp = \{w \in \mathbb{R}^2 : v \cdot w = 0\}$$

Write $C_{v,w}$ for the open cone spanned by vectors $v, w \in \mathbb{R}^2$, $v \neq w$:

$$C_{v,w} = \{av + bw : a, b > 0\}$$

(1) Let $u, v, w \in \mathbb{R}^2$ be three distinct vectors. Prove that the following statements are equivalent:
  (a) $-u \in C_{v,w}$
  (b) $-u \in C_{v,w}$, $-w \in C_{u,v}$ and $-v \in C_{w,u}$
  (c) $H_u \cap H_v \cap H_w = \varnothing$
  (d) $H_u \cup H_v \cup H_w = \mathbb{R}^2$

  If you're having trouble writing up a formal argument here, you can draw me a convincing, pretty diagram instead. (If you write up a proof, you can optionally include a picture!)

(2) Let $x \in \mathbb{R}$. Show that the sets

$$\left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \lim_{t\to\infty} \sigma(ta + tbx) = 0 \right\} \quad \text{and} \quad \left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \lim_{t\to\infty} \big(1 - \sigma(ta + tbx)\big) = 0 \right\}$$

  have the form $H_{\pm v}$, where $v = \begin{bmatrix} 1 \\ x \end{bmatrix}$.

(3) For $K > 0$, identify the level curves

$$\left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \sigma(a + bx) = K \right\} \quad \text{and} \quad \left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \big(1 - \sigma(a + bx)\big) = K \right\}$$

  as lines and the level sets

$$\left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \sigma(a + bx) \leq K \right\} \quad \text{and} \quad \left\{ \begin{bmatrix} a \\ b \end{bmatrix} : \big(1 - \sigma(a + bx)\big) = K \right\}$$

  as half-plane translates of the form $y + H_v$ and $z + H_w$.

(4) Consider a three point data set $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$, where $x_1 < x_2 < x_3$ and let

$$L(a, b) = \prod_{i=1}^{3} p(x_i, y_i | a, b),$$

  where

$$p(x_i, y_i | a, b) = \sigma(a + bx_i)^{y_i} (1 - \sigma(a + bx_i)^{y_i})^{1-y_i}.$$

  Show that

$$\lim_{a^2+b^2\to\infty} L(a, b) = 0$$

if and only if $y_1 = y_3 = 1$ and $y_2 = 0$ or $y_1 = y_3 = 0$ and $y_2 = 0$. For $0 < K < 1$, consider the sets

$$S_{i,K} = \{(a, b) : p_i(x_i, y_i|a, b) > K\}, \qquad i = 1, 2, 3.$$

Prove that $S_{1,K} \cap S_{2,K} \cap S_{3,K}$ is bounded if and only if either $y_1 = y_3 = 0$ and $y_2 = 1$ or $y_1 = y_3 = 1$ and $y_2 = 0$.

Let

$$v_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}, \qquad i = 1, 2, 3.$$

Prove that

$$L(a, b)$$

(a) Suppose $y_1 = y_3 = 0$ and $y_2 = 1$.

Consider fitting a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ to the following simplified logistic regression model:

$$Y_i|X = x_i \sim p(x_i, y_i) := \sigma(bx_i)^{y_i}(1 - \sigma(bx_i))^{1-y_i}$$

Here, σ denotes the sigmoid function: $\sigma(t) = (1+e^{-t})^{-1}$. Let $L(b)$ the the likelihood function associated this dataset/model.

Suppose that our dataset has the following property:

$$(*) \qquad x_i < 0 \text{ if and only if } y_i = 0$$

Prove that

$$\lim_{b\to\infty} L(b) = 1.$$

What can we say about $\operatorname{argmax}_b L(b)$ in this case?

Solution:

Suppose that $x_i < 0$ and $y_i = 0$. Then $e^{-bx_i} \to \infty$ as $b \to \infty$. It follows that

$$\lim_{b\to\infty} p(x_i, y_i) = \lim_{b\to\infty} \big(1 - \sigma(bx_i)\big) = 1 - 0 = 1.$$

If $x_i > 0$ and $y_i = 1$, then $e^{-bx_i} \to 0$ as $b \to \infty$. It follows that

$$\lim_{b\to\infty} p(x_i, y_i) = \lim_{b\to\infty} \sigma(bx_i) = 1.$$

Therefore, by property $(*)$,

$$\lim_{b\to\infty} p(x_i, y_i) = 1$$

for all $i = 1, \ldots, n$. Consequently,

$$\lim_{b\to\infty} L(b) = \lim_{b\to\infty} \prod_{i=1}^{n} p(x_i, y_i) = 1.$$

Suppose, now, that there are indices i and j such that

$$x_i < 0 \text{ and } y_i = 1 \quad \text{and} \quad x_j > 0 \text{ and } y_j = 0.$$

In particular, property $(*)$ is not satisfied. Prove that

$$\lim_{b\to\pm\infty} L(b) = 0.$$

What can we say about $\operatorname{argmax}_b L(b)$ in this case?

Solution: As $y_i = 0$, $p(x_i, y_i|b) = \sigma(bx_i)$. As $x_i < 0$, $\sigma(bx_i) \to 0$ as $b \to \infty$. Thus

$$\lim_{b\to\infty} p(x_i, y_i|b) = \lim_{b\to\infty} \sigma(bx_i) = 0.$$

As $y_j = 1$, $p(x_j, y_j|b) = 1 - \sigma(bx_j)$. As $x_j > 0$, $1 - \sigma(bx_j) \to 0$ as $b \to \infty$. Thus

$$\lim_{b\to\infty} p(x_j, y_j|b) = \lim_{b\to\infty} \sigma(bx_j) = 0.$$

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random samples from normally distributed populations with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. Let S_X^2 and S_Y^2 be the standard unbiased estimators of σ_X^2 and σ_Y^2, respectively.

(1) Suppose $\sigma_X^2 = \sigma_Y^2$ and write σ^2 for this common value.

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

is an unbiased estimator of σ^2. It's called the *pooled variance estimator*.

Solution: Since S_X^2 and S_Y^2 are unbiased estimators of σ_X and σ_Y, respectively,

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{m-1}{m+n-2}\,\mathbb{E}[S_X^2] + \frac{n-1}{m+n-2}\,\mathbb{E}[S_Y^2] \\
&= \frac{m-1}{m+n-2}\sigma_X^2 + \frac{n-1}{m+n-2}\sigma_Y^2 \\
&= \frac{m-1}{m+n-2}\sigma^2 + \frac{n-1}{m+n-2}\sigma^2 \qquad\qquad (\text{as } \sigma_X^2 = \sigma^2 = \sigma_Y^2) \\
&= \sigma^2
\end{aligned}$$

(2) Suppose, in addition to having common variance, that the X_i are independent of the Y_i. What is the distribution of

$$\frac{(m+n-2)S^2}{\sigma^2}?$$

What is the variance of S^2?

Solution: Since S_X^2 and S_Y^2 are independent,

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2_{m-1} \quad \text{and} \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1},$$

and it follows from general properties of χ^2-distributions that

$$\frac{(m+n-2)S^2}{\sigma^2} = \frac{1}{\sigma^2}\big((m-1)S_X^2 + (n-1)S_Y^2\big) \sim \chi^2_{m+n-2}.$$

Since χ^2_{m+n-2}-distributed random variable has variance $2(m+n-2)$, it follows that

$$\operatorname{Var} S^2 = \frac{2\sigma^4}{m+n-2}.$$

(3*) Generalize these results from the case of $K = 2$ populations to that of an arbitrary K. Compare with equation (4.15) in [2].

(4*) Can you prove analogous results with covariance matrices in place of scalar variances?

Since S_X^2 and S_Y^2 are independent,

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2_{m-1} \quad \text{and} \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1},$$

and it follows from general properties of χ^2-distributions that

$$\frac{(m+n-2)S^2}{\sigma^2} = \frac{1}{\sigma^2}\big((m-1)S_X^2 + (n-1)S_Y^2\big) \sim \chi^2_{m+n-2}.$$

Since χ^2_{m+n-2}-distributed random variable has variance $2(m+n-2)$, it follows that

$$\operatorname{Var} S^2 = \frac{2\sigma^4}{m+n-2}.$$

[1, Exercise 12.16] This exercise examines an extreme case in which the likelihood equations for logistic regression have no solution.

Consider the following 20-point data set:

$$(0,1),(0,1),(0,1),(0,1),(0,1),(0,1),(0,1),(0,1),(0,1),(0,1)$$
$$(1,0),(1,0),(1,0),(1,0),(1,0),(1,1),(1,1),(1,1),(1,1),(1,1)$$

(1) Observe that, empirically, $\mathrm{Prob}(Y = 1|X = 0) = 1$ and $\mathrm{Prob}(Y = 1|X = 1) = 0.5$. Let $\sigma(t) = (1 + e^{-t})^{-1}$ be the sigmoid function. Are there a and b such that $\sigma(a + b \cdot 0) = 1$ or $\sigma(a + b \cdot 1) = 0.5$?

Solution:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Since the exponential function takes values in $(0, \infty)$,

$$0 < \sigma(x) < 1,$$

for every $x \in \mathbb{R}$. In particular, there are no numbers a and b such that $\sigma(a+b\cdot 0) = 1$. Clearly, $\sigma(0) = \sigma(0 + 0 \cdot 1) = 0.5$.

(2) Let $\mathscr{L}(a, b)$ be the likelihood function associated to fitting a logistic regression model to this data set. Show that

$$L := \lim_{b\to\infty} \mathscr{L}(-b, b) = \sup_{(a,b)\in\mathbb{R}^2} \mathscr{L}(a, b) < \infty$$

and that $\mathscr{L}(a,b) \neq L$ for any $(a,b) \in \mathbb{R}^2$. What are

$$\lim_{b\to\infty} \sigma(-b + b \cdot 0) \quad \text{and} \quad \lim_{b\to\infty} \sigma(-b + b \cdot 1)?$$

Let $(\boldsymbol{X}, Y)$ be jointly distributed, where $\boldsymbol{X}$ is a p-dimensional random vector and Y takes values in $\{1, \ldots, K\}$. Suppose that, for each k, $\boldsymbol{X}|Y = k$ has Gaussian distribution with mean $\boldsymbol{\mu}_k$ and and variance Σ, with the latter independent of k.

Consider a data set $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^{1\times p}$ and $y^{(i)} \in \{1, \ldots, K\}$. For $1 \leq k \leq K$, let

$$I_k = \{i : y^{(i)} = k\}, \quad n_k = |I_k|, \quad \widehat{\pi}_k = \frac{n_k}{n}.$$

Define sample means $\boldsymbol{\mu}_k$ and a pooled sample covariance $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\mu}}_k = \widehat{\boldsymbol{\mu}}_{k,\boldsymbol{x}} = \frac{1}{n_k} \sum_{i\in I_k} \boldsymbol{x}^{(i)} \in \mathbb{R}^{p\times 1},$$

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i\in I_k} (\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_k)^T (\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_k) \in \mathbb{R}^{p\times p}.$$

Define linear discriminant functions, $\delta_k = \delta_{k,\boldsymbol{x}}$, by

$$\delta_k(\boldsymbol{v}) = \delta_{k,\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{v}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\mu}}_k^T - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\mu}}_k^T + \log \widehat{\pi}_k, \quad \boldsymbol{v} \in \mathbb{R}^{p\times 1}.$$

Let $\boldsymbol{a} \in \mathbb{R}^{p\times 1}$ and let

$$\boldsymbol{w}^{(i)} = \boldsymbol{x}^{(i)} - \boldsymbol{a}.$$

$$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{w}} = \widehat{\boldsymbol{\mu}}_{k,\boldsymbol{x}} - \boldsymbol{a}, \quad \Sigma_w = \Sigma_x$$

$$\delta_{k_1,w}(v-a) - \delta_{k_2,w}(v-a) = \delta_{k_1,x}(v) - \delta_{k_2,x}(v)$$

Let $U \in \mathbb{R}^{p\times p}$ be an orthogonal matrix and let $w^{(i)} = Ux^{(i)}$. Then

$$\delta_{k,Ux}(Uv) = \delta_x(v).$$

$$\sum_{k=1}^{K} \pi_k \mu_k = \mu$$

Suppose

$$Y \sim \text{Categorical}\left(\tfrac{1}{K}, \ldots, \tfrac{1}{K}\right),$$
$$\boldsymbol{X}|Y = k \sim N(\boldsymbol{\mu}_k, \sigma I).$$

Show that

$$\operatorname*{argmin}_k p(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \operatorname*{argmin}_k \|\boldsymbol{x} - \boldsymbol{\mu}_k\|.$$

Logistic regression (with and without ridge regularization, with and without PCA), LDA, Gaussian naïve Bayes, for breast cancer data set. Plot in 2d with decision boundary. Optional: Lasso

Document classification with multinomial naïve Bayes

Ridge regression via constrained optimization.

1. Total variation = variance within + variance between

Let X and Y be jointly distributed random variables, where

$$\begin{aligned} Y &\sim \text{Categorical}(\pi_1, \ldots, \pi_K), & \sum \pi_k &= 1, \\ X \mid Y = k &\sim N(\mu_k, \sigma_k^2), & k &= 1, \ldots, K. \end{aligned}$$

Assume that the μ_k are pairwise distinct. Prove that

$$\mathbb{E}[X] = \sum_{k=1}^{K} \pi_k \mu_k$$

and that

$$\operatorname{Var} X = \sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k (\mu_k - \mu)^2.$$

To establish the decomposition of the variance, you might want to use the *law of total variation:*

$$\operatorname{Var} X = \mathbb{E}(\operatorname{Var}(X|Y)) + \operatorname{Var} \mathbb{E}(X|Y).$$

The random variable X has marginal density $p(x)$, where

$$p(x) = \sum_{k=1}^{K} p(x, k) = \sum_{k=1}^{K} p(k) p(x|k) = \sum_{k=1}^{K} \pi_k p(x|k).$$

Therefore,

$$\mathbb{E}[X] = \sum_{k=1}^{K} \pi_k \, \mathbb{E}[X_k]$$

where X_k is a random variable with density $p(x|k)$. By hypothesis, X_k has expected value μ_k. Therefore,

$$\mathbb{E}[X] = \sum_{k=1}^{K} \pi_k \mu_k.$$

Since $\operatorname{Var}(X|Y = k) = \sigma_k^2$,

$$\mathbb{E}[\operatorname{Var}(X|Y)] = \sum_{k=1}^{K} p(Y = k) \operatorname{Var}(X|Y = k) = \sum_{k=1}^{K} \pi_k \sigma_k^2.$$

Let $\mu = \mathbb{E}[X]$. By the *law of total expectation*,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] = \mu.$$

Let $Z = \mathbb{E}[X|Y]$. Then set of values of Z is $\{\mu_1, \ldots, \mu_K\}$. Since the μ_k are pairwise distinct,

$$\operatorname{Prob}[Z = \mu_k] = p(Y = k) = \pi_k.$$

Therefore,

$$\operatorname{Var} Z = \sum_{k=1}^{K} \operatorname{Prob}[Z = \mu_k](\mu_k - \mu)^2 = \sum_{k=1}^{K} \pi_k \big(\mu_k - \mu\big)^2.$$

Let $x_1, \ldots, x_n$ be numbers and set

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Suppose that

$$\{1, \ldots, n\} = \bigsqcup_{k=1}^{K} I_k \qquad \text{(disjoint union)}$$

with

$$n_k := |I_k| > 0.$$

Set

$$\mu_k = \frac{1}{n_k} \sum_{i \in I_k}^{n} x_i, \qquad \sigma_k^2 = \frac{1}{n_k} \sum_{i \in I_k}^{n} (x_i - \mu_k)^2$$

Prove that

$$\sigma^2 = \sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k (\mu_k - \mu)^2, \quad \text{where} \quad \pi_k = \frac{n_k}{n}.$$

(This is an "algebraic version" of the law of total variance)

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}(x_i-\mu)^2 \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}(x_i-\mu_k+\mu_k-\mu)^2 \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\left\{(x_i-\mu_k)^2+(\mu_k-\mu)^2+2(x_i-\mu_k)(\mu_k-\mu)\right\} \\
&= \frac{1}{n}\sum_{k=1}^{K}n_k\frac{1}{n_k}\sum_{i\in I_k}(x_i-\mu_k)^2+\frac{1}{n}\sum_{k=1}^{K}(\mu_k-\mu)^2\sum_{i\in I_k}1 \\
&\qquad +\frac{2}{n}\sum_{k=1}^{K}(\mu_k-\mu)\sum_{i\in I_k}(x_i-\mu_k) \\
&= \sum_{k=1}^{K}\frac{n_k}{n}\sigma_k^2+\sum_{k=1}^{K}\frac{n_k}{n}(\mu_k-\mu)^2+0 \\
&= \sum_{k=1}^{K}\pi_k\sigma_k^2+\sum_{k=1}^{K}\pi_k(\mu_k-\mu)^2
\end{aligned}
$$

Define a probability space (Ω,μ) by

$$
\Omega=\{x_1,\ldots,x_n\}\times\{1,\ldots,K\},\quad \mu\big(\{(x_i,k)\}\big)=p_{ik}:=\begin{cases}\dfrac{1}{n} & \text{if } i\in I_k,\\ 0 & \text{otherwise.}\end{cases}
$$

Note that μ is, indeed, a probability measure on Ω:

$$
\sum_{(i,k)\in\Omega}p_{ik}=\sum_{k=1}^{K}\sum_{i=1}^{n}p_{ik}=\sum_{k=1}^{K}\sum_{i\in I_k}\frac{1}{n}=\sum_{k=1}^{n}\frac{n_k}{n}=1
$$

Define random variables X and Y on Ω by

$$
X(x_i,k)=x_i \quad\text{and}\quad Y(x_i,k)=k.
$$

Then

$$
\mu=\mathbb{E}[X]=\sum_{k=1}^{K}\sum_{i=1}^{n}p_{ik}x_i=\sum_{k=1}^{K}\sum_{i\in I_k}\frac{1}{n}x_i=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}x_i=\frac{1}{n}\sum_{i=1}^{n}x_i.
$$

Let $Z=\mathbb{E}[X|Y]$, so that

$$
Z(k)=\mathbb{E}[X|Y=k]=\frac{\sum_{i=1}^{n}p_{ik}x_i}{\sum_{i=1}^{n}p_{ik}}=\frac{\sum_{i\in I_k}x_i}{\sum_{i\in I_k}^{n}1}=\frac{1}{n_k}\sum_{i\in I_k}x_i=\mu_k.
$$

$$
\begin{aligned}
\operatorname{Var}[X|Y=k] &= \mathbb{E}[(X-Z)^2|Y=k] \\
&= \frac{\sum_{i=1}^n p_{ik}(X(i)-Z(k))^2}{\sum_{i=1}^n p_{ik}} \\
&= \frac{\frac{1}{n}\sum_{i\in I_k}(x_i-\mu_k)^2}{\sum_{i\in I_k}\frac{1}{n}} \\
&= \frac{1}{n_k}\sum_{i\in I_k}(x_i-\mu_k)^2 \\
&= \sigma_k^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\operatorname{Var}[X|Y]\right] &= \sum_{k=1}^K \operatorname{Prob}(Y=k)\operatorname{Var}[X|Y=k] \\
&= \sum_{k=1}^K \left(\sum_{(i,k)} p_{ik}\right)\sigma_k^2 \\
&= \sum_{k=1}^K \frac{n_k}{n}\sigma_k^2 \\
&= \sum_{k=1}^K \pi_k\sigma_k^2
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\operatorname{Var} Z &= \mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right] \\
&= \mathbb{E}\left[(Z-\mu)^2\right] && \text{(by the law of total expectation)} \\
&= \sum_{k=1}^K \operatorname{Prob}(Z=\mu_k)(\mu_k-\mu)^2 \\
&= \sum_{k=1}^K \operatorname{Prob}(Y=k)(\mu_k-\mu)^2 && (\text{as } k=\ell \iff \mu_k=\mu_\ell) \\
&= \sum_{k=1}^K \left(\sum_{(i,k)} p_{ik}\right)(\mu_k-\mu)^2 \\
&= \sum_{k=1}^K \pi_k(\mu_k-\mu)^2.
\end{aligned}
$$

1.1. **Matrix version.**

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}(x_i-\mu)(x_i-\mu)^T \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}((x_i-\mu_k)+(\mu_k-\mu))((x_i-\mu_k)+(\mu_k-\mu))^T \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\big\{(x_i-\mu_k)(x_i-\mu_k)^T+(\mu_k-\mu)(\mu_k-\mu)^T \\
&\qquad\qquad + (x_i-\mu_k)(\mu_k-\mu)^T+(\mu_k-\mu)(x_i-\mu_k)^T\big\} \\
&= \frac{1}{n}\sum_{k=1}^{K}n_k\frac{1}{n_k}\sum_{i\in I_k}(x_i-\mu_k)(x_i-\mu_k)^T+\frac{1}{n}\sum_{k=1}^{K}(\mu_k-\mu)(\mu_k-\mu)^T\sum_{i\in I_k}1 \\
&\qquad\qquad + \frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{i\in I_k}(x_i-\mu_k)\right\}(\mu_k-\mu)^T \\
&\qquad\qquad + \frac{1}{n}\sum_{k=1}^{K}(\mu_k-\mu)\sum_{i\in I_k}(x_i-\mu_k)^T \\
&= \sum_{k=1}^{K}\frac{n_k}{n}\Sigma_k^2+\sum_{k=1}^{K}\frac{n_k}{n}(\mu_k-\mu)(\mu_k-\mu)^T+0+0 \\
&= \sum_{k=1}^{K}\pi_k\Sigma_k^2+\sum_{k=1}^{K}\pi_k(\mu_k-\mu)(\mu_k-\mu)^T
\end{aligned}
$$

References

[1] Casella, Berger, *Statistical Inference (2nd ed.)*, Duxbury, 2002.
[2]