Problem 0. *Binary logistic regression*
Consider a two-class classification problem. The training data consists of n independent and identically distributed observations $(x_1, y_1), \ldots, (x_n, y_n)$, where each $y_i \in \{0, 1\}$ and $x_i \in \mathbb{R}^d$. We consider classification made using logistic regression. The posterior probabilities $\mathbf{P}(Y = k \mid X = x)$ for $k = 0, 1$ are modelled as follows,

$$\log\left(\frac{\mathbf{P}(Y = 1 \mid X = x)}{\mathbf{P}(Y = 0 \mid X = x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \beta^t x\,,$$

where $\beta^t = (\beta_1, \ldots, \beta_d)$ and $x^t = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

(a) Show that

$$\mathbf{P}(Y = 1 \mid X = x) = \sigma(\beta_0 + \beta^t x) = 1 - \mathbf{P}(Y = 0 \mid X = x)\,,$$

where $\sigma(u) = e^u/(1 + e^u)$ is the sigmoid function.

(b) Coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated using maximum likelihood. Consider the log likelihood function,

$$\ell(\beta_0, \beta) := \log\left(\prod_{i=1}^{n} p(y_i \mid x_i, \beta_0, \beta)\right)\,,$$

where we used the convenient notation $p(y_i \mid x_i, \beta_0, \beta) = \mathbf{P}(Y = y_i \mid X = x_i)$. Show that

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} y_i \log \sigma_i + (1 - y_i) \log(1 - \sigma_i)\,,$$

where we defined $\sigma_i := \sigma(\beta_0 + \beta^t x_i)$.

(c) Show that for $i = 1, \ldots, n$ and $j = 0, \ldots, d$,

$$\frac{\partial \log \sigma_i}{\partial \beta_j} = x_{ij}(1 - \sigma_i)\,,$$

and

$$\frac{\partial \log(1 - \sigma_i)}{\partial \beta_j} = -x_{ij}\sigma_i\,,$$

where $x_{i0} \equiv 1$ for all $i = 1, \ldots, n$. Deduce that

$$\frac{\partial \ell(\beta_0, \beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \sigma_i) x_{ij}\,.$$

(d) Deduce from question *(c)* that the gradient $\nabla_{\beta_0,\beta}\ell(\beta_0, \beta)$ of ℓ with respect to (β_0, β) can be written in the matrix form

$$\nabla_{\beta}\ell(\beta_0, \beta) = X^t(y - \sigma)\,,$$

where X is an $n \times (d + 1)$ matrix, and σ and y are column vectors that you specify.

(e) Show that for $j, k = 0, \ldots, d$,

$$\frac{\partial \ell(\beta_0, \beta)}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} \sigma_i (1 - \sigma_i)\,.$$

Deduce that the Hessian can be written as

$$\nabla^2_\beta \ell(\beta_0, \beta) = -X^t W X\,,$$

for a matrix W that you will specify.

(f) Put $b := (\beta_0, \beta)$, and $\hat{b} := (\hat{\beta}_0, \hat{\beta})$, the maximum likelihood estimator of b. Deduce from the previous questions the asymptotic distribution of $n^{1/2}(\hat{b} - b)$.

(g) Recall what Newton method for unconstrained minimisation problems is. Write down a generic expression for Newton algorithm.

(h) We numerically solve $\nabla_\beta \ell(\beta_0, \beta) = 0$ using Newton method. Show that a single step in Newton algorithm can be written

$$\tilde{\beta}^{(t+1)} = (X^t W X)^{-1} X^t W z^{(t)}\,,$$

where $z^{(t)}$ denotes the adjusted response, function of the current parameter estimates $\tilde{\beta}^{(t)}$. Give the expression of $z^{(t)}$.

(i) Deduce from *(h)* why Newton algorithm for logistic regression is commonly referred to as an iterative reweighed least square algorithm.

Problem 1. *Optimal Linear Risk*

The risk of a fixed binary classifier f under the 0/1 loss ℓ_0 is

$$\mathcal{R}(f) = \mathbf{E}\{\ell_0(Y,\, f(X))\} = \mathbf{P}(Y \neq f(X))\,, \tag{1}$$

where $Y \in \{0, 1\}$ and $X \in \mathbb{R}^d$. Given $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$, a linear classifier $f_{\beta_0,\beta}$ is such that

$$f_{\beta_0,\beta}(x) = \begin{cases} 1 & \text{if } \beta_0 + \beta^t x \geq 0 \\ 0 & \text{if } \beta_0 + \beta^t x < 0 \end{cases}$$

The optimal linear risk $\bar{R}$ is defined by $\bar{R} = \inf_{\beta_0,\beta} \mathcal{R}(f_{\beta_0,\beta})\,.$

(i) Suppose in questions *(i)*, *(ii)* and *(iii)* that X is univariate. For $y' \in \{0, 1\}$ and $x' \in \mathbb{R}$, we define a linear discrimination rule as

$$f_{x',y'}(x) = \begin{cases} y' & \text{if } x \leq x' \\ 1 - y' & \text{if } x > x'\,. \end{cases}$$

According to (1), the goal is the find the values of x' and y' which minimise the misclassification error,

$$(x^*, y^*) = \arg\min_{(x',y')} \mathbf{P}(Y \neq f_{x',y'}(X))\,.$$

Suppose that $\mathbf{P}(Y=1)=p=1-\mathbf{P}(Y=0)$, $X|Y=j \sim F_j$, $m_j=\mathbf{E}(X|Y=j)$ and $\sigma_j^2=\mathrm{var}(X|Y=j)$. Check that the optimal linear risk can be written

$$\bar{R}=\inf_{(x',y')} \mathbf{1}_{\{y'=0\}}\left\{pF_1(x')+(1-p)(1-F_0(x'))\right\}+\mathbf{1}_{\{y'=1\}}\left\{p(1-F_1(x'))+(1-p)F_0(x')\right\} .$$

(ii) Prove the Chebyshev-Cantelli inequality, which states that for any $u \geq 0$,

$$\mathbf{P}(X-\mathbf{E}X>u) \leq \frac{\mathrm{var}(X)}{\mathrm{var}(X)+u^2} .$$

Argue that a similar inequality holds for $\mathbf{P}(X-\mathbf{E}X \leq -u)$.

(iii) Deduce from *(i)* and *(ii)* that

$$\bar{R} \leq \left(1+\frac{(m_0-m_1)^2}{(\sigma_0+\sigma_1)^2}\right)^{-1} .$$

(iv) Generalise the upper bound derived in *(iii)* for multivariate feature points $X \in \mathbb{R}^d$.

Problem 2. *Probit regression*
We consider the problem of two-class classification using probit regression. It will be convenient to code the two classes associated with $x_i \in \mathbb{R}^d$ with 0/1 responses y_i. Under the probit model,

$$p(y_i \mid x_i, \beta)=\Phi_i^{y_i}(1-\Phi_i)^{1-y_i} ,$$

where $\Phi_i=\Phi(\beta_0+\beta^t x_i)$, and Φ is the standard normal cdf. Let $(x_1,y_1),\ldots,(x_n,y_n)$ be our learning sample.

(i) Give an interpretation of the probit model using a latent variable formulation, similar to the one presented on page 8, Chapter 5 of the lecture notes.

(ii) Write down the log-likelihood $\ell(\beta)$.

(iii) Show that

$$\frac{\partial \ell(\beta_0,\beta)}{\partial \beta_j}=\sum_{i=1}^n \frac{\phi_i(y_i-\Phi_i)}{\Phi_i(1-\Phi_i)} x_{ij} ,$$

where $\phi_i := \phi(\beta_0+\beta^t \mathbf{x}_i)$, with ϕ the standard normal pdf, and $x_{i0} \equiv 1$ for all $i=1,\ldots,n$.

(iv) Show that

$$\frac{\partial \ell(\beta_0,\beta)}{\partial \beta_k \partial \beta_j}=-\sum_{i=1}^n x_{ij}x_{ik}\phi_i\left(y_i\frac{\phi_i+(\beta_0+\beta^t x_i)\Phi_i}{\Phi_i^2}+(1-y_i)\frac{\phi_i-(\beta_0+\beta^t x_i)(1-\Phi_i)}{(1-\Phi_i)^2}\right) .$$

(v) Deduce from *(iii)* that the Fisher information matrix is $I = (I_{jk})$, with

$$I_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \frac{\phi_i^2}{\Phi_i(1-\Phi_i)} \,,$$

and re-express the right-hand side in matrix form. Deduce the expression of the asymptotic covariance matrix of the maximum likelihood estimator and give an estimate of it.

Problem 3. *Multiclass logistic regression*
We consider logistic regression with $K > 2$ classes. We use the notation

$$\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,(K-1)})^t \in \mathbb{R}^{K-1} \,,$$

where response $y_{i,k} = 1$ if observation i belongs to class k, for $k = 1, \ldots, K-1$, and 0 otherwise. The i-th input vector is denoted $\mathbf{x}_i = (x_{i,0}, \ldots, x_{i,d})^t \in \mathbb{R}^{d+1}$, for $i = 1, \ldots, n$, with $x_{i,0} = 1$. Let $\beta_k = (\beta_{k,0}, \ldots, \beta_{k,d})^t \in \mathbb{R}^{d+1}$ be the parameter vector corresponding to class k, for $k = 1, \ldots, K-1$. Finally, put $\theta := (\beta_1^t, \ldots, \beta_{K-1}^t)^t$.

(i) Recall the expression of the posterior probabilities

$$\mathbf{P}(Y = k \mid \mathbf{X} = \mathbf{x}, \theta)$$

under the multi-class logistic regression model.

(ii) Show that the log-likelihood can be written as

$$\ell(\theta) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K-1} y_{i,j}\, \beta_j^t\, \mathbf{x}_i - \log\left(1 + \sum_{\ell=1}^{K-1} \exp(\beta_\ell^t\, \mathbf{x}_i)\right) \right\} .$$

We introduce further notation: for $1 \le k \le K-1$,

$$\begin{aligned}
\mathbf{z}_k &:= (y_{1,k},\, \ldots,\, y_{n,k})^t \in \mathbb{R}^n \\
\mathbf{p}_k &:= (\mathbf{P}(Y = k \mid \mathbf{x}_1),\, \ldots,\, \mathbf{P}(Y = k \mid \mathbf{x}_n))^t \in \mathbb{R}^n \,,
\end{aligned}$$

and

$$\mathbf{X} := \begin{pmatrix} \cdots & \mathbf{x}_1^t & \cdots \\ & \vdots & \\ \cdots & \mathbf{x}_n^t & \cdots \end{pmatrix} \in \mathbb{R}^{n\times(d+1)} \,, \qquad \mathcal{X}^t := \begin{pmatrix} \mathbf{X}^t & & \\ & \ddots & \\ & & \mathbf{X}^t \end{pmatrix} \in \mathbb{R}^{(K-1)(d+1)\times(K-1)n} \,,$$

where the matrix $\mathcal{X}^t$ is a $(K-1)\times(K-1)$ diagonal bloc matrix, with diagonal blocs $\mathbf{X}^t$.

(iii) Show that the gradient of $\ell(\theta)$ is given by

$$\nabla_\theta \ell(\theta) = \mathcal{X}^t \begin{pmatrix} \mathbf{z}_1 - \mathbf{p}_1 \\ \vdots \\ \mathbf{z}_{K-1} - \mathbf{p}_{K-1} \end{pmatrix} .$$

(iv) Show that the Hessian can be written in the form

$$\nabla^2_\theta \ell(\theta) = -\mathcal{X}^t \mathbf{W} \mathcal{X}\,,$$

where $\mathbf{W}$ is a *non-diagonal* bloc matrix. Show that $\mathbf{W}$ can be expressed in terms of $K-1$ diagonal matrices $\mathbf{Q}_l \in \mathbb{R}^{n\times n}$, $l = 1, \ldots, K-1$, and $K-1$ diagonal matrices $\mathbf{R}_l \in \mathbb{R}^{n\times n}$, $l = 1, \ldots, K-1$,

$$\mathbf{W} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{R}_1\mathbf{R}_2 & \ldots & \\ \mathbf{R}_2\mathbf{R}_1 & \mathbf{Q}_2 & \ldots & \\ \vdots & \vdots & \ddots & \\ & & & \mathbf{Q}_n \end{pmatrix}.$$

Give the expression of the matrices $\mathbf{Q}_l$ and $\mathbf{R}_l$.

(v) We use a Newton procedure to iteratively minimise the log-likelihood. Show that at each iteration, we are solving a new non-diagonal weighted least square problem. Specify the value of the working response and the weight matrix.

Problem 3.

Suppose that within each class $\{1, \ldots, K\}$, the data follow a multinomial distribution. Specifically, the i-th observation (X_i, Y_i) is such that

$$\mathbf{P}(X_i = \mathbf{x}_i \mid Y_{ik} = 1) = \frac{x_i!}{x_{i1}! \ldots x_{im}!}\, p_{k1}^{x_{i1}} \ldots p_{km}^{x_{im}}\,, \quad k = 1, \ldots, K\,,$$

where $\mathbf{x}_i := (x_{i1}, \ldots .x_{im})$, $x_i := \sum_l x_{il}$, $\sum_l p_{kl} = 1$, and $Y_i := (Y_{i1}, \ldots, Y_{iK})$, where $Y_{ik} = 1$ if observation X_i is in class k, and 0 otherwise. Put $\pi_k = \mathbf{P}(Y_{ik} = 1)$. Our goal is to predict the class of a new observation, based on the model above.

(i) How many parameters do we need to estimate?

(ii) Write down the log-likelihood associated with a training sample $\mathcal{L}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ of size n.

(iii) Derive the maximum likelihood estimator (MLE) for each parameter of the model.

(iv) What happens for categories with zero count? Suggest an easy modification of the MLE which takes care of this problem.