Exercise 1 (Logistic regression)

Assume we are given a data set $\mathscr{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ of independent random variables distributed as the generic pair $(\mathbf{X}, Y)$ where $\mathbf{X}$ takes values in $\mathbb{R}^d$ and $Y \in \{0, 1\}$. We consider the following logistic model

$$\log\left(\frac{\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]}{\mathbb{P}[Y=0|\mathbf{X}=\mathbf{x}]}\right) = \mathbf{x}^T\beta\,.$$

1. Prove that

$$\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}] = \frac{e^{\mathbf{x}^T\beta}}{1+e^{\mathbf{x}^T\beta}}$$

2. Since the data are assumed to be independent, the likelihood writes

$$\mathscr{L}(\beta) = \prod_{i=1}^n \mathbb{P}[Y=y_i|X=x_i].$$

   Prove that the negative log likelihood f can be written as

$$f(\beta) = \sum_{i=1}^n \log\Big(1+\exp\left(-\tilde{y}_i\mathbf{x}_i^T\beta\right)\Big),$$

   where $\tilde{y}_i = 2y_i - 1$.

3. Prove that the gradient of the negative log likelihood satisfies

$$\nabla f(\beta) = -\frac{1}{n}\sum_{i=1}^n \frac{y_i e^{-\tilde{y}_i\mathbf{x}_i^T\beta}}{1+e^{-\tilde{y}_i\mathbf{x}_i^T\beta}}\mathbf{x}_i.$$

4. Prove that the Hessian matrix H of the negative log likelihood satisfies

$$H = \sum_{i=1}^n \frac{e^{-\tilde{y}_i\mathbf{x}_i^T\beta}}{\left(1+e^{-\tilde{y}_i\mathbf{x}_i^T\beta}\right)^2}\mathbf{x}_i\mathbf{x}_i^T.$$

5. Prove that the function f is convex. Recall that if the Hessian matrix H of f is positive, that is, for all $\mathbf{z} \in \mathbb{R}^d$,

$$\mathbf{z}^T H\mathbf{z} \geq 0,$$

   then the function f is convex.

Solution of exercise 1:

1. We have

$$\begin{aligned}
&\log\left(\frac{\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]}{1-\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]}\right)=\mathbf{x}^T\beta \\
\Leftrightarrow &\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]=\Big(1-\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]\Big)e^{\mathbf{x}^T\beta} \\
\Leftrightarrow &\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}]=\frac{e^{\mathbf{x}^T\beta}}{1+e^{\mathbf{x}^T\beta}}.
\end{aligned}$$

2. Note that

$$\begin{aligned}
\mathscr{L}(\beta)&=\prod_{i=1}^n \mathbb{P}[Y=y_i|\mathbf{X}=\mathbf{x}_i] \\
&=\prod_{i=1}^n \Big(\mathbb{P}[Y=1|\mathbf{X}=\mathbf{x}_i]\mathbb{1}_{y_i=1}+\mathbb{P}[Y=0|\mathbf{X}=\mathbf{x}_i])\mathbb{1}_{y_i=0}\Big) \\
&=\prod_{i=1}^n \Big(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\mathbb{1}_{y_i=1}+\frac{1}{1+e^{\mathbf{x}_i^T\beta}}\mathbb{1}_{y_i=0}\Big) \\
&=\prod_{i=1}^n \Big(\frac{e^{\mathbf{x}_i^T\beta}}{1+e^{\mathbf{x}_i^T\beta}}\mathbb{1}_{y_i=1}+\frac{e^{-\mathbf{x}_i^T\beta}}{1+e^{-\mathbf{x}_i^T\beta}}\mathbb{1}_{y_i=0}\Big) \\
&=\prod_{i=1}^n \Big(\frac{e^{(2y_i-1)\mathbf{x}_i^T\beta}}{1+e^{(2y_i-1)\mathbf{x}_i^T\beta}}\mathbb{1}_{y_i=1}\Big) \\
&=\prod_{i=1}^n \left(\frac{1}{1+e^{-(2y_i-1)\mathbf{x}_i^T\beta}}\right).
\end{aligned}$$

Consequently, the negative log likelihood satisfies

$$\begin{aligned}
f(\beta)&=-\log\left(\prod_{i=1}^n\left(\frac{1}{1+e^{-(2y_i-1)\mathbf{x}_i^T\beta}}\right)\right) \\
&=\sum_{i=1}^n \log\Big(1+e^{-(2y_i-1)\mathbf{x}_i^T\beta}\Big) \\
&=\sum_{i=1}^n \log\Big(1+e^{-\tilde{y}_i\mathbf{x}_i^T\beta}\Big)
\end{aligned}$$

3. Let $j\in\{1,\dots,d\}$, then

$$\frac{\partial f}{\partial\beta_j}(\beta)=\sum_{i=1}^n -\tilde{y}_i(\mathbf{x}_i)_j\frac{e^{-\tilde{y}_i\mathbf{x}_i^T\beta}}{1+e^{-\tilde{y}_i\mathbf{x}_i^T\beta}}.$$

Thus,

$$\nabla f(\beta) = -\sum_{i=1}^{n} \tilde{y}_i \frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}} \mathbf{x}_i.$$

4. Let $j,k \in \{1,\ldots,d\}$, then

$$\begin{aligned}
\frac{\partial^2 f}{\partial \beta_k \beta_j}(\beta) &= -\sum_{i=1}^{n} \tilde{y}_i (\mathbf{x}_i)_j \frac{\partial}{\partial \beta_k} \left(\frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}} \right) \\
&= \sum_{i=1}^{n} \tilde{y}_i^2 (\mathbf{x}_i)_j (\mathbf{x}_i)_k \frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{\left(1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}\right)^2} \\
&= \sum_{i=1}^{n} \frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{\left(1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}\right)^2} (\mathbf{x}_i \mathbf{x}_i^T)_{jk}.
\end{aligned}$$

Consequently, the Hessian matrix takes the form

$$H = \sum_{i=1}^{n} \frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{\left(1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}\right)^2} \mathbf{x}_i \mathbf{x}_i^T.$$

Note that one can also write

$$\begin{aligned}
H &= \sum_{i=1}^{n} \frac{e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}} \frac{1}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}} \mathbf{x}_i \mathbf{x}_i^T \\
&= \sum_{i=1}^{n} \left(1 - \frac{1}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}\right) \left(\frac{1}{1+e^{-\tilde{y}_i \mathbf{x}_i^T \beta}}\right) \mathbf{x}_i \mathbf{x}_i^T \\
&\leq \frac{1}{4} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T
\end{aligned}$$

5. Let $\mathbf{z} \in \mathbb{R}^d$, then

$$\mathbf{z}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{z} = (\mathbf{x}_i^T \mathbf{z})^T (\mathbf{x}_i^T \mathbf{z}) \geq 0.$$

Hence

$$\mathbf{z}^T H \mathbf{z} \geq 0,$$

and the function f is convex.

Exercise 2 (Ridge and Lasso estimates)

We assume to be in a regression setting where we are given a dataset $\mathscr{D}_n = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \dots, n\}$. Recall that the proximal function writes, for any $g : \mathbb{R}^d \to \mathbb{R}$ convex, and any $\beta \in \mathbb{R}^d$,

$$\text{prox}_g(\beta) = \underset{\beta' \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} \|\beta' - \beta\|_2^2 + g(\beta') \right\}$$

Ridge penalization

1. Let $\lambda \geq 0$ and set $g : \beta \to \lambda \|\beta\|_2^2$, and $\lambda \geq 0$. Set $\beta \in \mathbb{R}$ and solve the following optimization problem:

$$\hat{\beta} = \underset{\beta' \in \mathbb{R}}{\text{argmin}} \left\{ \frac{1}{2} (\beta' - \beta)^2 + \lambda \beta'^2 \right\}$$

2. Using the previous question, find the solution of

$$\underset{\beta' \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} \|\beta' - \beta\|_2^2 + \lambda \|\beta'\|_2^2 \right\}$$

   Note that the previous quantity is nothing but $\text{prox}_g(\beta)$.

Lasso penalization

3. Prove that the solution of the following optimization problem

$$\hat{\beta} \in \underset{\beta' \in \mathbb{R}}{\text{argmin}} \left\{ \frac{1}{2} (\beta' - \beta)^2 + \lambda |\beta'| \right\};$$

   can be written as

$$\hat{\beta} = \text{sign}(\beta) \left(|\beta| - \lambda \right)_+ ,$$

   where $x_+ = \max(x, 0)$.

4. Using the previous question, solve

$$\underset{\beta' \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} \|\beta' - \beta\|_2^2 + \lambda \|\beta'\|_1^2 \right\}$$

5. Plot $\hat{\beta}_j$ as a function β. What is the influence of λ on the estimation?

Solution of exercise 2:

1. Set
$$f : \beta' \mapsto \frac{1}{2}(\beta' - \beta)^2 + \lambda(\beta')^2.$$
One has
$$\begin{aligned} & f'(\hat{\beta}) = 0 \\ \Leftrightarrow\ & (\hat{\beta} - \beta) + 2\lambda\hat{\beta} = 0 \\ \Leftrightarrow\ & \hat{\beta} = \frac{\beta}{1+2\lambda}. \end{aligned}$$

2. According to the previous question, we have,
$$\underset{\beta' \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\beta' - \beta\|_2^2 + \lambda \|\beta'\|_2^2 \right\}$$
which is the same as solving, for all $1 \le j \le d$,
$$\underset{\beta'_j \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2}(\beta'_j - \beta)^2 + \lambda(\beta'_j)^2 \right\}.$$
Thus,
$$\operatorname{prox}_g(\beta) = \frac{1}{1+2\lambda}\beta.$$

3. Set
$$f : \beta' \mapsto \frac{1}{2}(\beta' - \beta)^2 + \lambda|\beta'|.$$
Note that , for all $\beta' \neq 0$,
$$f'(\beta') = \begin{cases} \beta' - \beta - \lambda & \text{if } \beta' < 0 \\ \beta' - \beta + \lambda & \text{if } \beta' > 0. \end{cases}$$
The function f is strictly convex and tends to infinity when $\beta' \to \infty$. Thus, it admits a unique minimum. If this minimum $\hat{\beta} \neq 0$ it must satisfy
$$f'(\hat{\beta}) = 0,$$
that is
$$\hat{\beta} = \begin{cases} \beta + \lambda & \text{if } \hat{\beta} < 0 \\ \beta - \lambda & \text{if } \hat{\beta} > 0, \end{cases}$$
which is possible if
$$\beta + \lambda < 0 \quad \text{or} \quad \beta - \lambda > 0,$$

that is $|\beta| > \lambda$. If $|\beta| \leq \lambda$, the minimum does not belong to $(-\infty, 0) \cup (0, \infty)$ and thus, since it must exist, it is equal to 0. Gathering the previous facts, we get

$$\hat{\beta} = \begin{cases} \beta + \lambda & \text{if } \beta < -\lambda \\ \beta - \lambda & \text{if } \beta > \lambda \\ 0 & \text{if } |\beta| \leq \lambda. \end{cases}$$

This can also be written as

$$\hat{\beta} = \operatorname{sign}(\beta)(|\beta| - \lambda)_+.$$

4. According to the previous question and similarly to question 2, we have

$$\operatorname{prox}_g(\beta) = \operatorname{sign}(\beta) \odot (|\beta| - \lambda)_+.$$

5. According to the previous result, we see that $\hat{\beta}_j = 0$ if $|\beta_j| \leq \lambda$. The parameter λ is thus the threshold of the procedure.

Exercise 3 (Descent lemma)

Prove the descent lemma: if f is L-smooth, that is, for all $\beta, \beta' \in \mathbb{R}^d$,

$$\|\nabla f(\beta) - \nabla f(\beta')\| \leq L\|\beta - \beta'\|,$$

then, for any $\beta, \beta' \in \mathbb{R}^d$,

$$f(\beta') \leq f(\beta) + \langle \nabla f(\beta), \beta' - \beta \rangle + \frac{L}{2}\|\beta - \beta'\|_2^2.$$

Solution of exercise 3:

Use the fact that

$$\begin{aligned} f(\beta') &= f(\beta) + \int_0^1 \langle \nabla f(\beta + t(\beta' - \beta)), \beta' - \beta \rangle dt \\ &= f(\beta) + \langle \nabla f(\beta), \beta' - \beta \rangle \\ &\quad + \int_0^1 \langle \nabla f(\beta + t(\beta' - \beta)) - \nabla f(\beta), \beta' - \beta \rangle dt, \end{aligned}$$

so that

$$\begin{aligned} &|f(\beta') - f(\beta) - \langle \nabla f(\beta), \beta' - \beta \rangle| \\ &\qquad \leq \int_0^1 |\langle \nabla f(\beta + t(\beta' - \beta)) - \nabla f(\beta), \beta' - \beta \rangle dt| \\ &\qquad \leq \int_0^1 \|\nabla f(\beta + t(\beta' - \beta)) - \nabla f(\beta)\| \|\beta' - \beta\| dt \\ &\qquad \leq \int_0^1 Lt\|\beta' - \beta\|^2 dt = \frac{L}{2}\|\beta' - \beta\|^2 \end{aligned}$$

which proves the descent lemma. □

Exercise 4 (Convergence of gradient descent)

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex, differentiable and L-smooth. Then the gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(\beta^{(k)}) - f(\beta^\star) \leq \frac{\|\beta^{(0)} - \beta^\star\|_2^2}{2tk},$$

where

$$\beta^\star \in \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} f(\beta).$$

Solution of exercise 4: The descent lemma gives us

$$f(\beta') \leq f(\beta) + \langle \nabla f(\beta), \beta' - \beta \rangle + \frac{L}{2}\|\beta - \beta'\|_2^2.$$

Let $\beta' = \beta^+ = \beta - t\nabla f(\beta)$, we have

$$\begin{aligned} f(\beta^+) &\leq f(\beta) + \langle \nabla f(\beta), \beta - t\nabla f(\beta) - \beta \rangle + \frac{L}{2}\|\beta - (\beta - t\nabla f(\beta))\|_2^2 \\ &= f(\beta) - \Big(1 - \frac{Lt}{2}\Big) t \|\nabla f(\beta)\|_2^2 \end{aligned}$$

Taking $0 \leq t \leq 1/L$, we have

$$f(\beta^+) \leq f(\beta) - \frac{t}{2}\|\nabla f(\beta)\|_2^2.$$

Since f is convex,

$$f(\beta) \leq f(\beta^\star) + \langle \nabla f(\beta), \beta - \beta^\star \rangle,$$

and

$$\begin{aligned} f(\beta^+) &\leq f(\beta) - \frac{t}{2}\|\nabla f(\beta)\|_2^2 \qquad (1) \\ &\leq f(\beta^\star) + \langle \nabla f(\beta), \beta - \beta^\star \rangle - \frac{t}{2}\|\nabla f(\beta)\|_2^2 \\ &= f(\beta^\star) + \frac{1}{2t}\Big(\|\beta - \beta^\star\|^2 - \|\beta - \beta^\star - t\nabla f(\beta)\|^2\Big) \\ &= f(\beta^\star) + \frac{1}{2t}\Big(\|\beta - \beta^\star\|^2 - \|\beta^+ - \beta^\star\|^2\Big). \end{aligned}$$

Summing over iterations, we have

$$\begin{aligned} \sum_{i=1}^{k} (f(\beta^{(i)}) - f(\beta^\star)) &\leq \frac{1}{2t}\sum_{i=1}^{k}\Big(\|\beta^{(i-1)} - \beta^\star\|^2 - \|\beta^{(i)} - \beta^\star\|^2\Big) \\ &\leq \frac{1}{2t}\Big(\|\beta^{(0)} - \beta^\star\|^2 - \|\beta^{(k)} - \beta^\star\|^2\Big) \\ &\leq \frac{1}{2t}\|\beta^{(0)} - \beta^\star\|^2. \end{aligned}$$

According to (1), $f(\beta^{(k)})$ is nonincreasing. Thus,

$$f(\beta^{(k)}) - f(\beta^\star) \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(\beta^{(k)}) - f(\beta^\star)\right) \leq \frac{\|\beta^{(0)} - \beta^\star\|^2}{2tk},$$

and, by choosing $t = 1/L$, we have

$$f(\beta^{(k)}) - f(\beta^\star) \leq \frac{L\|\beta^{(0)} - \beta^\star\|^2}{2k},$$

Exercise 5 (Optimality of $\eta = 1/L$)

Explain why do we set the step size $\eta = 1/L$ in the gradient descent algorithm.
Solution of exercise 5: According to the descent lemma, for any $\beta, \beta' \in \mathbb{R}^d$,

$$f(\beta') \leq f(\beta) + \langle \nabla f(\beta), \beta' - \beta \rangle + \frac{L}{2}\|\beta - \beta'\|_2^2.$$

It is then natural to minimize the right-hand side. At the kth iteration,

$$\begin{aligned}
&\underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}}\left\{f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)}\rangle + \frac{L}{2}\|\beta^{(k)} - \beta\|_2^2\right\}\\
\Leftrightarrow\ &\underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}}\left\{f(\beta^{(k)}) + \frac{L}{2}\|\beta - \beta^{(k)} + \frac{1}{L}\nabla f(\beta^{(k)})\|_2^2\right\}\\
\Leftrightarrow\ &\underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\|\beta - \left(\beta^{(k)} + \frac{1}{L}\nabla f(\beta^{(k)})\right)\|_2^2\right\}.
\end{aligned}$$

Thus, it is natural to choose

$$\beta^{(k+1)} = \beta^{(k)} + \frac{1}{L}\nabla f(\beta^{(k)}).$$