Machine Learning - 2020-2021
Universitat Pompeu Fabra

Problems 8: Logistic Regression and SoftMax

Goal

The goal of this practice is to understand two different classifiers, one for binary classification and the other for multiclass. This two classifiers, that does not allow closed form solutions take they main role in Neural Network. We begin with a general linear classifier. We assume $\mathbf{w} = (w_0, w_1 \ldots, w_D)^T$ and $\mathbf{x}^{(n)} = (1, x_1^{(n)}, \ldots, x_D^{(n)})^T$ when necessary.

- *LS Linear Classifier*:

Model	$h_{\mathbf{w}}(\mathbf{x})$	$= \text{sign}(\mathbf{w}^T\mathbf{x})$
Least Square Error	$\mathbb{E}(\mathbf{w})$	$= \frac{1}{2}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$
Gradient	$\frac{\partial \mathbb{E}(\mathbf{w})}{\partial w_j}$	$= \sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)x_j^{(n)}$

- *Logistic Regression*:

Model	$h_{\mathbf{w}}(\mathbf{x})$	$=P(\mathcal{C}_1\|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$ where $\sigma(a) = \frac{1}{1+e^{-a}}$ is the *Sigmoid* or *Logistic* function
Cross-entropy error	$\mathbb{E}(\mathbf{w})$	$= -\sum_{n=1}^{N}\left(y^{(n)}\ln h_{\mathbf{w}}(\mathbf{x}^{(n)}) + (1-y^{(n)})\ln(1-h_{\mathbf{w}}(\mathbf{x}^{(n)}))\right)$
Gradient	$\frac{\partial \mathbb{E}(\mathbf{w})}{\partial w_j}$	$=\sum_{1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)x_j^{(n)}$

- *Softmax function*:

Model	$h_k(\mathbf{x})$	$=P(\mathcal{C}_k\|\mathbf{x}) = \frac{e^{\mathbf{w}_k^T\mathbf{x}}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T\mathbf{x}}}$
Cross-entropy error	$\mathbb{E}(\mathbf{w}_1, \ldots, \mathbf{w}_K)$	$= -\sum_{n=1}^{N}\left(y_1^{(n)}\ln h_1^{(n)} + \ldots, y_K^{(n)}\ln h_K^{(n)}\right) = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)}\ln h_k^{(n)}$ where $h_k^{(n)} = h_k(\mathbf{x}^{(n)})$
Gradient	$\frac{\partial \mathbb{E}(\mathbf{w}_1,\ldots,\mathbf{w}_K)}{\partial \mathbf{w}_j}$	$=\sum_{n=1}^{N}\left(h_j^{(n)} - y_j^{(n)}\right)x_j^{(n)}$

Exercises

Linear and Logistic classification

1. Consider two classes: the class $C_1 = \{\mathbf{x}^{(1)} = \binom{1}{2}\}$ with a label 1 and the class $C_2 = \{\mathbf{x}^{(2)} = \binom{-1}{-2}, \mathbf{x}^{(3)} = \binom{-2}{-1}\}$ with label -1.

   Consider a **linear classifier** $h_{\mathbf{w}}(\mathbf{x})$.

   (a) Plot the points, write the parametric form of $h_{\mathbf{w}}(\mathbf{x})$ and guess -and draw- what could be a good solution.

   Solution: The classifier will be of the form: $h(\binom{x_1}{x_2}) = \text{sign}(w_0 + w_1x_1 + w_2x_2) = \text{sign}(\mathbf{w}^T\begin{pmatrix}x_1\\x_2\\1\end{pmatrix}) = \text{sign}(\mathbf{w}^T\mathbf{x})$

   (b) Write the error function (least square cost) that is minimised when estimating the linear classifier that separates the given data.

   Solution:

$$\mathbb{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=n}^{N}(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)})^2 = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(n)})^2$$
$$= \frac{1}{2}\Big((\mathbf{w}^T\mathbf{x}^{(1)} - 1)^2 + (\mathbf{w}^T\mathbf{x}^{(2)} + 1)^2 + (\mathbf{w}^T\mathbf{x}^{(3)} + 1)^2\Big)$$

(c) Find the closed-form solution for the linear classifier that separates both classes. To do so, derive the expression of the error from the previous exercise with respect to the weight vector and set it to 0.

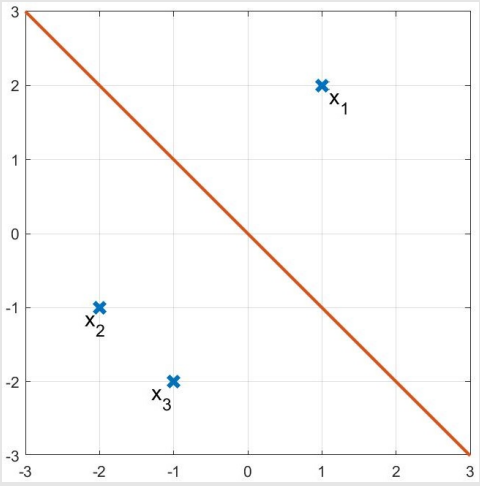
Solution:

$$\mathbb{E}(\mathbf{w}) = \frac{1}{2}\Big((\mathbf{w}^\top \mathbf{x}^{(1)} - 1)^2 + (\mathbf{w}^\top \mathbf{x}^{(2)} + 1)^2 + (\mathbf{w}^\top \mathbf{x}^{(3)} + 1)^2\Big) \implies$$
$$\implies \mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$$

(d) Estimate the linear classifier using the closed-form solution found in the previous exercise. Draw the data points and the resulting linear classifier in the same figure.

Solution:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} = \left[\begin{pmatrix} 1 & -1 & -2 \\ 2 & -2 & -1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 1 \\ -1 & -2 & 1 \\ -2 & -1 & 1 \end{pmatrix}\right]^{-1}\begin{pmatrix} 1 & -1 & -2 \\ 2 & -2 & -1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$$
$$= \begin{pmatrix} 6 & 6 & -2 \\ 6 & 9 & -1 \\ -2 & -1 & 3 \end{pmatrix}^{-1}\begin{pmatrix} 4 \\ 5 \\ -1 \end{pmatrix} = \frac{1}{36}\begin{pmatrix} 26 & -16 & 12 \\ -16 & 14 & -6 \\ 12 & -6 & 18 \end{pmatrix}\begin{pmatrix} 4 \\ 5 \\ -1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$



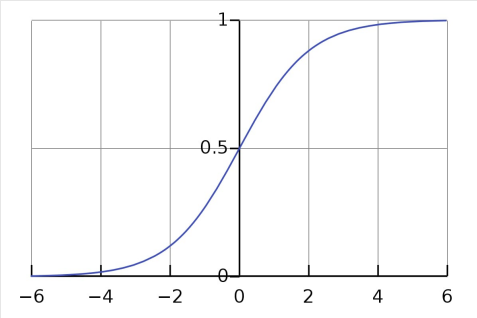
2. Consider two classes: the class $C_1 = \{\mathbf{x}^{(1)} = \binom{1}{2}\}$ with a label 1 and the class $C_2 = \{\mathbf{x}^{(2)} = \binom{-1}{-2}, \mathbf{x}^{(3)} = \binom{-2}{-1}\}$ with label -1.

(a) Plot the points and write the parametric form of a **logistic classifier** $h_{\mathbf{w}}(\mathbf{x})$. What is the difference with a linear classifier?

Solution: The classifier will be of the form: $h_{\mathbf{w}}(\binom{x_1}{x_2}) = \sigma(\mathbf{w}^T\mathbf{x}))$ where $\mathbf{w}^T\mathbf{x} = w_0 + w_1x_1 + w_2x_2$ and $\sigma(a) = \frac{1}{1+\exp(-a)}$.

(b) Labels 1 and -1 are not adequate when solving a two-class problem with logistic regression. Why? Which labels should we choose?

Solution: The sigmoid function never returns negative values, i.e, $\sigma(z) > 0 \ \ \forall z \in \mathbb{R}$. Therefore, label -1 can never be predicted by the classifier $h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$ for any data point $\mathbf{x}$. To solve this issue, we take label 0 for class C_2.



(c) Write the error function that is minimised when estimating the logistic classifier that separates the given data and develop it.

Solution:
$\mathbb{E}(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{n=1}^{N} y^{(n)} \ln h^{(n)} + (1-y^{(n)}) \ln (1-h^{(n)})$ where $h^{(n)} = h_{\mathbf{w}}(\mathbf{x}^{(n)}) = \sigma(\mathbf{w}^T \mathbf{x}^{(n)})$
In this special case, $y^{(1)} = 1, 1-y^{(2)} = 1; 1-y^{(3)} = 1$ the rest of terms are zero:

$$\mathbb{E}(\mathbf{w}) = -\ln(\sigma(\mathbf{w}^T \mathbf{x}^{(1)})) - \ln(1-\sigma(\mathbf{w}^T \mathbf{x}^{(2)})) - \ln(1-\sigma(\mathbf{w}^T \mathbf{x}^{(3)}))$$

(d) (Optional) Derive the expression of the error from the previous exercise with respect to the weight vector.

Solution:

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{\partial \ln(\sigma(\mathbf{w}^\top \mathbf{x}^{(1)}))}{\partial \mathbf{w}} - \frac{\partial \ln(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(2)}))}{\partial \mathbf{w}} - \frac{\partial \ln(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(3)}))}{\partial \mathbf{w}} \\ &= -\frac{1}{\sigma(\mathbf{w}^\top \mathbf{x}^{(1)})}\frac{\partial \sigma(\mathbf{w}^\top \mathbf{x}^{(1)})}{\partial \mathbf{w}} - \frac{1}{1-\sigma(\mathbf{w}^\top \mathbf{x}^{(2)})}\frac{\partial (1-\sigma(\mathbf{w}^\top \mathbf{x}^{(2)}))}{\partial \mathbf{w}} - \frac{1}{1-\sigma(\mathbf{w}^\top \mathbf{x}^{(3)})}\frac{\partial (1-\sigma(\mathbf{w}^\top \mathbf{x}^{(3)}))}{\partial \mathbf{w}}\end{aligned}$$

Since

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}}\sigma(\mathbf{w}^\top \mathbf{x}) &= \frac{\partial}{\partial \mathbf{w}}\frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x})} = -(1+\exp(-\mathbf{w}^\top \mathbf{x}))^{-2}\exp(-\mathbf{w}^\top \mathbf{x})(-\mathbf{x}) = \frac{\mathbf{x}\exp(-\mathbf{w}^\top \mathbf{x})}{(1+\exp(-\mathbf{w}^\top \mathbf{x}))^2} \\ &= \mathbf{x}\sigma(\mathbf{w}^\top \mathbf{x})(1-\sigma(\mathbf{w}^\top \mathbf{x}))\end{aligned}$$

then

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{\mathbf{x}^{(1)}\sigma(\mathbf{w}^\top \mathbf{x}^{(1)})(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(1)}))}{\sigma(\mathbf{w}^\top \mathbf{x}^{(1)})} - \frac{-\mathbf{x}^{(2)}\sigma(\mathbf{w}^\top \mathbf{x}^{(2)})(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(2)}))}{1-\sigma(\mathbf{w}^\top \mathbf{x}^{(2)})} - \frac{-\mathbf{x}^{(3)}\sigma(\mathbf{w}^\top \mathbf{x}^{(3)})(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(3)}))}{1-\sigma(\mathbf{w}^\top \mathbf{x}^{(3)})} \\ &= -\mathbf{x}^{(1)}(1-\sigma(\mathbf{w}^\top \mathbf{x}^{(1)})) + \mathbf{x}^{(2)}\sigma(\mathbf{w}^\top \mathbf{x}^{(2)}) + \mathbf{x}^{(3)}\sigma(\mathbf{w}^\top \mathbf{x}^{(3)}) \\ &= \sum_{n=1}^{3} \mathbf{x}^{(n)}(\sigma(\mathbf{w}^\top \mathbf{x}^{(n)}) - y^{(n)})\end{aligned}$$

(e) (Jupyter) Generate a gradient descend algorithm with ridge regularisation. To do so, you only have to add to the gradient an extra term λw, where λ is the regularisation parameter. Plot the data points and the resulting linear classifier in the same figure.

(f) (Jupyter, optional) Estimate the logistic classifier using the function sklearn.linear_model.LogisticRegression. Draw the data points and the resulting linear classifier in the same figure.

3. Consider three classes:
Class 0: $\{\mathbf{x}^{(1)} = \binom{0.5}{0.4}\, \mathbf{x}^{(2)} = \binom{0.8}{0.3}, \mathbf{x}^{(3)} = \binom{0.3}{0.8}\}$
Class 1: $\{\mathbf{x}^{(4)} = \binom{-0.4}{0.3}, \mathbf{x}^{(5)} = \binom{-0.3}{0.7}, \mathbf{x}^{(6)} = \binom{0.7}{0.2}\}$
Class 2: $\{\mathbf{x}^{(7)} = \binom{0.7}{-0.4}, \mathbf{x}^{(8)} = \binom{0.5}{-0.6}, \mathbf{x}^{(9)} = \binom{-0.4}{-0.5}\}$

And assume that the values obtained for the three discriminants are: $\mathbf{w}_0 = \begin{pmatrix} -0.3 \\ 0.87 \\ 1.47 \end{pmatrix}$, $\mathbf{w}_1 = \begin{pmatrix} -0.01 \\ 0.58 \\ 1.02 \end{pmatrix}$ and $\mathbf{w}_3 = \begin{pmatrix} 0.43 \\ -1.90 \\ 0.33 \end{pmatrix}$

(a) Draw the points and codify the class of the vectors according to 1 of K coding.

Solution: $y^{(1)} = y^{(2)} = y^{(3)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$; $y^{(4)} = y^{(5)} = y^{(6)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$; $y^{(7)} = y^{(8)} = y^{(9)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

(b) Determine at which class the calculated softmax will classify the points $x^{(1)}$, $x^{(4)}$ and $x^{(7)}$.

Solution: $W = \begin{pmatrix} w_{10} & w_{11} & w_{12} \\ w_{20} & w_{21} & w_{22} \\ w_{30} & w_{31} & w_{32} \end{pmatrix} = \begin{pmatrix} -0.3 & 0.87 & 1.47 \\ -0.01 & 0.58 & 1.02 \\ 0.43 & -1.90 & 0.33 \end{pmatrix}$
For the point $\mathbf{x}^{(1)}$ we get $\mathbf{a} = W * \begin{pmatrix} 1 \\ 0.5 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.72 \\ 0.69 \\ -0.39 \end{pmatrix}$ and $b = \exp a = \begin{pmatrix} 2.06 \\ 1.99 \\ 0.68 \end{pmatrix}$ $s = \sum(\mathbf{b}) = 4.73$
$\text{sofmax}(\mathbf{x}^{(1)}) = b/s = \begin{pmatrix} 0.44 \\ 0.42 \\ 0.14 \end{pmatrix}$
$\text{softmax}(x^{(4)}) = \begin{pmatrix} 0.15 \\ 0.19 \\ 0.66 \end{pmatrix}$
$\text{softmax}(x^{(7)}) = \begin{pmatrix} 0.36 \\ 0.47 \\ 0.17 \end{pmatrix}$
And this classifier would assign $\mathbf{x}^{(1)}$ to the class 0; $\mathbf{x}^{(4)}$ to the class 2; $\mathbf{x}^{(7)}$ to the class 1.

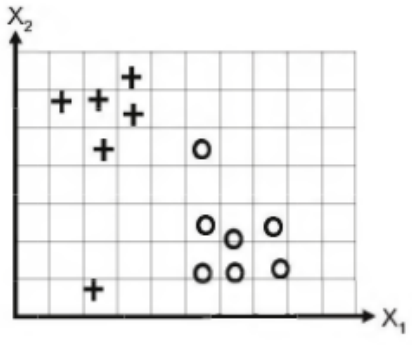
(c) Calculate and plot the discriminating surfaces.

Solution: let $g_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x}$ We will have three DS.
The DS_{12} would be $\{\mathbf{x} | g_1(\mathbf{x}) = g_2(\mathbf{x})\} = \{\mathbf{x} | -0.29 + 0.29x_1 + 0.45x_2 = 0\}$
Analogous for DS_{13} and DS_{23}

(d) Calculate the error for the given solutions.

Solution: We know $\mathbf{W}$ and let $\mathbf{p}^{(n)}$ be: $\mathbf{p}^{(n)} = \begin{pmatrix} h_1^{(n)} \\ h_2^{(n)} \\ h_3^{(n)} \end{pmatrix}$,then
$\mathbf{p}^{(1)} = \begin{pmatrix} 0.44 \\ 0.42 \\ 0.14 \end{pmatrix}$ and we know $y^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
$\mathbb{E}(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3) = -\sum_{n=1}^{9} \left(y^{(n)T} \ln \mathbf{p}^{(n)} \right) = (1, 0, 0) \begin{pmatrix} \ln 0.44 \\ \ln 0.42 \\ \ln 0.14 \end{pmatrix} + \ldots =$

4. We will work with the data of the figure:



(a) (Jupyter) We will fit the model $p(y = 1 | \mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$ by minimizing the Cross-Entropy error $\mathbb{E}(\mathbf{w})$. For the solution you have obtained, How many points are wrong classified?

(b) (Jupyter) Now suppose we regularize only the w_0 parameter, i.e, we minimize

$$\mathbb{E}_T(\mathbf{w}) = \mathbb{E}(\mathbf{w}) + \lambda w_0^2$$

Use the Gradient Descent to estimate the classifier. Comment the results.