Artificial Intelligence I 2022/2023

Week 7 Tutorial and Additional Exercises

Logistic Regression

School of Computer Science

March 24, 2023

In this tutorial. . .

In this tutorial we will be covering

- Univariate and multivariate logistic regression.
- Geometric concepts.
- Optional theoretical exercises.

Univariate logistic regression

Recall the formal statement of *univariate logistic regression*:

- Given a training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$, where $y^{(i)} \in \{0, 1\}$ for all $i = 1, \ldots, n$, train weights w_0, w_1 that minimise a loss function.
- Given this training set, and weights w_0, w_1, the *logistic loss* (or *cross-entropy loss*) function is given as

$$g(w_0, w_1) = -\frac{1}{n} \sum_{i=1}^{n} \Big(y^{(i)} \ln(\sigma(w_0 + w_1 x^{(i)})) \\ + (1 - y^{(i)}) \ln(1 - \sigma(w_0 + w_1 x^{(i)}) \Big)$$

- where $\sigma(x) = \dfrac{1}{1 + e^{-x}}$ is the *sigmoid* function.

Multivariate logistic regression

Recall the formal statement of *multivariate logistic regression*:

- Given a training set $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$, where $y^{(i)} \in \{0, 1\}$ for all $i = 1, \ldots, n$, train a weight vector $\mathbf{w}$ that minimizes a loss function.
- If we have d variables, then for all $i = 1, \ldots, n$, we write

$$\mathbf{x}^{(i)} = (1, x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)}) \text{ and } \mathbf{w} = (w_0, w_1, w_2, \ldots, w_d).$$

- Given this training set and a weight vector $\mathbf{w}$, the *logistic loss* (or cross-entropy loss) function is given as

$$g(\mathbf{w}) = -\frac{1}{n} \sum_{i=1}^{n} \Big(y^{(i)} \ln(\sigma(\mathbf{w}^T \mathbf{x}^{(i)})) + (1 - y^{(i)}) \ln(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})) \Big) .$$

Exercise 1

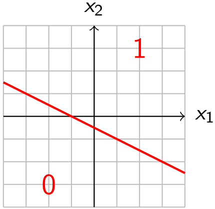
- Consider a logistic regression model with 2 variables that given an instance $\mathbf{x} = (x_1, x_2)$ and weights w_0, w_1, w_2, it predicts the label of $\mathbf{x}$ to be

$$\hat{y} = \begin{cases} 1 \text{ if } w_0 + w_1 x_1 + w_2 x_2 > 0 \\ 0 \text{ if } w_0 + w_1 x_1 + w_2 x_2 < 0 \end{cases}$$

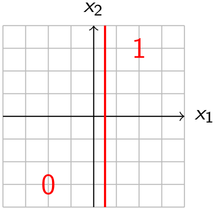
- For each of the following cases, draw the decision boundary in the $x_1 x_2$-plane. This is the line where $w_0 + w_1 x_1 + w_2 x_2 = 0$. Also draw the labels corresponding to the two resulting areas.
  1. $w_0 = 1,\ w_1 = 1,\ w_2 = 2.$
  2. $w_0 = 0,\ w_1 = -3,\ w_2 = 1.$
  3. $w_0 = -2,\ w_1 = 4,\ w_2 = 0.$
  4. $w_0 = -2,\ w_1 = 0,\ w_2 = -1.$

Exercise 1: Solution

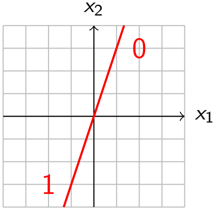
- $w_0 = 1, w_1 = 1, w_2 = 2.$



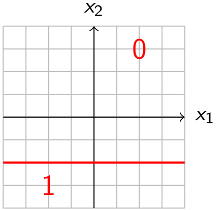
- $w_0 = -2, w_1 = 4, w_2 = 0.$



- $w_0 = 0, w_1 = -3, w_2 = 1.$



- $w_0 = -2, w_1 = 0, w_2 = -1.$

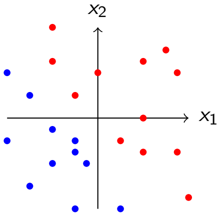


Exercise 2

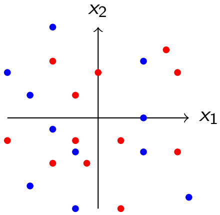
- Logistic regression creates a decision boundary (e.g. a line for two variables) and predicts the label of an instance according to which side of the boundary it falls into.
- Assume each instance has two variables (x_1, x_2), and a label $y \in \{0, 1\}$. Design two training sets that logistic regression can separate with a line, and two training sets that logistic regression cannot separate with a line. For each point, write the values of its two variables and its label.
- Hint: You might want to plot the points of the training set in the x_1x_2-plane to determine whether a line can separate all instances with one label and all instances with the other label.

Exercise 2: Solution

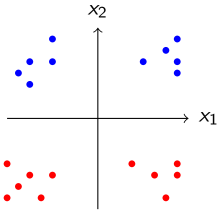
- Linearly separable.



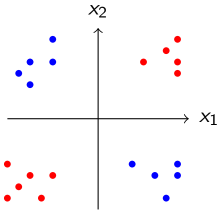
- Non-separable (overlapping).



- Linearly separable.



- Non-linearly separable.

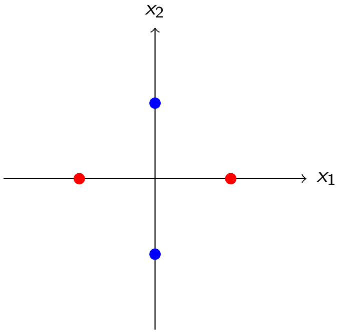


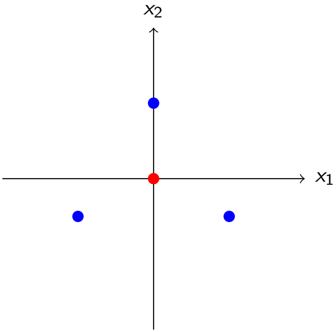
Exercise 3

- This exercise studies the power of a linear decision boundary, as the maximum number of instances it can separate.
- Reconsider the case where each instance has two variables (x_1, x_2) and a label of either 0 or 1.
- Can you plot three instances in the x_1x_2-plane, not all three in the same line, such that no line can separate the two labels? You can freely choose the label of each instance.
- Can you plot four instances in the x_1x_2-plane, no three in the same line, such that no line can separate the two labels? You can freely choose the label of each instance.
- In learning theory, this notion is called the *VC-dimension* (out of the scope of this module).

Exercise 3: Solution

- Three instances can always be separated by a line (proof omitted).
- Four instances cannot always be separated by a line. Two examples are:





Optional Exercise 1

- Consider the *sigmoid* function

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

- Show that σ is increasing and only takes values in $[0, 1]$.
- Can σ take on the values of 0 or 1 for some x?
- Hint: To show that σ is increasing, show that $\sigma'(x) > 0$, for all x. To show that σ takes values in $[0, 1]$, find the limits

$$\lim_{x \to -\infty} \sigma(x) \quad \text{and} \quad \lim_{x \to \infty} \sigma(x).$$

- Hint: To find whether σ takes on the values of 0 or 1, solve

$$\sigma(x) = 0 \quad \text{and} \quad \sigma(x) = 1.$$

Optional Exercise 1: Solution

- First, σ is increasing since, for all x, we have

$$\left(\frac{1}{1+e^{-x}}\right)' = -\frac{1}{(1+e^{-x})^2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2} > 0.$$

- Also, σ only takes values in $[0,1]$ since it is increasing and

$$\lim_{x\to-\infty}\frac{1}{1+e^{-x}} = 0 \quad \text{and} \quad \lim_{x\to\infty}\frac{1}{1+e^{-x}} = 1.$$

- Setting $\sigma(x) = 0$ and $\sigma(x) = 1$ gives respectively

$$1 = 0 \quad \text{and } e^{-x} = 0.$$

- Both are impossible, so σ cannot take on the values of 0 or 1.

Optional Exercise 2

- Let $(\mathbf{x}, y)$ be a data point and $\mathbf{w}$ be the weight vector to be optimised in a multivariate logistic regression model with d variables. Assume that $\mathbf{x}$ and $\mathbf{w}$ are of the form[1]

$$\mathbf{x} = (x_0, x_1, \ldots, x_d) \text{ and } \mathbf{w} = (w_0, w_1, \ldots, w_d).$$

- Let $\sigma(x) = \frac{1}{1+e^{-x}}$ and g be the logistic loss function

$$g(\mathbf{w}) = -\left(y \ln(\sigma(\mathbf{w}^T\mathbf{x})) + (1-y)\ln(1-\sigma(\mathbf{w}^T\mathbf{x}))\right).$$

- Use the derivative rules to prove that

$$\nabla g(\mathbf{w}) = -(y - \sigma(\mathbf{w}^T\mathbf{x}))\mathbf{x}.$$

- Hint: $\frac{\partial \sigma}{\partial w_i}(\mathbf{w}^T\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})(1-\sigma(\mathbf{w}^T\mathbf{x}))x_i,\ i = 0, \ldots, d.$

[1] We usually take $x_0 = 1$, but we leave it as x_0 here.

Optional Exercise 2: Solution

- The partial derivative of g with respect to w_i, $0 \leq i \leq d$, is

$$\begin{aligned}
\frac{\partial g}{\partial w_i}(\mathbf{w}) &= -\left(\frac{y}{\sigma(\mathbf{w}^T\mathbf{x})} - \frac{1-y}{1-\sigma(\mathbf{w}^T\mathbf{x})}\right)\frac{\partial \sigma}{\partial w_i}(\mathbf{w}^T\mathbf{x}) \\
&= -\left(\frac{y-\sigma(\mathbf{w}^T\mathbf{x})}{\sigma(\mathbf{w}^T\mathbf{x})(1-\sigma(\mathbf{w}^T\mathbf{x}))}\right)\sigma(\mathbf{w}^T\mathbf{x})(1-\sigma(\mathbf{w}^T\mathbf{x}))x_i \\
&= -(y-\sigma(\mathbf{w}^T\mathbf{x}))x_i.
\end{aligned}$$

- Therefore, the gradient vector of g is

$$\begin{aligned}
\nabla g(\mathbf{w}) &= (-(y-\sigma(\mathbf{w}^T\mathbf{x}))x_0, \ldots, -(y-\sigma(\mathbf{w}^T\mathbf{x}))x_d) \\
&= -(y-\sigma(\mathbf{w}^T\mathbf{x}))(x_0, \ldots, x_d) \\
&= -(y-\sigma(\mathbf{w}^T\mathbf{x}))\mathbf{x}
\end{aligned}$$

Q&A

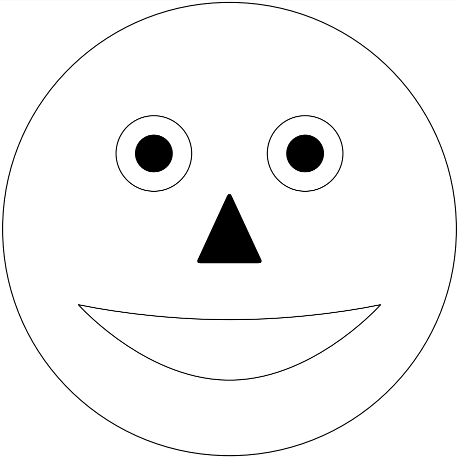
Any questions?

Some closing words. . .

—Regression analysis is the hydrogen bomb of the statistics arsenal.

Charles Wheelan

Until the next time...



Thank you for your attention!