Advanced Institute for Artificial Intelligence

Regressão Logística

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

Introdução

Existem problemas para os quais desejamos estimar a **classe** (rótulo) de uma determinada amostra com base no seu conjunto de dados de entrada, isto é, as **características** do seu problema.

Problemas de classificação são similares aos de regressão pois desejamos aprender uma função que, dada uma entrada, estime um valor de saída. A diferença é que nossa saída será mapeada para uma **probabilidade** (*soft classification*) ou **rótulo** (*hard classification*) da amostra.

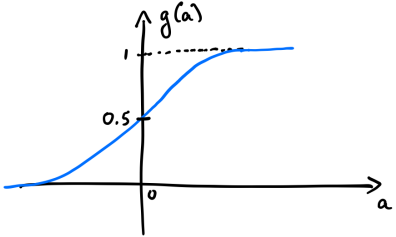
A técnica de Regressão Logística é uma das mais utilizadas na literatura, principalmente por ser simples e dar bons resultados em diversas situações. Ela tem esse nome pelo fato de estimar a probabilidade de uma dada amostra pertencer à uma classe em específico. Portanto, é uma técnica de classificação do tipo *soft*.

Definição do problema: seja um conjunto de dados $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_z, y_z)\}$ tal que $\boldsymbol{x}_i \in \mathbb{R}^{n+1}$ corresponde ao dado de entrada e $y_i \in [0, 1]$ denota o seu respectivo valor de saída. Temos, ainda, que $\mathcal{X}$ pode ser **particionado** da seguinte forma: $\mathcal{X} = \mathcal{X}^1 \cup \mathcal{X}^2$, em que $\mathcal{X}^1$ e $\mathcal{X}^2$ denotam os conjuntos de dados de **treinamento** e **teste**, respectivamente. Nosso objetivo é, dado o conjunto de treinamento, aprender uma função $h : \mathbb{R}^{n+1} \rightarrow [0, 1]$ que consiga estimar a probabilidade de uma amostra pertencer à uma dada classe. Por que usar uma função logística (sigmoide)?

A Função Logística possui algumas propriedades interessantes com a seguinte formulação:

$$g(a) = \frac{1}{1 + e^{-a}}, \tag{1}$$

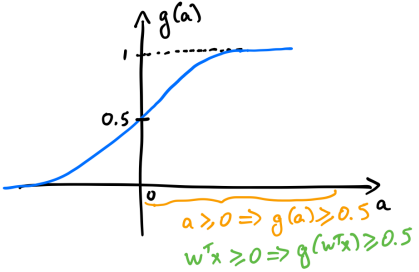
tal que $g(a) \in [0, 1]$.



Como funciona o Regressor Logístico? Dado que queremos obter uma saída $h_{\boldsymbol{w}}(\boldsymbol{x}) \in [0, 1]$, basta modificarmos nossa entrada na Equação 1 como segue:

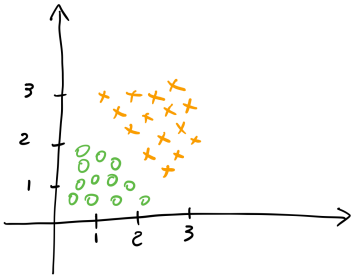
$$\begin{aligned} h_{\boldsymbol{w}}(\boldsymbol{x}) &= g(\boldsymbol{w}^T \boldsymbol{x}) \\ &= \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}}. \end{aligned} \tag{2}$$

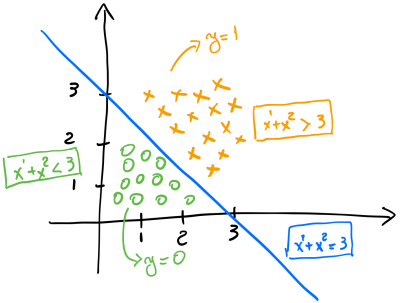
Agora, o termo $\boldsymbol{w}^T \boldsymbol{x}$ é chamado de **função base**.



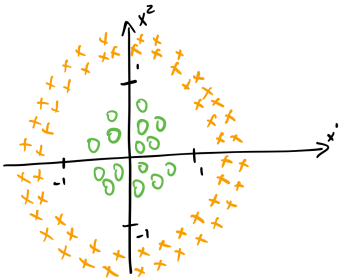
Na prática, queremos que $h_{\boldsymbol{w}}(\boldsymbol{x}) \geq 0.5$ quando $\boldsymbol{w}^T \boldsymbol{x} \geq 0$. De maneira análoga, temos que $h_{\boldsymbol{w}}(\boldsymbol{x}) < 0.5$ quando $\boldsymbol{w}^T \boldsymbol{x} < 0$.

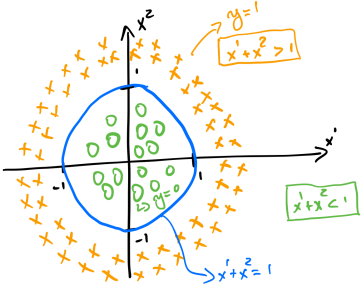
Suponha que tenhamos a seguinte situação, em que $h_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{w}^T\boldsymbol{x})$ e $\boldsymbol{w} = [-3\ 1\ 1]$. Queremos atribuir $y = 1$ se $\boldsymbol{w}^T\boldsymbol{x} > 0$, ou seja, se $-3 + x^1 + x^2 > 0 \Rightarrow x^1 + x^2 > 3$. De maneira análoga, queremos atribuir $y = 0$ se $\boldsymbol{w}^T\boldsymbol{x} < 0$, ou seja, se $-3 + x^1 + x^2 < 0 \Rightarrow x^1 + x^2 < 3$.





Suponha agora outra situação, em que $h_{\boldsymbol{w}}(\boldsymbol{x}) = g(w_0 + w_1 x^1 + w_2 x^2 + w_3 (x^1)^2 + w_4 (x^2)^2)$ e $\boldsymbol{w} = [-1\ 0\ 0\ 1\ 1]$. Queremos atribuir $y = 1$ se $-1 + (x^1)^2 + (x^2)^2 > 0 \Rightarrow (x^1)^2 + (x^2)^2 > 1$. De maneira análoga, queremos atribuir $y = 0$ se $1 + (x^1)^2 + (x^2)^2 < 0 \Rightarrow (x^1)^2 + (x^2)^2 < 1$.





Desta forma, podemos notar que, dependendo da função base escolhida, diferentes superfícies de separação podem ser obtidas. Usualmente, quanto maior o grau do polinômio, mais complexa será a sua superfície.

Assim sendo, temos as seguintes informações até o momento:

- Função hipótese: $h_{\boldsymbol{w}}(\boldsymbol{x}) = \dfrac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}}$.
- Função base: $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$ (padrão).
- Função de custo: ?

Por que não é interessante utilizar MSE como função de custo para o Regressor Logístico?

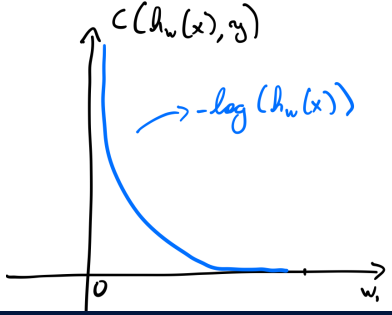
Temos dois problemas principais com MSE quando usamos a técnica de Regressão Logística:

1. Suponha que o rótulo verdadeiro de uma amostra $\boldsymbol{x} \in \mathcal{X}$ qualquer seja $y = 1$, e nosso classificador tenha resultado como saída $h_{\boldsymbol{w}}(\boldsymbol{x}) = 0$. Neste caso (para $m = 1$), temos que $J(\boldsymbol{w}) = \frac{1}{2}(1-0)^2 = 0.5$. Note que essa é uma penalização muito pequena para um **erro** de classificação. Assim sendo, MSE não penaliza muito fortemente erros de classificação e pode levar a um aprendizado insuficiente.
2. A função de custo MSE **não é convexa** para o Regressor Logístico (provado matematicamente).

Vamos, então, definir uma nova função de custo que seja convexa. Nossa função será dividida, inicialmente, em duas partes:

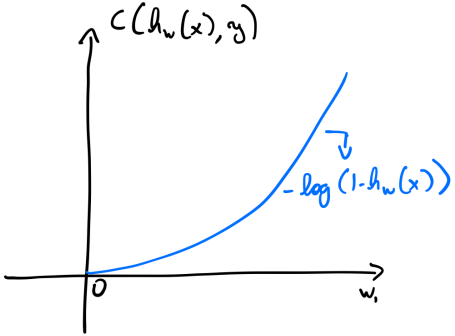
$$C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = \begin{cases} -\log(h_{\boldsymbol{w}}(\boldsymbol{x})) \text{ se } y = 1 \\ -\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x})) \text{ se } y = 0. \end{cases} \tag{3}$$

Vamos, agora, analisar as duas situações separadamente. Como primeiro caso, temos a situação em que $y = 1$.



Neste caso, quando temos um erro, ou seja, $h_{\boldsymbol{w}}(\boldsymbol{x}) = 0$, então $C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = -\log(0) = \infty$. No caso de um acerto, ou seja, $h_{\boldsymbol{w}}(\boldsymbol{x}) = 1$, temos que $C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = -\log(1) = 0$

A próxima situação ocorre quando $y = 0$.



Neste caso, quando temos um erro, ou seja, $h_{\boldsymbol{w}}(\boldsymbol{x}) = 1$, então $C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = -\log(1-1) = -\log(0) = \infty$. No caso de um acerto, ou seja, $h_{\boldsymbol{w}}(\boldsymbol{x}) = 0$, temos que $C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = -\log(1-0) = -\log(1) = 0$.

Podemos escrever a Equação 3 agrupando as duas situações descritas anteriormente:

$$C(h_{\boldsymbol{w}}(\boldsymbol{x}), y) = -y \log(h_{\boldsymbol{w}}(\boldsymbol{x})) - (1 - y) log(1 - h_{\boldsymbol{w}}(\boldsymbol{x})). \quad (4)$$

Quando $y = 1$, termo vermelho é zerado e a equação acima torna-se a Equação 3 para este caso. Por outro lado, quando $y = 0$, termo azul é zerado e a equação acima torna-se a Equação 3 para este caso.

A nossa função de custo final para o Regressor Logístico é dada por:

$$
\begin{aligned}
J(\boldsymbol{w}) &= \frac{1}{m}\sum_{i=1}^{m} C(h_{\boldsymbol{w}}(\boldsymbol{x}_i), y_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} [-y_i \log(h_{\boldsymbol{w}}(\boldsymbol{x_i})) - (1-y_i)\log(1-h_{\boldsymbol{w}}(\boldsymbol{x}_i))]. \qquad (5)
\end{aligned}
$$

A função acima é também conhecida por **entropia cruzada binária**.

Basta, agora, aprendermos o conjunto de parâmetros $\boldsymbol{w}$ de maneira semelhante à Regressão Linear, ou seja, podemos fazer uso da técnica de gradiente descendente. A derivada da função de custo apresentada na Equação 5 é dada como segue:

$$\frac{\partial J(\boldsymbol{w})}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}[(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)x_i^j]. \tag{6}$$

Muito embora a formulação acima seja idêntica à Regressão Linear, devemos lembrar que a função hipótese $(h_{\boldsymbol{w}}(\boldsymbol{x})$ é diferente.

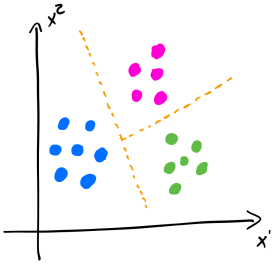
Desta forma, o algoritmo do gradiente descendente pode ser sumarizado da seguinte forma:

1. Atribua valores aleatórios para $\boldsymbol{w}$.
2. Avalie a função de custo $J(\boldsymbol{w})$.
3. Caso o **critério de parada tenha sido atingido**, vá para o passo 6.
4. $w_j^{(t+1)} = w_j^{(t)} - \alpha \frac{1}{m} \sum_{i=1}^{m} [(h_{\boldsymbol{w}}(x_i) - y_i)^2 x_i^j]$.
5. Retorne ao Passo 2.
6. Fim do algoritmo.

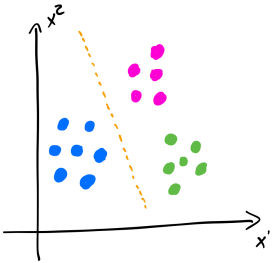
Note que os pesos em $\boldsymbol{w}$ precisam ser atualizados simultaneamente!

No entanto, note que o Regressor Logístico é, naturalmente, um **classificador binário**, isto é, $y_i \in \{0, 1\}$, $\forall i = 1, 2, \ldots, z$. Desta forma, como podemos atuar em problemas que possuem **múltiplas classes**?

Agora, nossos rótulos podem ser representados da seguinte forma: $y_i \in \{0, 1, \ldots, c-1\}$, em que c corresponde ao número de classes. Suponha a seguinte situação em que $c = 3$.

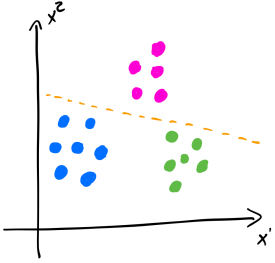


Existem algumas abordagens para tratar esse problema, tais como OVO (*one-versus-one*) e OVA (*one-versus-all*). Vamos considerar a abordagem OVA. Neste caso, para um problema de classificação com c classes, temos que criar c classificadores. Vejamos o exemplo abaixo em que $c = 3$.



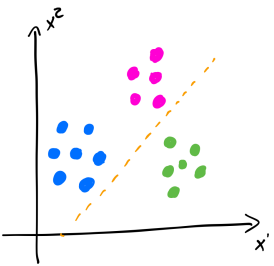
Azul versus todos

$h^1_{\boldsymbol{w}}(\boldsymbol{x})$



Rosa versus todos

$h^2_{\boldsymbol{w}}(\boldsymbol{x})$



Verde versus todos

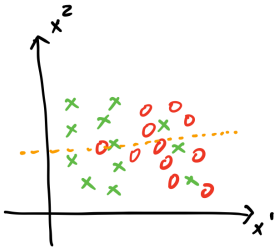
$h^3_{\boldsymbol{w}}(\boldsymbol{x})$

Basta, então, treinarmos cada classificador $h^i_{\boldsymbol{w}}(\boldsymbol{x})$, $\forall i = 0, 1, \ldots, c-1$ em $\mathcal{X}^1$. Dada uma amostra $\boldsymbol{x} \in \mathcal{X}^2$, seu rótulo y será aquele que obedece à seguinte regra de decisão:

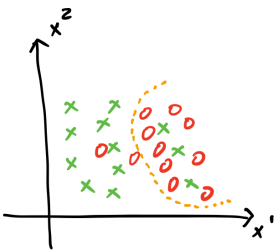
$$y = \arg\max_i \{h^i_{\boldsymbol{w}}(\boldsymbol{x})\}. \tag{7}$$

Regularização

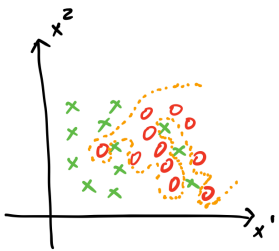
Como vimos anteriormente, a ideia da regularização é evitar com que a técnica fique muito **especializada** (viciada) no conjunto de treinamento e, portanto, não consiga generalizar muito bem no conjunto de teste. Vejamos as três principais situações que podem ocorrer em classificadores de padrões.



Subtreinamento



Bom treinamento



Supertreinamento

Com isso, podemos modificar a Equação 5 da função de custo como segue:

$$J(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}[-y_i \log(h_{\boldsymbol{w}}(\boldsymbol{x_i})) - (1-y_i)\log(1-h_{\boldsymbol{w}}(\boldsymbol{x}_i))] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2, \tag{8}$$

em que λ corresponde à taxa de regularização. Já as derivadas parciais da nova função de custo equivalem à derivadas da MSE no caso da regressão linear, como segue:

$$\frac{\partial J(\boldsymbol{w})}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m}(h_{\boldsymbol{w}}(x_i) - y_i), \tag{9}$$

e

$$\frac{\partial J(\boldsymbol{w})}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}[(h_{\boldsymbol{w}}(x_i) - y_i)x_i^j] + \frac{\lambda}{m}w_j. \tag{10}$$

O algoritmo do gradiente descendente permanece o mesmo.