* In linear regression, some samples are very large  (>>1) or very small (<<0) can have the strong effect to the classification result. So we want to have the output from 0 to 1. We use sigmoid function (1)
* the cost function Logistic regression have many local minimum if we use least square error=> change to another cost function.
* When we take derivative of the course function, it has the same equation like linear regression, but the hypothesis is a sigmoid function.
* When we want to use logistic regression model in multiclass classification, we calculate the probability in each class, and choose the highest value.

1. Có thể hiểu như là classification ko phải là một hàm linear function (là hàm thì input tang thi output cũng tang tuyến tính theo) mà output nên cố định. Chính vì lý do này mà mình chuyển output sử dụng một hàm non-linear như sigmoid function (để input là các samples có đặc biệt như cực lớn thì output cũng phải là 1)