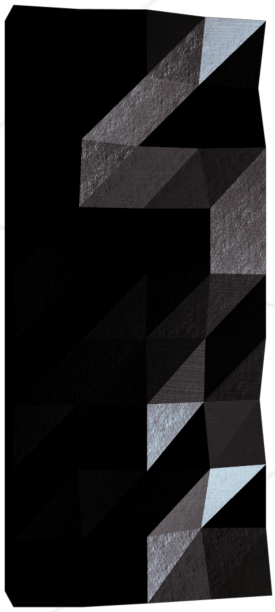


Introduction to Machine Learning

Classification and Logistic Regression

Jeson Lee ljunzhen@42.us.org
Jesonleejunzhen.com

Summary: Learn the most basic type of classification model, logistic regression!

Contents

`Eat, Sleep, Code, Repeat.`

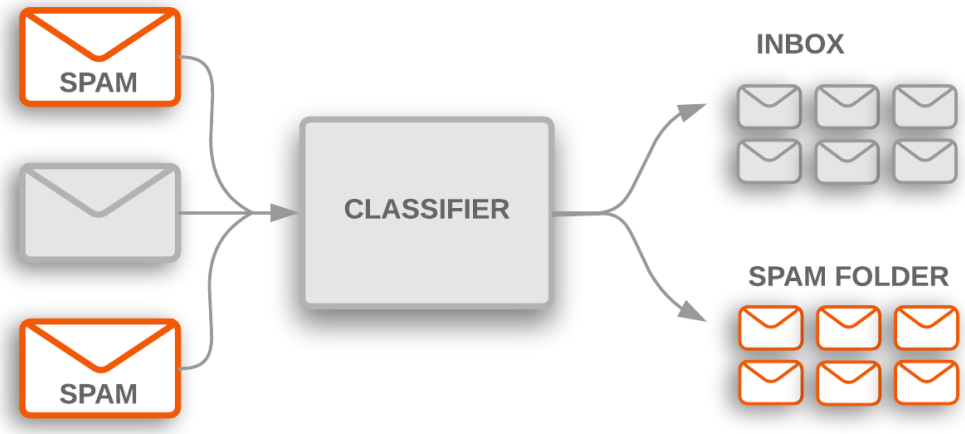
Chapter I

Concepts of Classification

What is classification in machine learning?

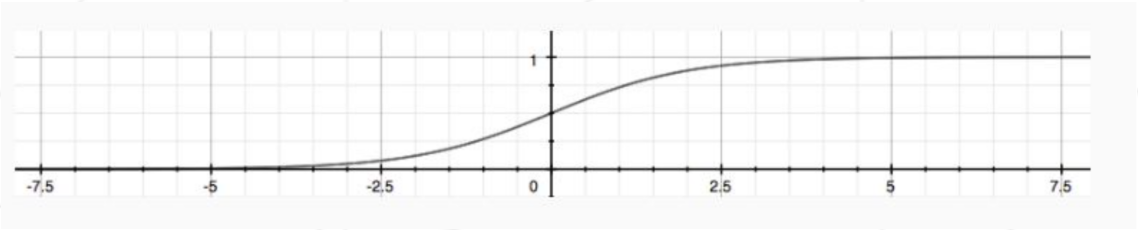
Classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify a new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too.

One of the most common classification models is logistic regression.



Hypothesis Representation

Sigmoid Function or also called Logistic Function is one of the functions that can be used to represent our hypothesis for Classification problems (Binary Classification to be exact)



The values of sigmoid function do not get bigger than 1 nor it gets lesser than 0. This is what makes it good for classification because our classification problem can only have values within the range (0,1).

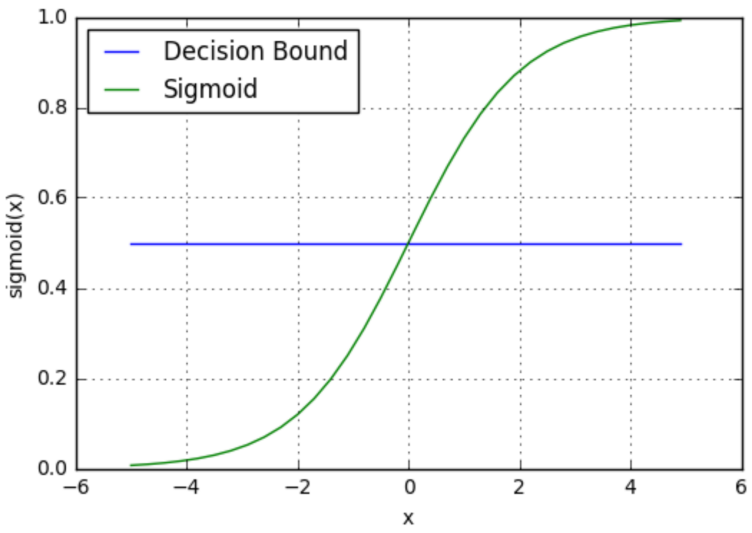
Here's the mathematical formula for it:

$$f(t) = \frac{1}{1+e^{-t}}$$

Decision Boundary

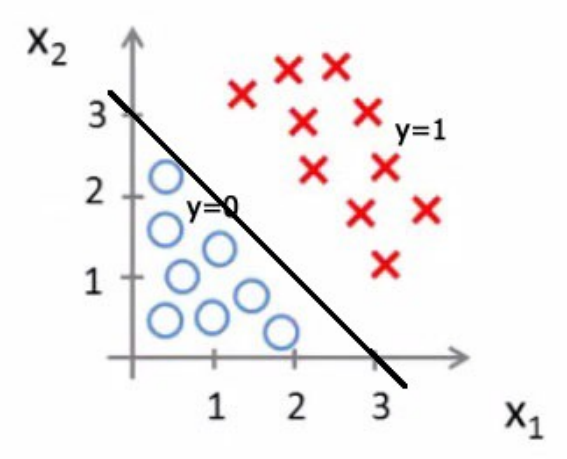
Decision boundary helps to differentiate probabilities into positive class and negative class. threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

p≥0.5,class=1

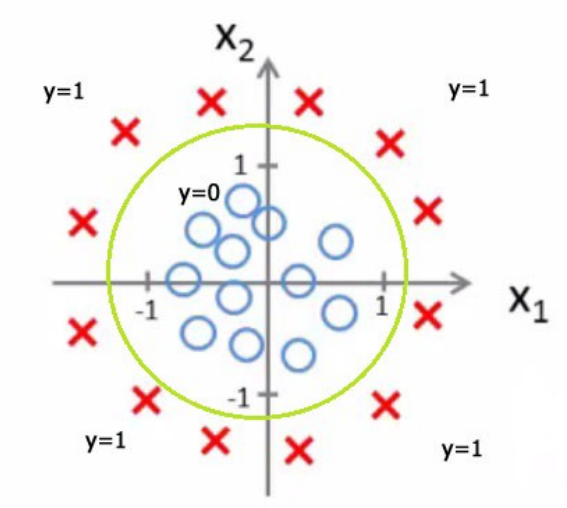


p<0.5,class=0

Linear Decision Boundary:

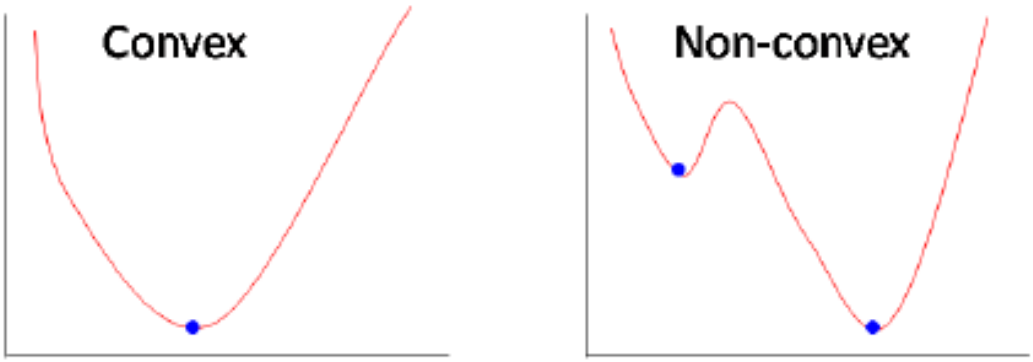


Non-linear Decision Boundary:



Cost Function

The cost function of logistic regression is not the same as the one we used for a linear regression. That's because the logistic function will cause the output to have many local optima. In other words, it will not be a convex function.



What makes a non-convex optimization hard is the presence of saddle points and local minima, where the gradient is (0,...,0) and that have an arbitrarily bad objective value.

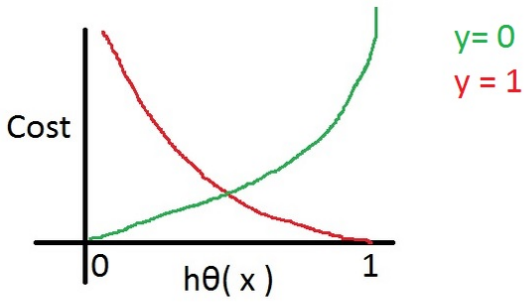
So instead, our cost function for logistic regression looks like:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y} = 1$$

$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y} = 0$$

In order to understand the above cost function in a better way, see the diagram below:



If our correct answer 'y' is 0, then the cost function will be 0 if our hypothesis function also outputs 0. If our hypothesis approaches 1, then the cost function will approach infinity.

If our correct answer 'y' is 1, then the cost function will be 0 if our hypothesis function outputs 1. If our hypothesis approaches 0, then the cost function will approach infinity.

Note that writing the cost function in this way guarantees that $J(\theta)$ is convex for logistic regression.

More information!

- Machine Learning Classifiers
- Logistic Regression - Fun and Easy Machine Learning (video)
- Logistic Regression
- The third week of Andrew Ng's Machine Learning course on Coursera

Chapter II

Ask Your Peers

1. Explain what classification is. What type of machine learning does it fall under?
2. What is logistic regression? What is the function used for it?
3. What is the difference between linear regression and logistic regression?
4. Explain what is the use of the decision boundary.
5. Why is the cost function used for logistic regression different from linear regression?
6. What are the examples you can use a logistic function for?
7. Besides logistic regressions, what are other classification models? List 3 of them.

Chapter III

Exercise 00: Classification Project

	Exercise
Classification Project	
Topics to study : `logistic regression, scikit-learn, pandas, numpy, matplotlib`	
Files to turn in : `suv_launch_prediction.py`	
Forbidden functions : `None`	
Notes : `n/a`	

Scenario:

Imagine that you're a marketing associate for BMW, waiting for the working hour to end. Suddenly, you hear your boss knocking on your office door.

`"Knock Knock"`

You open the door.

`Jeson (CEO): Hey, we are launching a new SUV next week and we need to create an Instagram ad to target people that will most likely buy the new SUV we are making.`

`You: Yes Jeson. I will get it done as soon as possible.`

The good news is you have the dataset of users who bought the previous model of SUV and you happen to know a little bit of machine learning (thanks to Hack High School).

Now, use logistic regression to classify and predict which customers would buy this new BMW SUV. You must be able to visualize the result by using the Matplotlib library. The image below is the end result of the test set that is expected to be displayed.

Using Jupyter notebook to build this project will be highly recommended. Download the dataset to begin. Happy hacking!

Use the scikit-learn library to perform feature scaling and logistic regression.

Chapter IV

Exercise 01: Bonus Project 1

	Exercise
Bonus Project 1	
Topics to study : `linear regression, scikit-learn, pandas, numpy, matplotlib`	
Files to turn in : `suv_launch_prediction_1.py`	
Forbidden functions : `None`	
Notes : `n/a`	

Using the same dataset from the linear regression project, try performing other regression algorithms like support vector machine, decision tree regression or random forest regression and see the difference in them!

Work on at least 1 regression model other than linear regression.

Chapter V

Exercise 02: Bonus Project 2

	Exercise
Bonus Project 2	
Topics to study : `logistic regression, scikit-learn, pandas, numpy, matplotlib`	
Files to turn in : `suv_launch_prediction_2.py`	
Forbidden functions : `None`	
Notes : `n/a`	

Using the same dataset from the logistic regression project, try performing other classification algorithms like K-Nearest-Neighbors, naive bayes, decision tree classification or random forest classification and see the difference in them!

Work on at least 1 classification model other than logistic regression.

Chapter VI

Conclusion

I hope this course gave you a basic understanding and fundamentals of machine learning. You have learned the essential python libraries through exercises to building your own machine learning model. I hope you enjoyed what we have put together and use the skills to further improve yourself.

If you like this course, I will try to work on creating a more advanced course called "intro to deep learning" where you will learn about neural network and eventually creating something really cool that you can show off to your classmates.

Please fill out this feedback form here to let me know what you think about the course and what I can improve.

References and image sources:

- Towards Data Science
- Udemy course (Machine Learning A to Z)
- Hackernoon
- Machine Learning Plus
- Coursera Stanford Machine Learning course
- Kaggle