Would you survive the Titanic?

Machine Learning assignment

Marija Maneva

Train a model

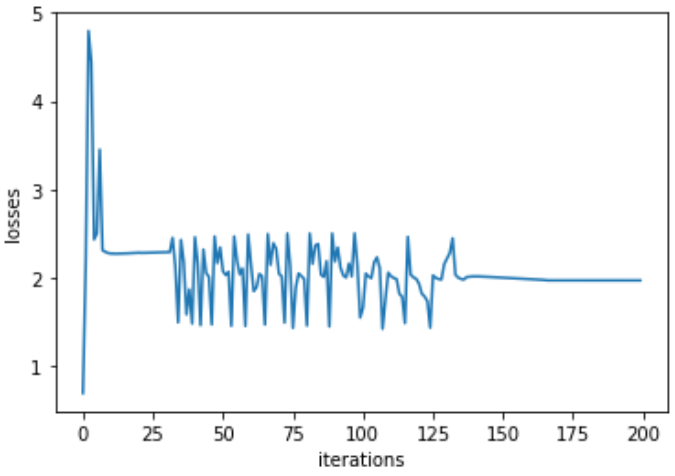
1. Which is a good value for the learning rate?

A good value for the learning rate is 0.005.

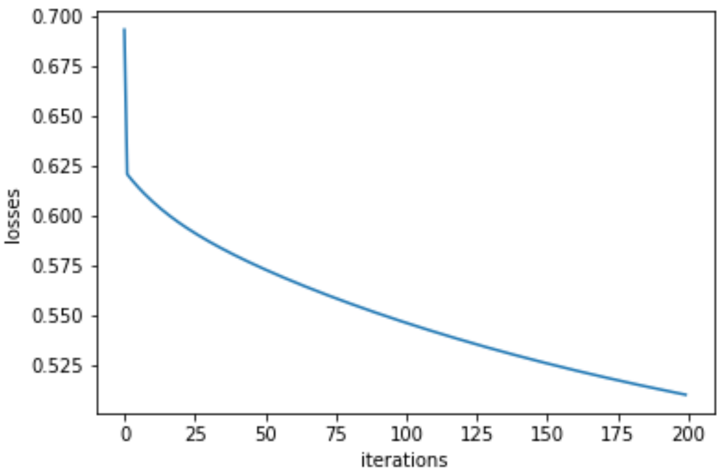
The learning rate gives the rate of speed where the gradient moves during gradient descent. By setting it too high would make the convergence unstable (we would have a “zig zag” plot), otherwise, by setting it too low, it would make the convergence slow.

Here there are some examples on the plots containing iterations vs losses.

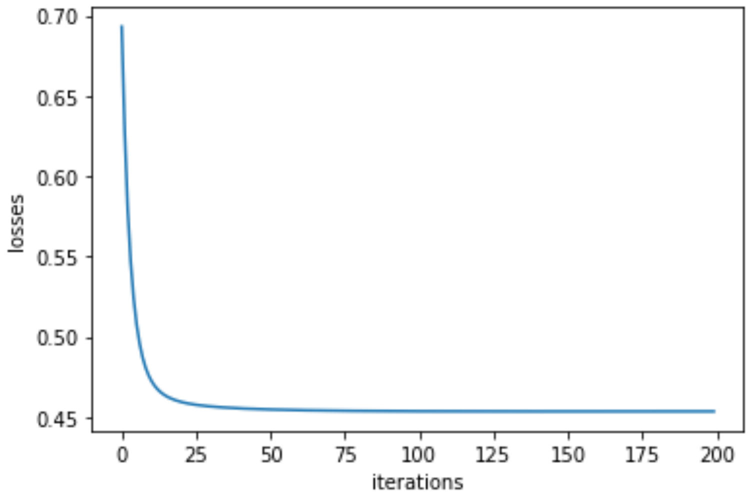
a. learning rate = 0.05 (too high)



b. learning rate = 0.00005 (too low)



c. learning rate = 0.005 (good value)



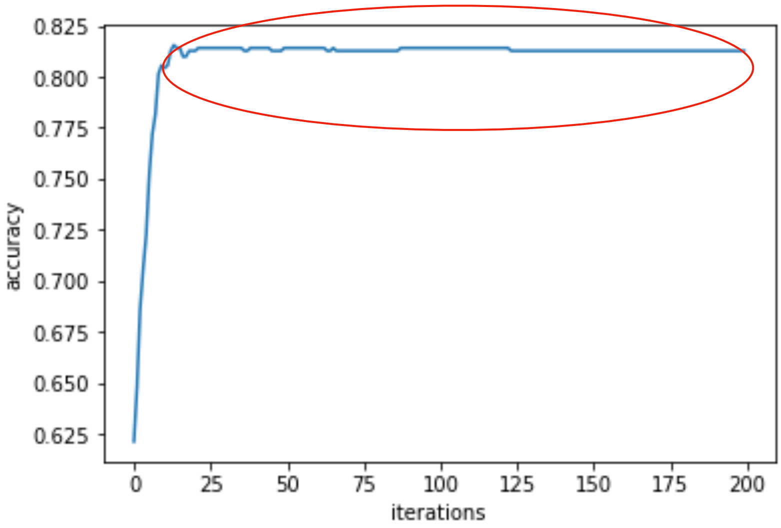
The same trend,by changing the learning rates, can also be seen on the plots of the accuracies (contained in the zip).

2. How many iterations are required to converge?
The number of iterations required to converge is 200 000.
The number of iterations is hard to estimate in advance, the result can better be seen by plotting different numbers until we come to a point where the convergence is reached.

Analyze the model

1. What would be your probability to survive?
My probability to survive would be 71.61% with the following features :
class =2 ; sex =1 ; age = 23; siblings/spouses aboard = 1 ; parents/children aboard = 2; fare = 50 (guessed).

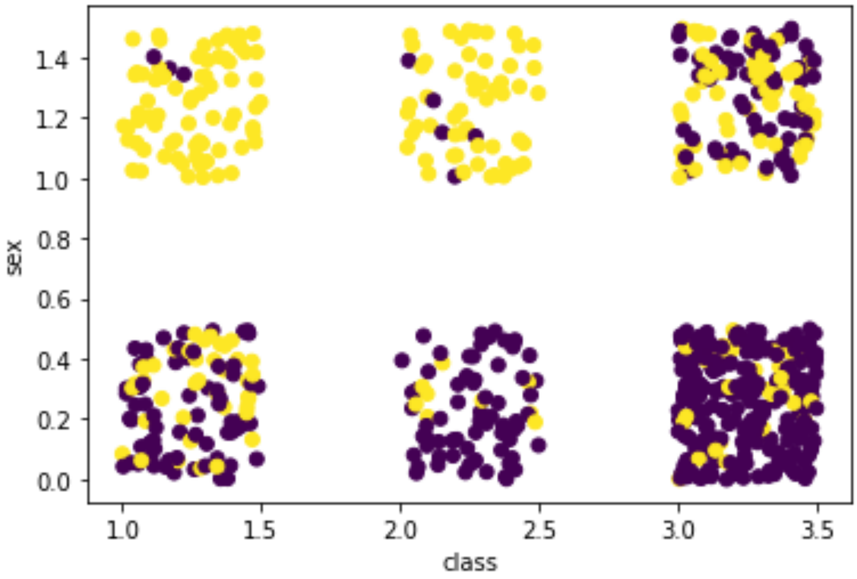
2. What is the training accuracy of the trained model?
The training accuracy of the trained model is 0.81267 (81%).
It can be seen also in the following plot how the convergence reaches the quoted value.



3. Looking at the learned weights, how the individual features influence the probability of surviving?
The weights are values that specify the impact of the features considered.
By looking at the learned weights, it can be seen that the biggest weight is put on the gender. The gender is the feature that influences mostly the results with a weight of 2.99.
The second feature that has a big impact on the probability of surviving is the economic class with a weight of 1.25.
Then there's the feature about having siblings/spouses aboard (w=0.4) and having parents/children aboard (w=0.15).
The features that have a minimal impact are the age (w=0.05) and the the fare (w=0.0044).

4. What kind of passengers was most likely to survive? And what kind to die?
The passengers that were more likely to survive were females, younger passengers from higher class, with lower fare and with less siblings/spouses or parents/children aboard.
The passenger that were more likely to die were males, older passengers from lower classes, with higher fare and with more siblings/spouses or parents/children aboard.

5. Draw a scatter plot showing the distribution of the two classes in the plane defined by the two most influential features. Comment the plot.



The scatter plot is a visual representation of the relationship between two variables or is used to identify different patterns in data. In this case we are examining the features sex and class.

The variable sex contains two values: 1 and 0. This is why on the y-axis the data is distribuееted on these two values.

The variable class contains three values: 1, 2 and 3. As a consequence, the data is divided between these values.

As we can see from the plot, there's not relationship between the two variables. It can be seen how a big part of the passengers are males from the 3rd social class. Women are most likely to be from the first class.

Evaluate the model

1. What is the test accuracy of the model?

The test accuracy of the model is 79.096%.

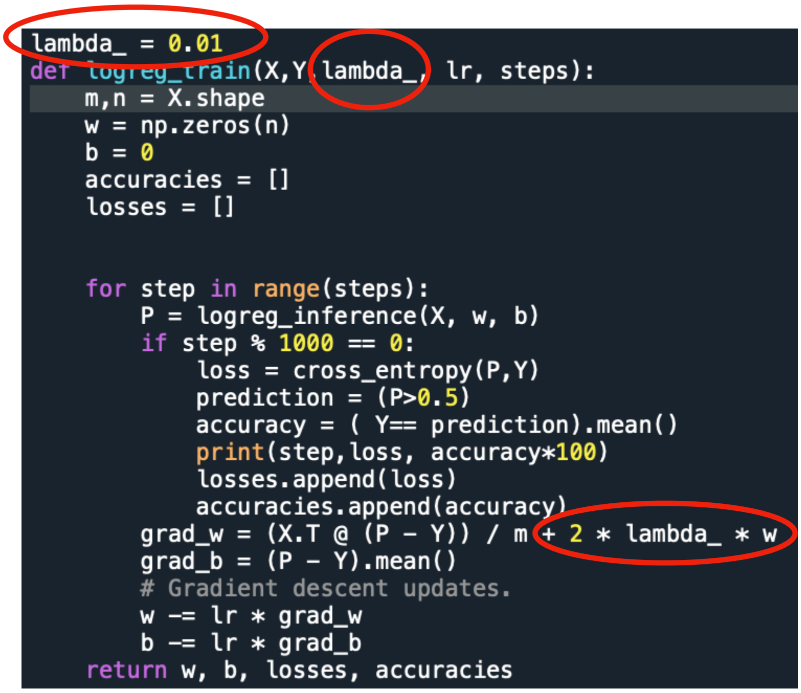
2. Is the model overfitting or underfitting the training set?

A useful tool to detect overfitting and underfitting is the accuracy value on the training and testing set. If a model has high accuracy on the training set, but low accuracy on the test set, this means that you are overfitting your model,; in other words the model fits too closely the training data and it cannot be generalized. Vice versa, if the accuracy is low on the training set and high on the test set, this means that you are underfitting the model so the model is not efficient.

In this case, the test accuracy is 79% and the train accuracy is 81%. The differences between the two is not significant so this means that the model is neither overfitting nor underfitting.

3. How can you increase the performance of the model?

The performance of a model can be increased by using an approach called logistic regression L2 regularization. This classification method consists in minimizing the cross entropy with L2 regularization. L2 is the squared norm of w, which is multiplied by lambda and it consists in favoring certain weights over a big number of features. By calculating the derivatives of the equation, we obtain the term “ 2λw”, which has to be implemented in the code by adding it to the derivative with respect to w and not to the derivative with respect to b because b doesn’t occur in the regularization term.



```
lambda_ = 0.01
def logreg_train(X,Y, lambda_, lr, steps):
    m,n = X.shape
    w = np.zeros(n)
    b = 0
    accuracies = []
    losses = []


    for step in range(steps):
        P = logreg_inference(X, w, b)
        if step % 1000 == 0:
            loss = cross_entropy(P,Y)
            prediction = (P>0.5)
            accuracy = ( Y== prediction).mean()
            print(step,loss, accuracy*100)
            losses.append(loss)
            accuracies.append(accuracy)
        grad_w = (X.T @ (P - Y)) / m + 2 * lambda_ * w
        grad_b = (P - Y).mean()
        # Gradient descent updates.
        w -= lr * grad_w
        b -= lr * grad_b
    return w, b, losses, accuracies
```

Note: The modifications ,relative to the regularization, added to the code, are the ones circled in red.

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.