

Homework 1

SAAS DF

Spring 2025

Only problems marked with (*) are expected to be completed. However, the other ones could be fun, and a good primer for future classes or just for fun!

1 Teamwork Makes the Dream Work (*)

Teamwork Makes the Dream Work (*). For this homework, you will be part of a group of 5. Each person in the group should each train one classification model on the `Titanic-Dataset.csv` dataset using `scikit-learn` to predict the `survived` category.

Please see `transcript.txt`, which contains the roasts people mentioned on Wednesday's lab. This will help you ensure you don't get roasted for the same reasons in the future.

Please **report any statistics or plots you feel are relevant** for your model's performance. Additionally, each member of the group should **write 2 sentences** commenting on the validity of the model, or insights it gives. Discuss with your team to make sure each person hits different points. If you run out of ideas, ask for help!

2 More on Logistic Regression

Is Logistic Regression Interpretable (*). For a binary classification model to be calibrated, the confidence score it outputs should be close to the true proportion of samples of class 1. It turns out that using entropy as the loss function for logistic regression actually encourages it to be "calibrated". Let us make an argument for that here.

1. The loss function for logistic regression is binary cross-entropy: $L(y, f(\theta)) = -y \log f(x) - (1 - y) \log(1 - f(\theta))$. How can we find the minimum? Note that y is a constant given by our data.
2. Now it turns out that for the sigmoid function, $\sigma'(\theta x) = x \sigma(\theta x)(1 - \sigma(\theta x))$. Plug in $\sigma(\theta x)$ for f above (this is logistic regression) and see what drops out.
3. You should have found that $y = \sigma(\theta x)$. Now in reality, the loss above would sum over all y_i and inputs, so your result would actually be $\sum_i y_i = \sum_i \sigma(\theta x_i)$. Explain whether or not this means the logistic regression model is calibrated or not.

If you are interested, please check out these additional references for more insight on calibration:

1. Scikit-learn Docs
2. StackExchange Forum, make sure to read critically with a skeptical eye, but the second response essentially covers the solution to this question more generally.

Optimal Logistic Regression (*). In the previous problem, you should have found the derivative of the logistic regression loss function with respect to θ . Suppose your data is nicely separated so that when $x_i < 0$, $y_i = 0$, and when $x_i > 0$, $y_i = 1$. What is the sign of the derivative?

What does this mean about gradient descent with logistic regression?

3 Exploring Other Methods

Reading Exercise. We didn't get to cover much on certain classification metrics. Please look into the following:

- A Classification Method: k -nearest neighbors
- A Clustering Method: Agglomerative Clustering

Write a brief description of how these two methods are alike! What method can you use to determine a good value for k in k -nearest neighbors, or how many clusters to use with agglomerative clustering? See our lecture slides for some slides on agglomerative clustering which we didn't get to touch on during lecture.

4 Hints

More Logistic Regression Troubles: Is Logistic Regression Interpretable. *As a hint for the first part, try taking the derivative with respect to θ and setting it to zero!*

Optimal Logistic Regression. *You should find that the sign of the derivative is always positive.*