DISCORSO PRESENTAZIONE

Slide 2 - Dataset

Our project consists of predicting if a passenger of a flight is going to be "satisfied" or not. The dataset that we used is relatively large and contains many features that we can make use of for our goal, such as the age and gender of the customer, and ratings of various services gathered from surveys after the flight.

Slide 3 - Preprocessing

The first thing we look for is correlation between the features of this dataset to the customer satisfaction, which is our target. We found that most features are highly correlated with satisfaction, but 5 of them have almost no correlation, so we decided to drop them.

About 400 rows contain missing values, and since our dataset contains more than 120 thousand rows, we decided to drop these rows too.

We used One-Hot encoding to encode the qualitative features of our dataset, and we rescaled every feature into the [0, 1] range, which results in much faster convergence for our methods.

Slide 4 – Model & Training

We've decided to use 70% of the dataset for training and 30% for testing. Since this is a binary classification task, we decided to use Logistic Regression for it.

In particular, we used three models for our training: Gradient ascent, Newton's method and Gaussian discriminant analysis.

For Gradient Ascent, we first tried without rescaling our values and the model was taking a lot of iterations to converge, and required an extremely low learning rate. After rescaling the values into the [0, 1] range, we manage to converge within one thousand iterations with a generous 0.3 learning rate.

When using Newthon method, we had to use a very small fraction of our dataset in order to be able to compute the inverse of the hessian matrix. Suprisingly, even with as little as 1% of the dataset used for training, we managed to achieve 80% accuracy. It must be noted that, because of this and because of the high number of features, the results are very susceptible to variations on the training set.

{{{{{{{This is because Newton's method converges to "saddle points", which proliferate in high-dimensional spaces, such as the one of our feature-set}}}}}}}

With Gaussian Discriminant Analysis we use a 70%/30% split in training and testing, and we achieve the highest accuracy with 83% correct predictions, with very short computation time.

Slide 5 – Results

As we can see from the results, Gradient Ascent seems to be the most reliable: it converges relatively fast, has very good accuracy and it has the greatest area under the precision/recall curve.

In the case of Gaussian Discriminant analysis, we can see that it has the worst area under the precision/recall curve of the 3 methods, and this might be because not all features follow a gaussian distribution.

Slide 6 – Conclusions

In conclusion we compared the 3 methods by looking at their precision/recall curves and found that Gradient Ascent seems to be the best one, but interestingly they all seem to “meet” at a point, where GDA peaks higher than the other two.

Anyway, all three methods obtain fairly good results, with more than 80% correct predictions overall.

Considering Newton’s unreliable results, we would avoid using it in a real scenario for this kind of problem. Since Gradient Ascent requires less assumptions and produces good results, we would settle using this method to generate our predictions.

Slide 7 – Literature

These were our references while working on the project.

We used Andrew Ng’s notes as reference for the theory behind the models used.

For considerations on Newton’s method’s unreliability when working in high dimensional spaces, we referred to the second link

And lastly, we took inspiration for our choice of features from the linked notebook on kaggle.

Thanks for listening.