CS-433 Machine Learning Project 1

Matthias Minder, Zora Oswald, Silvan Stettler

Abstract—

INTRODUCTION

Collision events in the Large Hadron Collider at CERN do not create directly observable results. An essential part of elementary particles such as the Higgs boson therefore rely on classifying the collisions based on a number of variables that can be measured in the collider. In our dataset, a vector containing 30 dimensions describes one collision event. A binary classification algorithm can be used to predict the presence of a Higgs boson based on this feature vector. The model is trained with data from collisions for which the existence of a Higgs boson is determined. Our classifications of the collision events consist of three main steps:

1) Normalization and imputation of missing data
2) Parameter optimization using cross-validation
3) Predictions on test data using the best-fitting binary model

Furthermore, the importance of variables for the final model was assessed using best subset selection.

METHODS

As a first step, all features were transformed to have mean zero and unit variance, disregarding missing values.

As a second step, missing values were imputed. These missing values are due to physical measures being made impossible under certain conditions. However, since the classification methods used in the scope of this project don't naturally support the presence of missing values, they were imputed using the following linear regression approach:

Analysis of the raw data showed that there are a total of six distinct patterns of missing features. A pattern of observation $\boldsymbol{x}$, denoted $P(\boldsymbol{x})$, is characterized by values of $\boldsymbol{x}$ missing in dimensions M and values being present in dimensions M'. Approximately 60'000 observations of the training data contained no missing values, i.e. $M_{P(\boldsymbol{x})} = \emptyset$. For each incomplete pattern $P_i(\boldsymbol{x})$ and every missing feature $l \in M_{P_i(\boldsymbol{x})}$, a linear regression was fitted to complete observations, taking all features $k \in M'_{P_i(\boldsymbol{x})}$ as observations and feature l as response. This fit was then applied to impute the missing feature l of all $\boldsymbol{x}$ corresponding to $P_i(\boldsymbol{x})$. All linear regression models were fitted using gradient descent. Missing values of the test data were imputed using the fits on the training data.

This method for missing value imputation was chosen because it follows the natural structure of the data: Observations corresponding to the same physical preconditions leading to a specific pattern of missing values were all subjected to the same fit for missing value imputation. Simple linear regression was chosen due to its easy interpretability. Fitting more complex regression models would only make limited sense, since the absence of "true" values for a given missing value pattern makes reasonable model comparison impossible. Finally, we chose this approach over "simple", constant imputation using the feature mean or median because it allows to capture more of the data variability.

However, by imputing missing values, the information about their underlying physical reasons is lost. To capture this information, dummy variables were created that encode every non-complete pattern of missing values.

Classifying the collision events comes down to finding a relationship between the continuous and independent inputs and a binary output describing the presence or absence of a Higgs boson. As seen in class, logistic regression lends itself well to this task and was thus the first model that we tested. The probability that a Higgs boson is present is given by the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in [0, 1]$$
$$z = \boldsymbol{X}\boldsymbol{w}$$

For values $\in (0.5, 1]$, the collision is classified to have yielded a Higgs boson and assigned the label 1. For values below 0.5, the label 0 is assigned respectively. A L_2-regularized cost function was used in order to avoid that the weight vector $\boldsymbol{w}$ tends to infinity.

$$\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} ln(1 + e^{\boldsymbol{x}^T \boldsymbol{w}}) - y_n \boldsymbol{x}^T \boldsymbol{w} + \frac{\lambda}{2} ||\boldsymbol{w}||^2 \qquad (1)$$

The above cost function was minimized using the stochastic gradient method.

The feature vector $\boldsymbol{X}$ in the above equations contains a constant term. In addition, the influence of enhancing the feature vector with the squares of each value was assessed. This enhanced feature vector has the form

$$\boldsymbol{x}_{enh.} = [1 \ \ \boldsymbol{x}_n \ \ \boldsymbol{x}_n^2] \qquad (2)$$

In addition to logistic regression, the collisions were classified with a Support Vector Machine (SVM) method.

The basic idea behind SVM methods is to find a line that separates the datapoint into two classes. This can be achieved by minimizing a cost function based on hinge loss.

$$\mathcal{L}(\boldsymbol{w}) = C \sum_{n=1}^{N} max(0, 1 - y_n \boldsymbol{x}_n^T \boldsymbol{w}) + \frac{1}{2} ||\boldsymbol{w}||^2 \quad y_i \in -1, 1$$

(3)

If the output is predicted on the right side of the line, meaning that $sgn(y_i) = sgn(\boldsymbol{x}_n^T \boldsymbol{w})$, then that particular datapoint does not contribute to the loss. Points on the wrong side of the line or too close to the line contribute, however. Therefore, the gradient of the hinge loss $\nabla \mathcal{L}$ either takes the value $\boldsymbol{0}$ or $-y_n \boldsymbol{x_n}$ plus the contribution of the penalty function $\lambda \boldsymbol{w}$. This expression for the gradient was used in order to minimize the SVM cost function by stochastic gradient descent.

Both L_2-regularized logistic regression and SVM depend on a hyperparameter which controls the penalization of false classification. This parameter was chosen to maximize accuracy, as determined in ten-fold cross-validation.
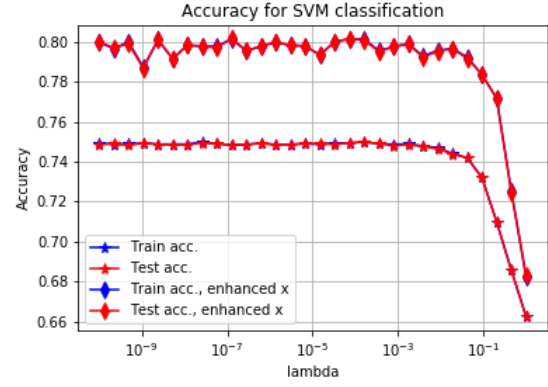

Figure 2. 10000 iterations, γ = 0.001

CONCLUSION

REFERENCES

[1]

RESULTS

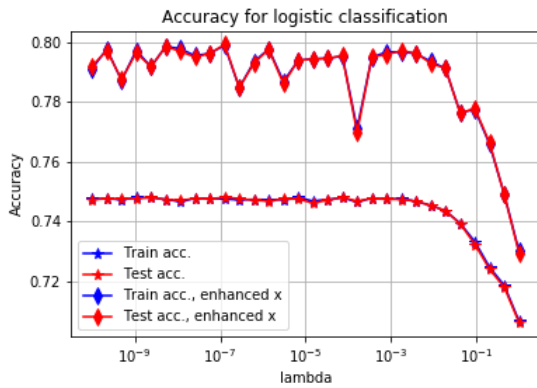The models were evaluated and tuned using ten-fold cross validation.

Figure 1. 10000 iterations, γ = 0.001