ML_Lab1

Marc Sparhuber

Table of contents

Task 1

Would you characterize this data set as being high-dimensional or low-dimensional? Based on this, do you expect that a standard logistic regression will work well for the purpose of prediction?

```
[1] 1003 1496
```

> I would consider this to be a high-dimensional data set as it contains more variables (columns) than observations (rows). I would expect a standard logistic regression to perform rather poorly and perhaps result in perfect fits for the given training data.

Task 2

a.

Extract the coefficients from the estimated model using the coef() function and inspect the coefficients that are placed 1010–1050 in the output from coef(). Do you notice anything special?

> Yes, this range of coefficients is exclusively NAs!

b.

Examine the training accuracy of the estimated model. What does this result suggest about the predictive capacity of the model?

```
[1] 1
```

> While the *training* accuracy is perfect, at 100%, this in no way means that its predictive capacity is similarly good. The *test* accuracy, which is what truly matters during a prediction task is likely much lower and worse due to the perfect *training* accuracy. This is due to overfitting, which entails f^ being so flexible that it models the noise contained in the data instead of successfully approximating the true f. In other words, the predictive capacity of this on new data would be bad because our classification model is too flexible which results in overfitting.

Task 3

Use the caret package to implement a 3-fold cross-validation procedure that estimates the test accuracy of a standard logistic regression. Report the accuracy. Does this result align with your expectations from #1 and #2? Do the results from #2 and #3 provide any indications of either over- or underfitting?

> The accuracy is ~0.52. This indicates an accuracy that is only slightly better than guessing randomly who wrote the tweet, which is in line with the previous answers to tasks 1 and 2. As stated in the answer to task 2b., the results indicate overfitting.

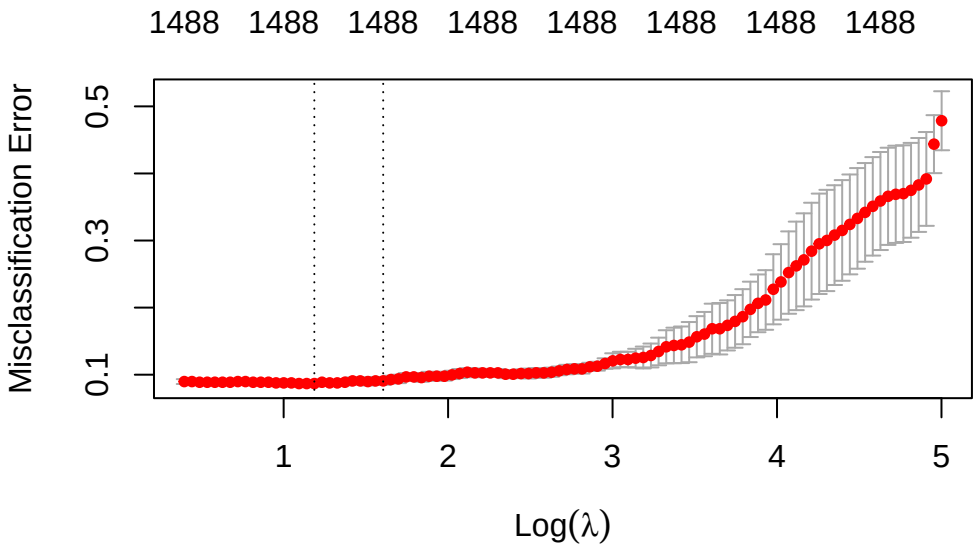
Task 4

Now we shall move beyond the standard logistic regression, and more specifically, turn to ridge regression for our prediction task. This importantly entails deciding on a value for the parameter . Use glmnet's function cv.glmnet to find the that minimizes the test error, and report the associated test accuracy. Is this a better or worse prediction model compared to the one in #2–3? Which of the two models do you believe have a higher variance? Why?

> Looking at the output of the cv.glmnet object, it can be seen that the lambda that minimizes the test error is 3.276. The test accuracy associated with this lambda is ~0.91, which seems quite good and a better prediction model than the previous. I calculated this by subtracting the misclassification rate ("Measure") from 1. As a flexible fit and high variance go hand in hand and the previous model was extremely flexible, I would reason that this one has a lower variance.

Task 5

Plot lambda against the misclassification error. Interpret the plot in terms of bias and variance.



The greater lambda gets, the greater the misclassification error. The dotted line on the left minimizes the mean squared error. As the bias-variance trade-off results in the error following a sort of U-shape it can be assumed that this point lies at the center of this U, with (squared) bias increasing to the left and variance increasing to the right.

Task 6

Lastly, extract the coefficients associated with the lowest test error. Have a closer look at the coefficients with the largest positive and largest negative values. What do they reveal? Do the words you find on either side confine to your expectations?

The words with high coefficients reveal, i.e., words expected to be in Trump tweets, what seems to be more conservative and vague language, whereas tweets which are expected to be in Bernie's tweets seem to contain more verbs. Further analysis here could be done to further distinguish the two's messaging on Twitter.

Task 6

Begin by importing the file "Kaggle_Social_Network_Ads.csv". Format the outcome variable Purchased as a factor variable (this is required for the subsequent analysis using the caret package).

Task 7

Use the caret package to implement a 5-fold cross-validation that assesses the test accuracy of a standard logistic regression model. Report its test accuracy.

> The test accuracy is ~0.85.

Task 8

To investigate whether GAMs can improve the performance over the standard logistic regression, implement three separate 5-fold cross-validations; each estimating a GAM with a different degree of freedom for the natural cubic splines (df {2, 3, 4}). Create splines for the two variables Age and Salary, but not for Gender, which is a categorical variable (hint: use the ns() function from the splines package to create splines). Again, to ensure identical folds, add set.seed(12345) above each train(). You may also re-use the trainControl object from the previous task. Report the accuracies of the different models.

> Across the three models with increasing degrees of freedom for the natural splines, the model with 2 degrees of freedom produces the highest estimated median accuracy, with ~0.91, whereas the other two are at 0.90 and ~0.90, respectively.

a.

Do you observe any improvement compared to the standard logistic regression?

> I do. All three of the GAMs outperform the standard logistic regression model.

b.

What does the difference in performance between the standard logistic regression and the GAMs suggest about the former? Is it over- or underfitted? Does it have high(er) bias or high(er) variance compared to the GAMs?

> IT suggests that the added complexity from the GAMs adds predictive power to the model. As the model was made more flexible by adding the natural splines compared to the standard logistic regression, complexity increased. As complexity increased and predictive accuracy rose, it can be concluded that the previous model was underfitted. Bias decreased because the GAMs were able to capture the more complex nature of the data and variance at least slightly increased because the more flexible a model is, the more sensetive it is to new training data.

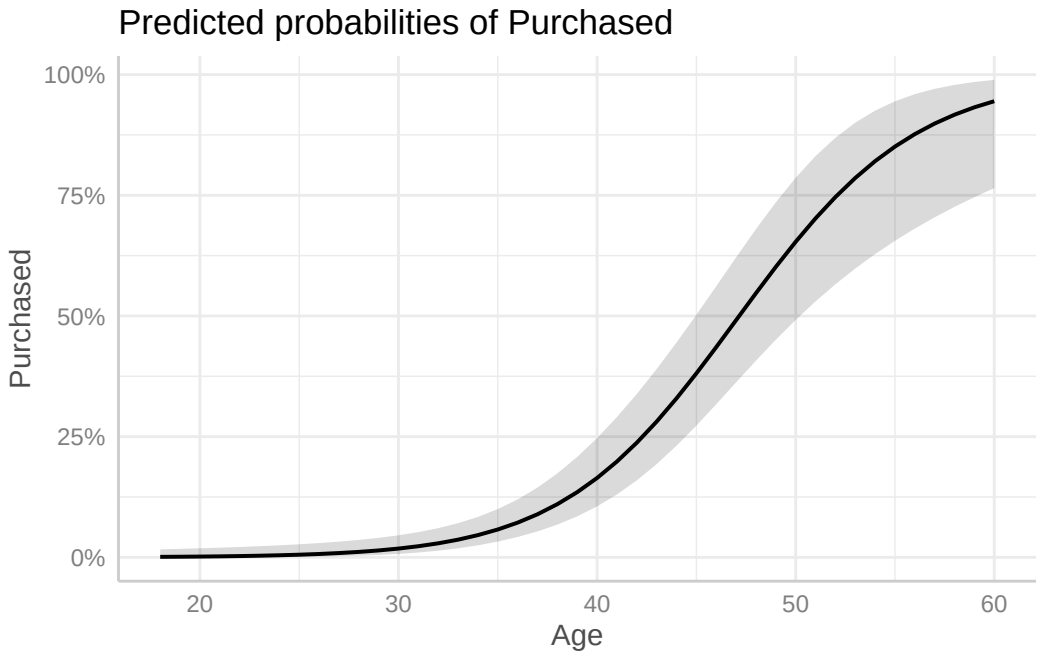
c.

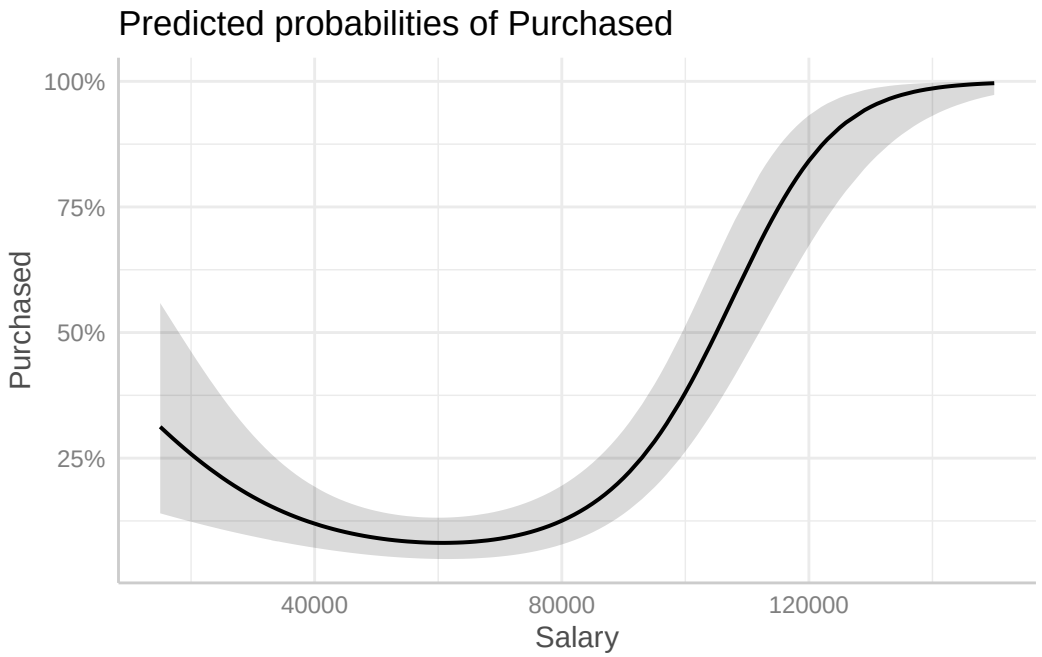
Which of the three GAMs do you prefer? Motivate.

> Of the three I would prefer the one with the fewest degrees of freedom, as it increases the estimated test accuracy by the most, based on the k-fold cross validation. The models with higher degrees of freedom seem to already add too much flexibility/complexity.

Task 9

Next, you shall examine the predictive relationships between the two continuous variables (Age and Salary) and the outcome. For this, you will use ggeffects's ggpredict() function which computes predictions while varying one variable and holding the remaining fixed at their means/mode. To do so, first re-estimate the GAM-specification that you found to be the best, on the full data using glm (ggeffects does not accept objects from caret). Interpret. Do you find any non-linear relationship?





Whereas the relationship between a positive purchasing decision and Age is nearly linear, the relationship between a positive purchasing decision and Salary isn't, declining from around ~1.32 at a salary of less than 20000 to a floor of ~1.17 at a salary of around 60000 to then increase near-linearly. These results suggest that purchases facilitated by online ads correlate with greater age but that individuals with mid-range salaries have different are less affected by ads in their purchasing decisions than people with more or less money at their disposal. Looking at the confidence bands suggest that greater the age, the less accurate the prediction may be, with the same effect occuring at the lower end of the salary distribution. This is perhaps due to fewer observations at the extremes.

Task 10

In this second part of the lab, we used GAMs to improve predictive performance. Would we expect to see similar improvements if we instead had used ridge and lasso regression? Why/why not?

We would expect the opposite, as both ridge and lasso approaches decrease complexity by, e.g., setting inconsequential coefficients to 0. This would have likely led to a less flexible fit and resulted in lower variance, higher bias and worse predictive accuracy.