Problem 1.1

Dataset

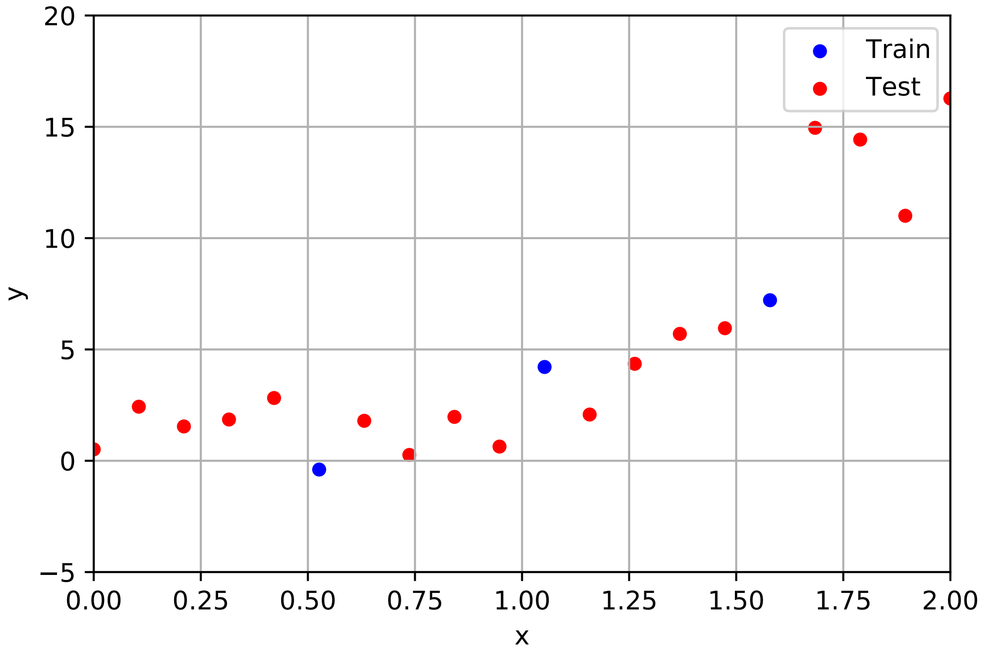


Figure 1: Scatterplot of train and test dataset

Error for different degree polynomials

In Figure 2 the stemplot of the MSE over the test set is shown, the errors are logarithmically scaled. The smallest error is achieved for the degree $m = 1$.

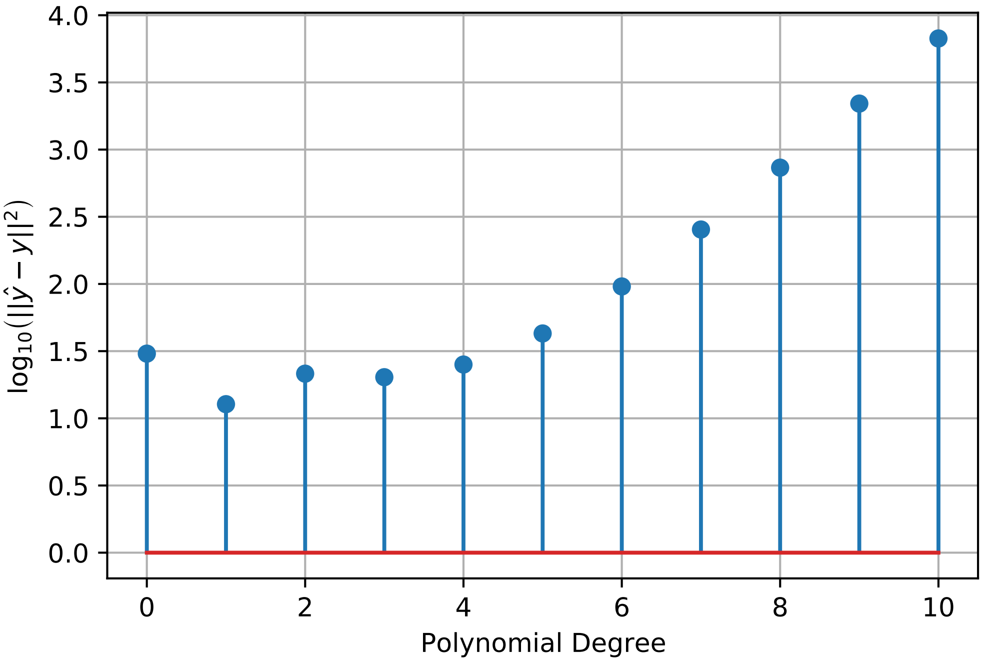
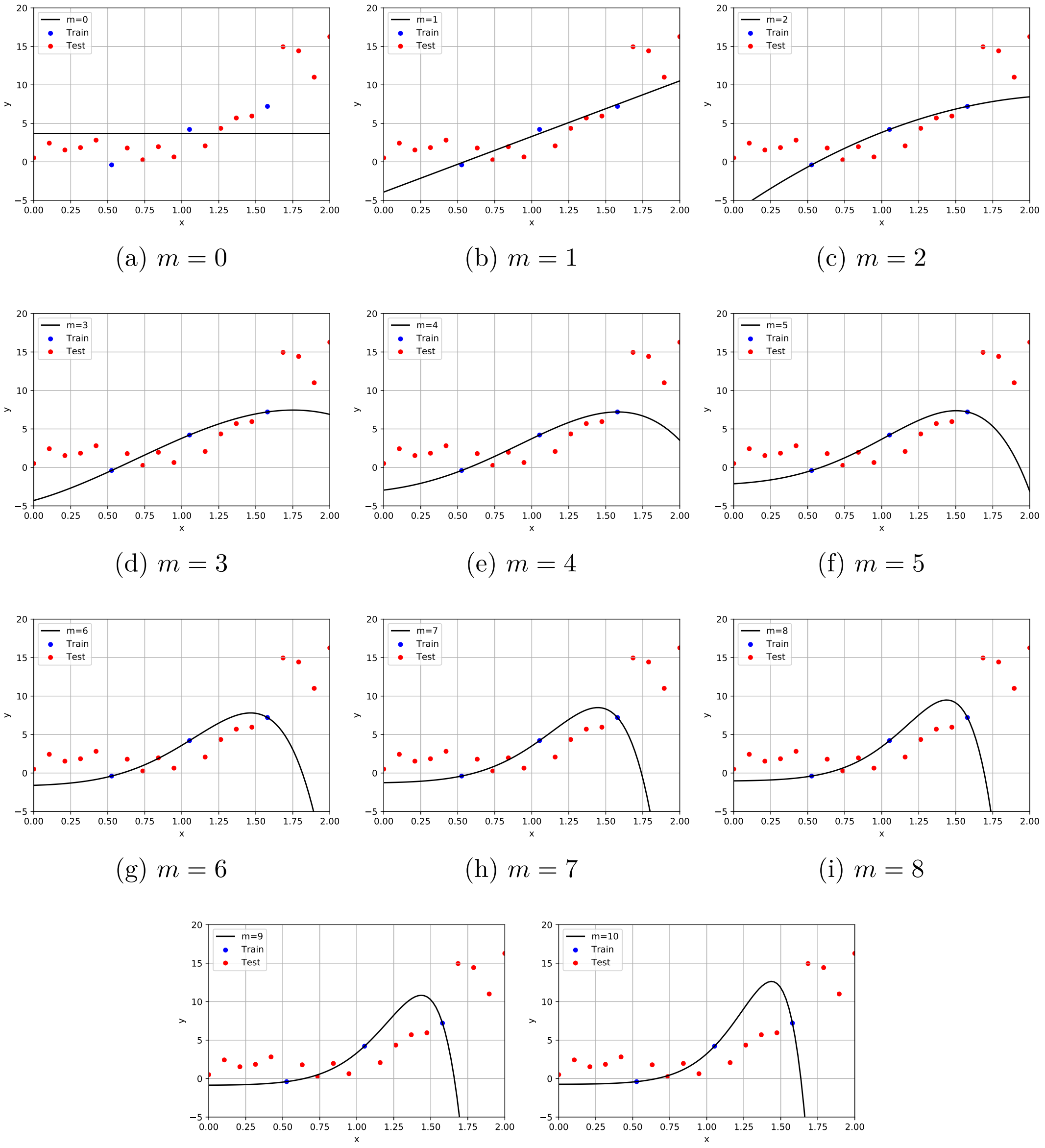


Figure 2: Stemplot showing the MSE for different degrees of the polynomial

Polynomial fit

In Figure 3 the polynomial fit is shown for different degrees on top of the two datasets. The value $m = 1$ leads to the smallest error as was shown above. For smaller values of m the regressor is to simple (underfitting). The train set contains only three points, therefore for polynomials of degree $m \geq 2$ the points in the train set can be captured exactly. Higher degree polynomials lead to very steep behaviour for $x > 1.5$, where there are no train datapoints anymore. Therefore the regressor generalizes poorly and has high error on the test set. So for $m \geq 2$ there is overfitting.



(a) $m = 0$ (b) $m = 1$ (c) $m = 2$

(d) $m = 3$ (e) $m = 4$ (f) $m = 5$

(g) $m = 6$ (h) $m = 7$ (i) $m = 8$

(j) $m = 9$ (k) $m = 10$

Figure 3: Polynomial fit for various different degrees m of polynomial

Ridge Regression Analytical

Cost function to be minimized

$$J(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda\mathbf{w}^T\mathbf{w} \tag{1}$$

with the matrix calculus rules given in the problem set:

$$\frac{\partial J\left(\mathbf{w}\right)}{\partial \mathbf{w}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} \tag{2}$$

by setting the derivative to zero the optimal solution $\hat{\mathbf{w}}$ can be obtained.

$$\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = \mathbf{X}^T\mathbf{y} \implies \hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

Ridge Regression Errors

The errors over the test set for different values of λ are shown in a log-log-plot in Figure 4. The smallest error is achieved for $\lambda = 5$.

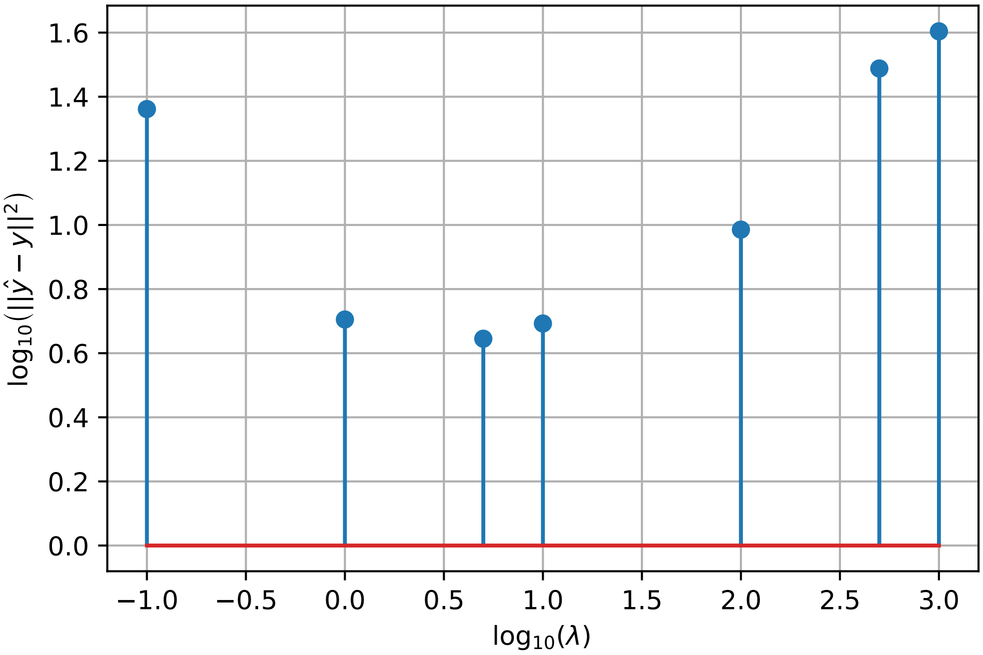
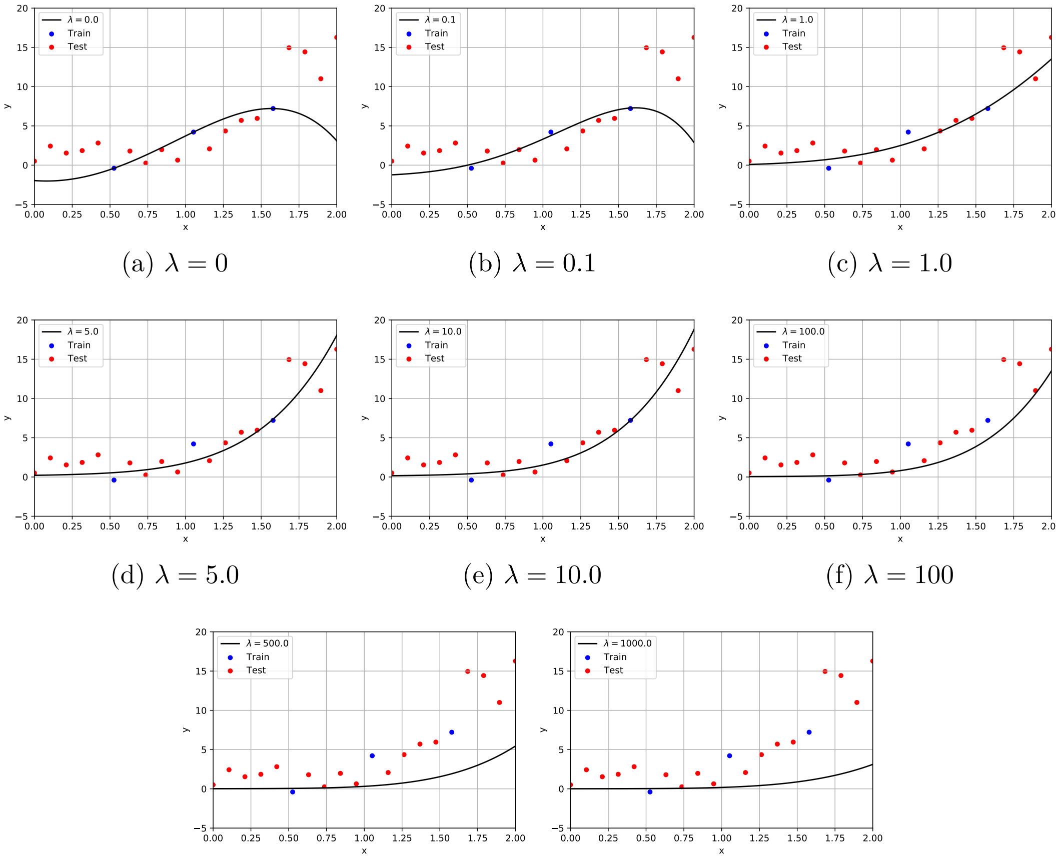


Figure 4: Stemplot showing the MSE for different values of λ

Ridge regression fit

For low values of λ (0 or 0.1) no regularizing effect is visible yet. The values 1.0 and 5.0 and 10.0 deliver relatively good results with small error. For higher values of λ the norm of $\mathbf{w}$ gets penalized more and more which forces the weights to become very small and therefore forces the polynomial towards zero. The error becomes higher for too high values of λ. A correct setting of λ can prevent overfitting.



(a) $\lambda = 0$

(b) $\lambda = 0.1$

(c) $\lambda = 1.0$

(d) $\lambda = 5.0$

(e) $\lambda = 10.0$

(f) $\lambda = 100$

(g) $\lambda = 500$

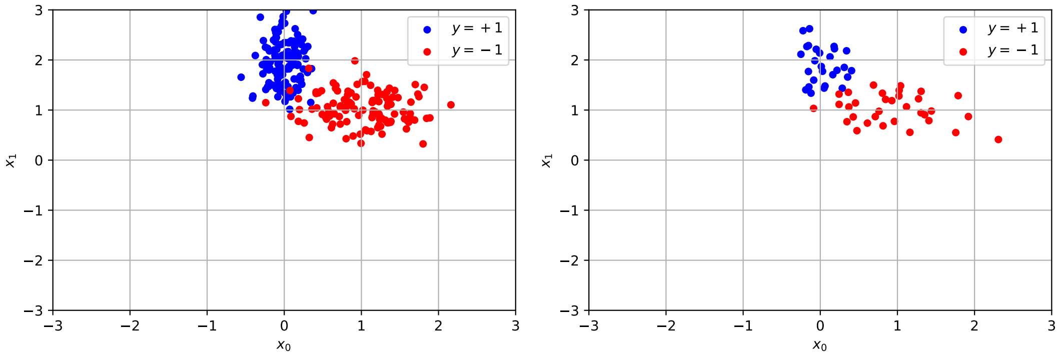
(h) $\lambda = 1000$

Figure 5: Ridge regressor fit for various different values of λ

Problem 1.2

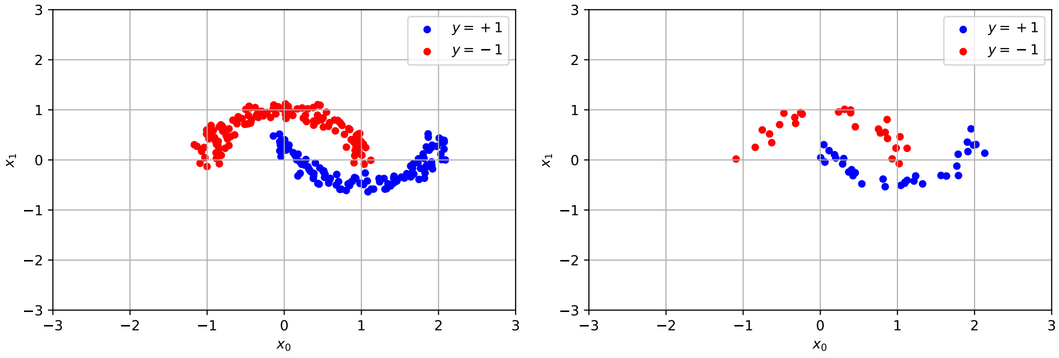
Dataset

The two datasets blob and moon are shown in Figure 6 and Figure 7. For each dataset the train and test set are shown separately and the two labels $\{-1, 1\}$ are shown in different colours.



(a) Train Data (b) Test Data

Figure 6: Scatterplot of train and test dataset blob



(a) Train Data (b) Test Data

Figure 7: Scatterplot of train and test dataset moon

Accuracy linear classifier

The accuracy of the linear classifier for the two datasets are shown in Table 1 and Table 2. For both, the accuracy is shown on the train and test set separately.

The accuracy for the moon dataset is worse because a linear classifier is not an optimal choice for this dataset as can be seen in the scatterplot.

Train Accuracy	97.92%
Test Accuracy	95.00%

Table 1: Accuracy of the linear classifier for blob dataset

Train Accuracy	87.08%
Test Accuracy	88.33%

Table 2: Accuracy of the linear classifier for moon dataset

Heatmap

A heatmap of the soft-label output is shown in Figure 8 and Figure 9 for the two datasets.

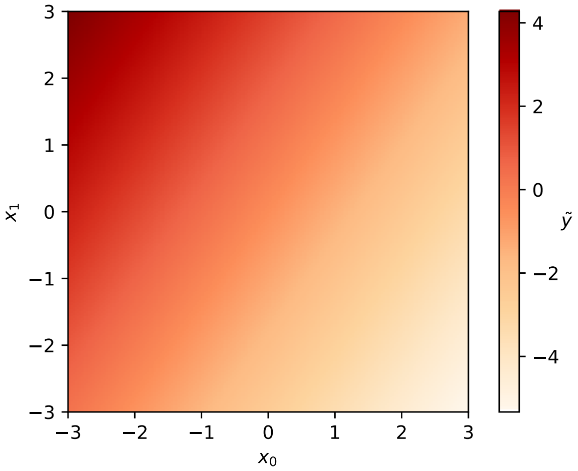


Figure 8: Heatmap blob dataset

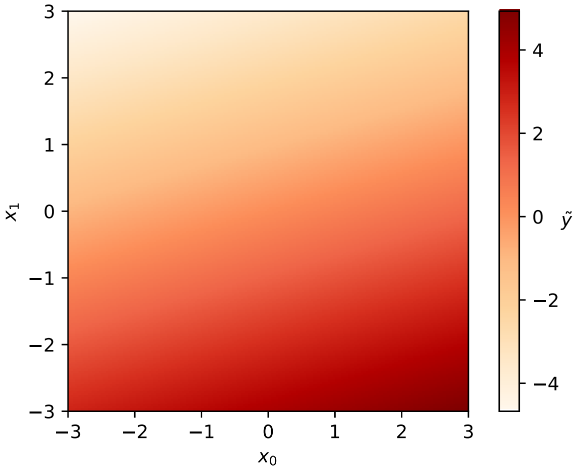
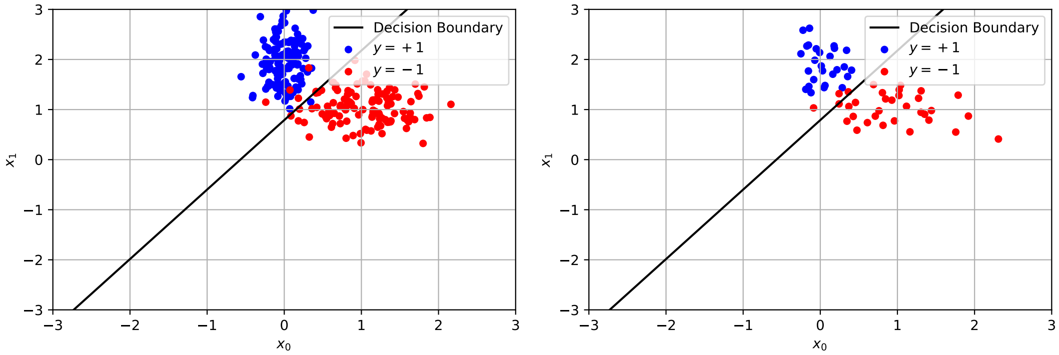


Figure 9: Heatmap moon dataset

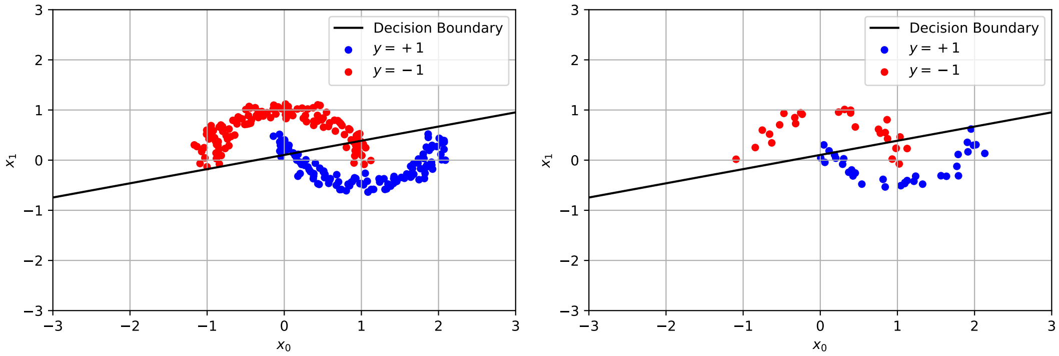
Decision Boundary

The decision boundary separating the two classes for the blob dataset is shown in Figure 10. For the moon dataset it is shown in Figure 11. As can be clearly seen, this dataset consists of two interleaving half circles. Therefore, a linear classifier is not the appropriate choice for this dataset. To better separate the two classes an S-shaped decision boundary would be needed, which cannot be achieved by a linear classifier.



(a) Train (b) Test

Figure 10: Scatterplot with decision Boundary



(a) Train (b) Test

Figure 11: Scatterplot with decision Boundary moon dataset

Loss comparison soft and hard labels

For the training of the classifier the MSE on the soft-labels has been chosen as loss function. For the two samples $s_1 = \{[-10, 10]^T, 1\}; s_2 = \{[0, 1.5]^T, -1\}$ the loss is shown in Table 3.

The sample s_1 has a loss value of 209.6 even though as can be seen from the hard-label loss, the sample is correctly classified. The sample s_2 has a loss of only 2.180 but is classified wrongly.

This indicates, that the MSE on the soft-labels is not ideal as loss function. A sample that is correctly classified but has a higher distance form the decision boundary contributes more the the loss function, than a sample that is close to the boundary and is classified wrongly.

		J_{LS}
s_1	Soft Label	209.6
	Hard Label	0
s_2	Soft Label	2.180
	Hard Label	4

Table 3: Loss comparison of soft and hard labels for the two samples s_1 and s_2

Problem 1.3

Dataset

As a quick sanity check, the correlation of the features with the median house value was computed. The values are reported in Table 4. The feature MedInc has the highest correlation with the target value.

MedInc	0.688 075
AveRooms	0.151 948
HouseAge	0.105 623
AveOccup	−0.023 737
Population	−0.024 650
Longitude	−0.045 967
AveBedrms	−0.046 701
Latitude	−0.144 160

Table 4: Correlation with MedHouseVal sorted highest to lowest

Scatterplot of Median House Value

A scatterplot of the datapoints in a longitude latitude diagram is shown in Figure 12, the color indicates the target value. The data outlines California and it can be seen that along the coast and near the large cities the target value is higher, further the density of data points is higher in those areas.

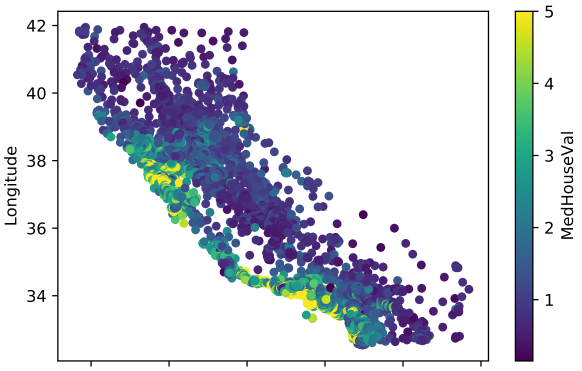
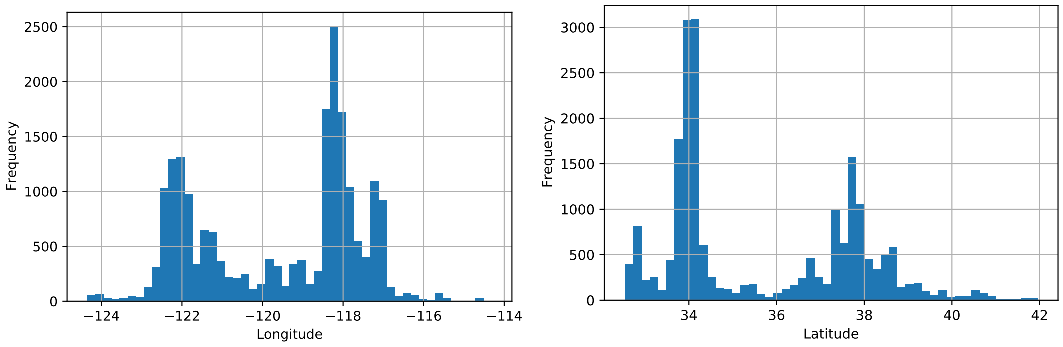


Figure 12: scatterplot longitude lattitude

In Figure 13 a histogram of longitude and latitude is shown. The two highest peaks are at approximately -122 and -118.2 for longitude and 34 and 37.8 for latitude. The four most populous cities in California are listed in Table 5. The cities Los Angeles and San Jose / San Francisco match up very good with the two highest peaks in the histograms.

City	Longitude	Latitude
Los Angeles	-118.24	34.05
San Diego	-117.16	32.72
San Jose	-121.89	37.34
San Francisco	-122.42	37.77

Table 5: Coordinates of the four most populous cities in California



(a) Longitude

(b) Latitude

Figure 13: Histograms of longitude and latitude

Weights of Ridge Regressor

In Table 6 the weights for the ridge regressor are shown. In Table 7 the intercept can be seen. The Features AveBedrms, MedInc, HouseAge have a positive impact on the house price, all the other features have negativ impact on the house price.

The intercept is very high, as for example the Longitude and Latitude have an Offset, those values are not centered around zero.

Feature	Weight
AveBedrms	$8.16 \cdot 10^{-1}$
MedInc	$4.48 \cdot 10^{-1}$
HouseAge	$9.59 \cdot 10^{-3}$
Population	$-6.09 \cdot 10^{-7}$
AveOccup	$-4.63 \cdot 10^{-3}$
AveRooms	$-1.33 \cdot 10^{-1}$
Latitude	$-4.63 \cdot 10^{-1}$
Longitude	$-4.36 \cdot 10^{-1}$

Table 6: Features and corresponding weights of the ridge regressor sorted by the weights

Intercept	-37.3

Table 7: Intercept of ridge regressor

Comparison of regressors

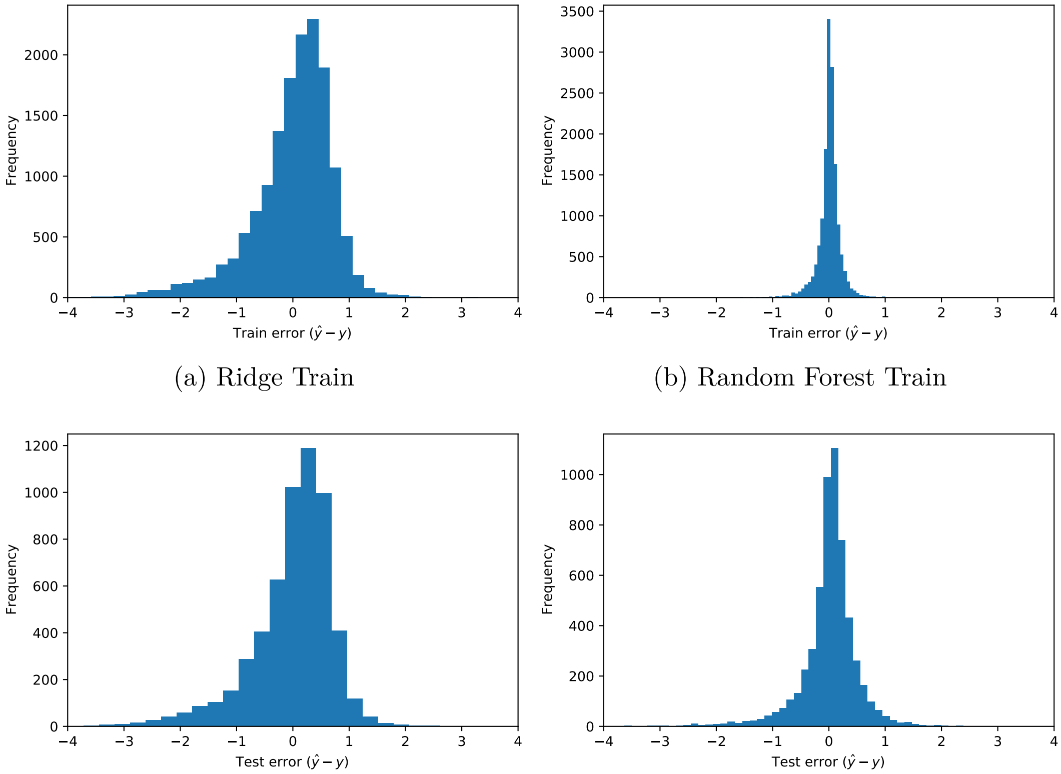
In Table 8 the mean average error has been computed on the train and test set for both regressors.

The ridge regressor performs approximately equal on the train and the test dataset. The random forest regressor has lower MAE on the train dataset than on the test dataset. This might indicate that the regressor is close to or already in an overfitting regime. Overall the random forest regressor performs better on the test set. This is to be expected because it can capture the more complex relationships in the data, which cannot be captured by a linear regressor.

		MAE
Ridge Regressor	Train	0.531
	Test	0.528
Random Forest Regressor	Train	0.125
	Test	0.331

Table 8: Mean Average Errors

In Figure 14 a histogram of the prediction error $\hat{y} - y$ is shown for the two regressors. For each a separate histogram is shown for the train and test set. The histograms of both regressors are approximately centered around zero, which indicates that both regressors are unbiased. The histogram of the random forest regressor is much narrower, the error variance is lower.



(a) Ridge Train

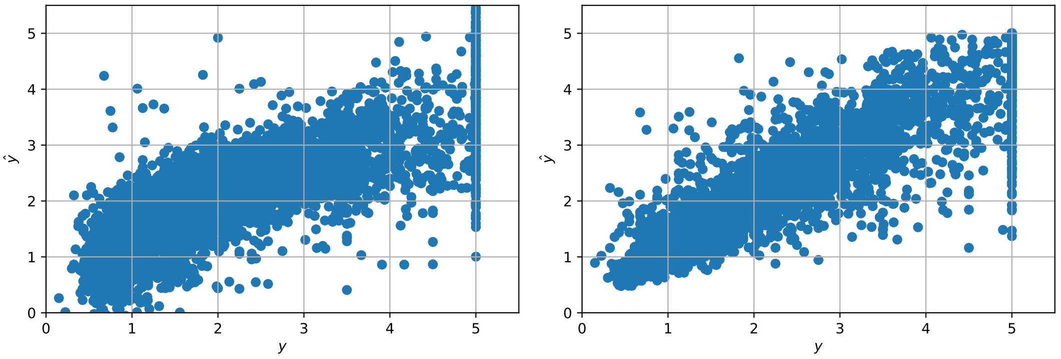
(b) Random Forest Train

(c) Ridge Test

(d) Random Forest Test

Figure 14: Error histgrams for ridge regressor and random forest regressor

In Figure 15 the predicted over the actual values are shown for both regressors over the test set. Ideally alle data points would lie on a straight 45° line. It can be seen that the random forest regressor performs better because the points are closer to the ideal ones.



(a) Ridge

(b) Random Forest

Figure 15: Scatter plot comparison of ridge regressor and random forest regressor over the test set

Cross Validation

The results of the 10-fold cross validation are reported in Table 9. The standard deviation is relatively low, indicating that the regressor performs comparable on all 10 splits of the cross validation.

Mean	0.811
Standard deviation	0.0097

Table 9: Mean and standard deviation of 10 fold cross-validation