DATA7703 Tutorial 10 Solutions

2021 Semester 2

1. Outliers and adversarial examples present very different challenges. Outliers are rare and very different from the inliers. They are considered as misleading data points and thus best to be removed. Adversarial examples are similar to regular observations, and the algorithms are expected to be able to give correct predictions on them.

2. One approach is to filter outliers first, then build a model.

Another approach consists of the subsampling methods. They make use of multiple random subsamples to find a robust model. The subsamples are typically chosen to be as small as possible so that they are less likely to contain an outlier. The Theil-Sen estimator and RANSAC are two examples of subsampling methods.

In addition, we also have the robust loss methods (aka M-estimators in statistics). These methods make use of a loss function which is robust against outliers, in the sense that they do not apply an excessively large penalty to outliers LAD and Huber regression are two examples of robust loss methods.

3. **(a)** We have one outlier $(4, 9)$, and three inliers $(1, 2), (2, 3)$ and $(5, 6)$. Draw a scatter plot for these points to convince yourself if needed.

(b) For the inliers, it is easy to see that the x and y values differ by 1 only, and thus the OLS model is $y = x + 1$.

For the entire dataset, some calculation is needed, and there are various ways to do this. For example, you can use the general closed-form formula for OLS to find the model. You can also use basic calculus to directly minimize the MSE to find the OLS model. Alternatively, you can use the formula for simple linear regression (https://en.wikipedia.org/wiki/Simple_linear_regression) to find the model. We illustrate the last approach here. The general formula is given by

$$y = \hat{\beta}x + \hat{\alpha},$$

where $\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, where $\bar{x}$ and $\bar{y}$ are the mean x and y values respectively.

With some calculation, we have $\bar{x} = 3$, $\bar{y} = 5$, $\sum_i (x_i - \bar{x})^2 = 10$, $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 14$. Hence $\hat{\beta} = 14/10 = 7/5$, and $\hat{\alpha} = 5 - 21/5 = 4/5$. That is, the OLS model is $y = \frac{7}{5}x + \frac{4}{5}$.

(c) For the inlier model $y = x + 1$, we can calculate the prediction $\hat{y}$ and residual r as follows

x	1	2	4	5
y	2	3	9	6
$\hat{y}$	2	3	5	6
r	0	0	4	0

The ℓ_1 loss of the inlier model is thus $\frac{|0|+|0|+|4|+|0|}{4} = 1$.

For the model $y = \frac{7}{5}x + \frac{4}{5}$, we can calculate the prediction $\hat{y}$ and residual r as follows

x	1	2	4	5
y	2	3	9	6
$\hat{y}$	$\frac{11}{5}$	$\frac{18}{5}$	$\frac{32}{5}$	$\frac{39}{5}$
r	$-\frac{1}{5}$	$-\frac{3}{5}$	$\frac{13}{5}$	$-\frac{9}{5}$

The ℓ_1 loss is thus $\frac{1}{4}\left(\frac{1}{5} + \frac{3}{5} + \frac{13}{5} + \frac{9}{5}\right) = 13/10$.

The inlier model is better than the model trained on the entire dataset in terms of the ℓ_1 loss.

(d) Recall that the Huber loss with a parameter δ is defined as

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta\left(|r| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases},$$

where r is the residual.

For the inlier model, the Huber loss $\frac{1}{4}(0 + 0 + 0.1(4 - 0.05) + 0) = 0.395/4$ when $\delta = 0.1$. Similarly, the Huber loss is $0.395/4$ when $\delta = 1$.

For the model trained on the entire dataset, when $\delta = 0.1$, the Huber loss is

$$\frac{1}{4}(0.1(0.2 - 0.05) + 0.1(0.6 - 0.05) + 0.1(2.6 - 0.05) + 0.1(1.8 - 0.05)) = 0.125$$

When $\delta = 1$, the Huber loss is

$$\frac{1}{4}\left(\frac{1}{2}0.2^2 + \frac{1}{2}0.6^2 + 1(2.6 - 0.5) + 1(1.8 - 0.5)\right) = 0.9.$$

The inlier model has smaller Huber loss (thus better) for both $\delta = 0.1$ and $\delta = 1$.

(e) There are 6 different pairs.

3 pairs only use the inliers, and they all lie on the straight line $y = x + 1$, thus the slopes for them are 1.

3 pairs use the outlier $(4, 9)$. The slopes are $\frac{9-2}{4-1} = \frac{7}{3}$, $\frac{9-3}{4-2} = 3$, $\frac{9-6}{4-5} = -3$.

Thus the 6 slopes are -3, 1, 1, 1, 7/3, 3. The slope m of the Theil-Sen model is the median slope $\frac{1}{2}(1 + 1) = 1$.

The $y_i - mx_i$ values are 1, 1, 5, 1. Thus the bias of the Theil-Sen model is their median $\frac{1}{2}(1 + 1) = 1$.

Hence the Theil-Sen model is $y = x + 1$.

4. (a) In general, choosing the model with maximum R^2 is different from choosing the model with minimum MSE, because the R^2 values are generally computed on different datasets.

To elaborate, recall that given n examples with labels $y_1, \ldots, y_n$, and if we predict $\hat{y}_1, \ldots, \hat{y}_n$ for these examples respectively, then the R^2 value is $R^2 = 1 - SSE/TSS$, where SSE (sum of squared error) is $\sum_i (y_i - \hat{y}_i)^2$, and TSS (total sum of squares) is $\sum_i (y_i - \bar{y})^2$, with $\bar{y}$ being the average of $y_1, \ldots, y_n$.

Note that $SSE = nMSE$. Thus if we are working with the same set of examples, then a model with larger R^2 has smaller MSE.

However, in RANSAC, assume that two candidate inlier models have R^2 values $1 - n_1 MSE_1/TSS_1$ and $1 - n_2 MSE_2/TSS_2$ respectively, where n_1 and n_2 are the numbers of inliers the two models are trained on respectively. We can see that $1 - n_1 MSE_1/TSS_1 < 1 - n_2 MSE_2/TSS_2$ is equivalent to $n_1 MSE_1/TSS_1 > n_2 MSE_2/TSS_2$, and this does not necessarily imply $MSE_1 > MSE_2$, because the values of n_1, n_2, TSS_1 and TSS_2 matter as well.

(b) A random subset may contain mostly outliers, and the outlier detector may pick mostly outliers to train a candidate inlier model. In this case, the candidate inlier model may have a good R^2 score, although it is far from being a good inlier model.

If we require the number of detected inliers to be sufficiently large, we can prevent this, because the poor outlier detector will fail to classify sufficiently many data points as inliers (since it mainly classifies outliers as inliers).