Regression metrics: (R)MSE, R-squared, MAE

Plan for the video

1) **Regression**
   - MSE, RMSE, R-squared
   - MAE
   - (R)MSPE, MAPE
   - (R)MSLE
2) **Classification:**
   - Accuracy, LogLoss, AUC
   - Cohen's (Quadratic weighted) Kappa

Plan for the video

1) **Regression**
   - MSE, RMSE, R-squared
   - MAE
   - (R)MSPE, MAPE
   - (R)MSLE
2) **Classification:**
   - Accuracy, LogLoss, AUC
   - Cohen's (Quadratic weighted) Kappa

Notation

- N – number of objects

- $y \in \mathbb{R}^N$ – target values

  $\hat{y} \in \mathbb{R}^N$ – predictions

- $\hat{y}_i \in \mathbb{R}$ – prediction for i-th object

  $y_i \in \mathbb{R}$ – target for i-th object

MSE: Mean Square Error

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

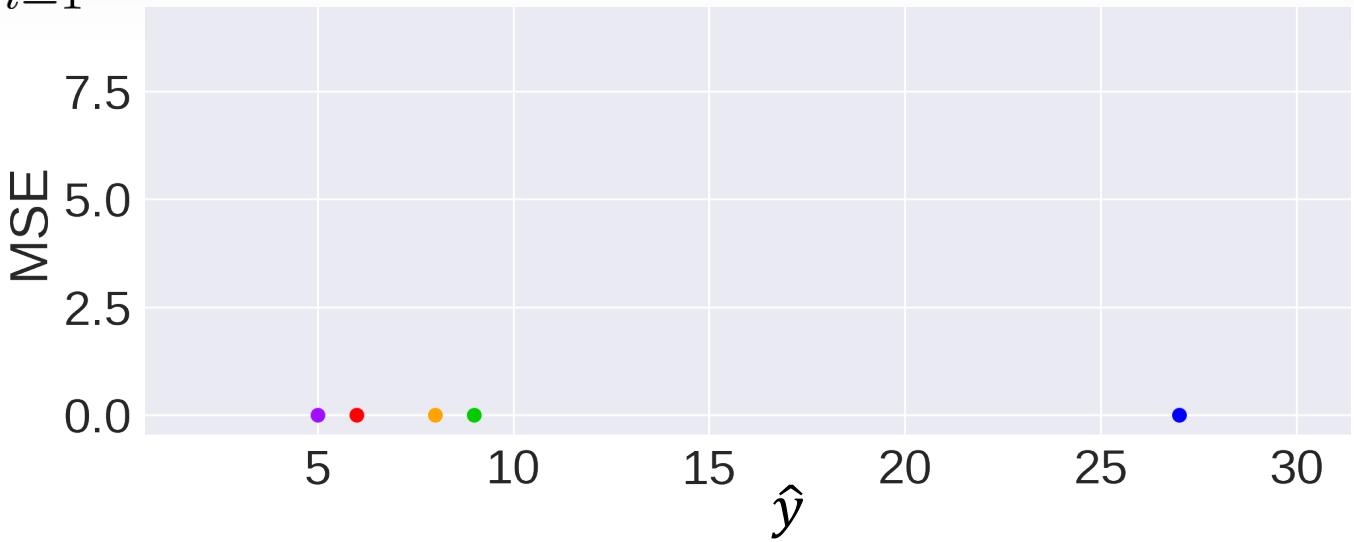
The first metric we will discuss is Mean Square Error. It is for sure the most common metric for regression type of problems. In data science, people use it when they don't have any specific preferences for the solution to their problem, or when they don't know other metric.

MSE: Mean Square Error

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Data:

X	Y
...	5
...	9
...	8
...	6
...	27

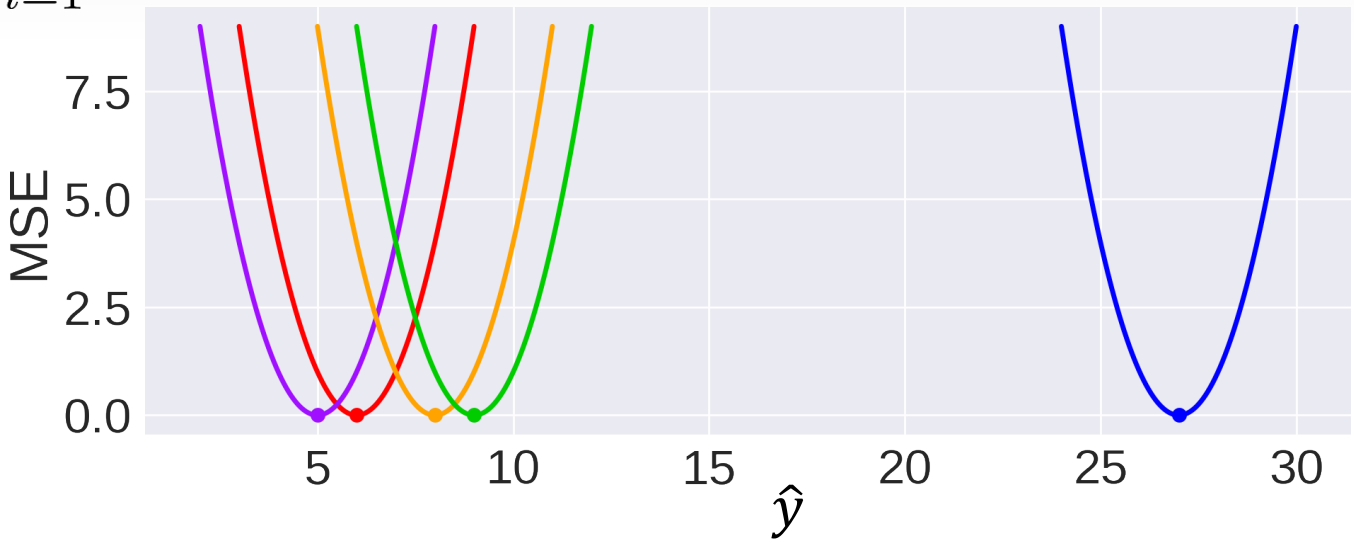


MSE: Mean Square Error

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Data:

X	Y
...	5
...	9
...	8
...	6
...	27



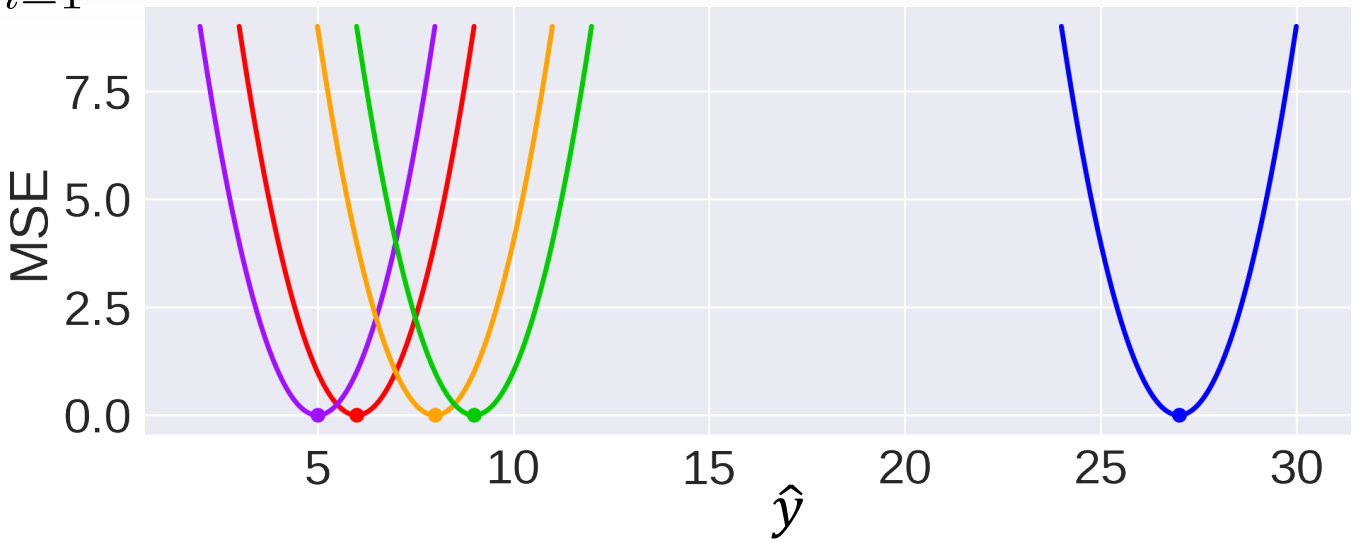
MSE: optimal constant

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha)^2$$

Best constant: ?

Data:

X	Y
...	5
...	9
...	8
...	6
...	27



But, what is the optimal constant? What constant minimizes the mean square error for our data set? In fact, it is easier to set the derivative of our total error with respect to that constant to zero, and find it from this equation. What we'll find is that the best constant is the mean value of the target column. If you think you don't know how to derive it, take a look at the reading materials. There is a fine explanation and links to related books

MSE: optimal constant

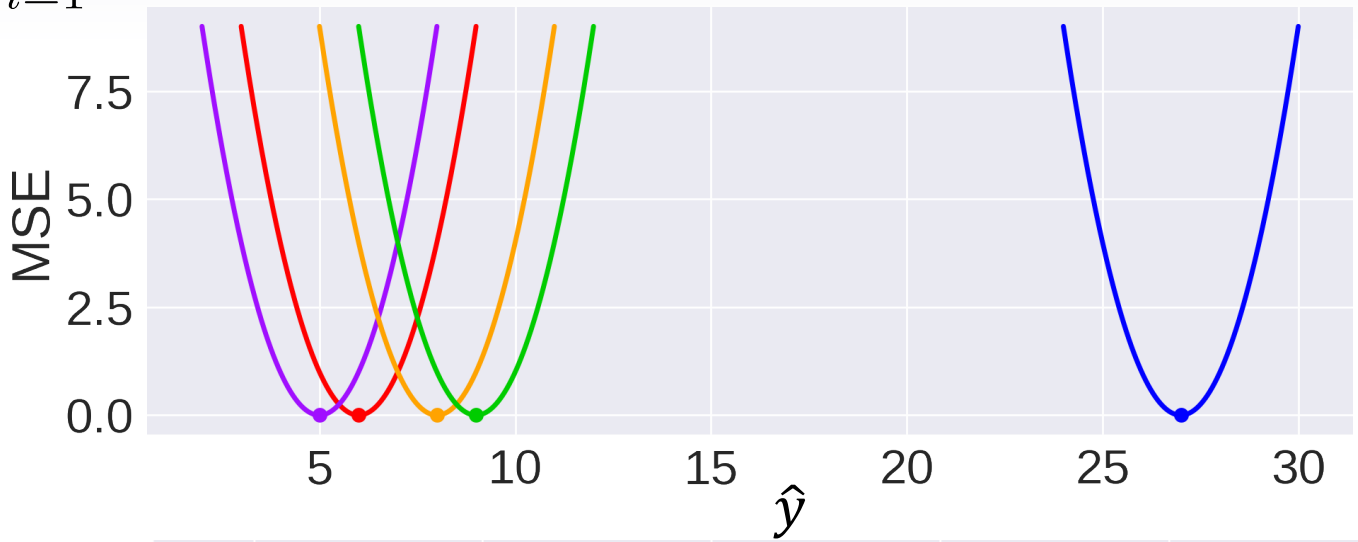
We can do it with a simple grid search over a given range by changing Alpha intuitively and recomputing an error.

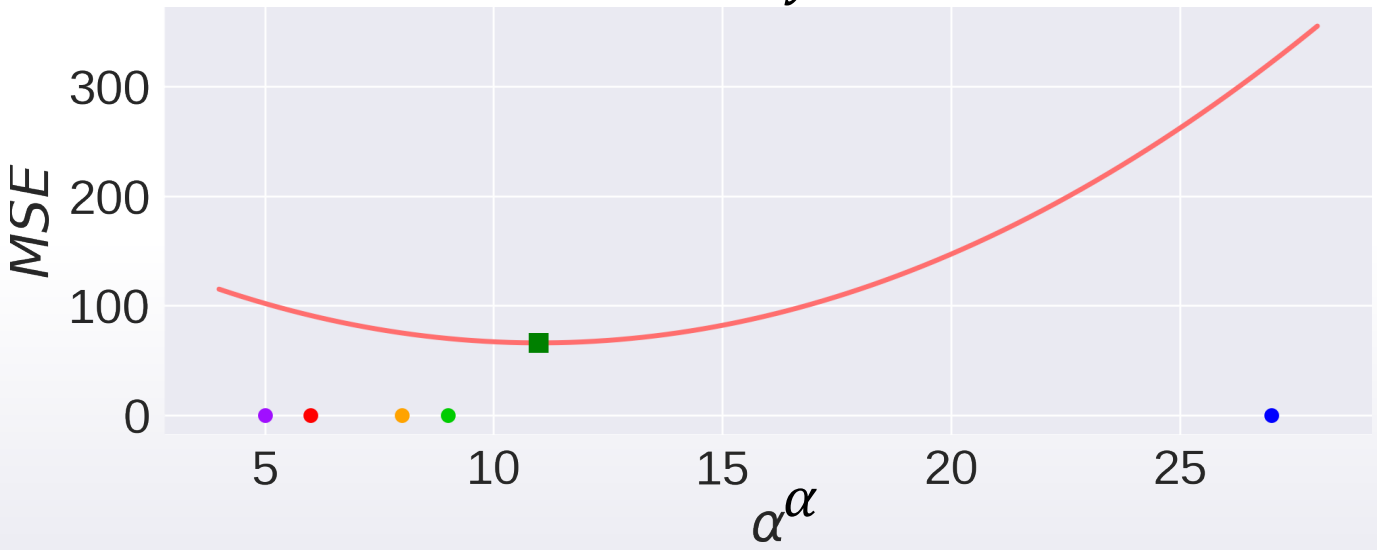
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha)^2$$

Best constant: target mean

Data:

X	Y
...	5
...	9
...	8
...	6
...	27





MSE notes: RMSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Root mean square error

- $$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}}$$

- $$\text{MSE}(a) > \text{MSE}(b) \iff \text{RMSE}(a) > \text{RMSE}(b)$$

The square root is introduced to make scale of the errors to be the same as the scale of the targets. For MSE, the error is squared, so taking a root out of it makes total error a little bit easier to comprehend because it is linear now.

MSE notes: RMSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Root mean square error

- $$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}}$$
- $$\text{MSE}(a) > \text{MSE}(b) \iff \text{RMSE}(a) > \text{RMSE}(b)$$
- $$\frac{\partial \text{RMSE}}{\partial \hat{y}_i} = \frac{1}{2\sqrt{MSE}}\frac{\partial \text{MSE}}{\partial \hat{y}_i}$$

Take a look at the gradient of RMSE with respect to i-th prediction. It is basically equal to gradient of MSE multiplied by some value. The value doesn't depend on the index I. It means that travelling along MSE gradient is equivalent to traveling along RMSE gradient but with a different flowing rate and the flowing rate depends on MSE score itself. it is kind of dynamic

MSE notes: RMSE

Actually, it's hard to realize if our model is good or not by looking at the absolute values of MSE or RMSE. It really depends on the properties of the dataset and their target vector. How much variation is there in the target vector.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

R-squared:

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

MAE: Mean Absolute Error

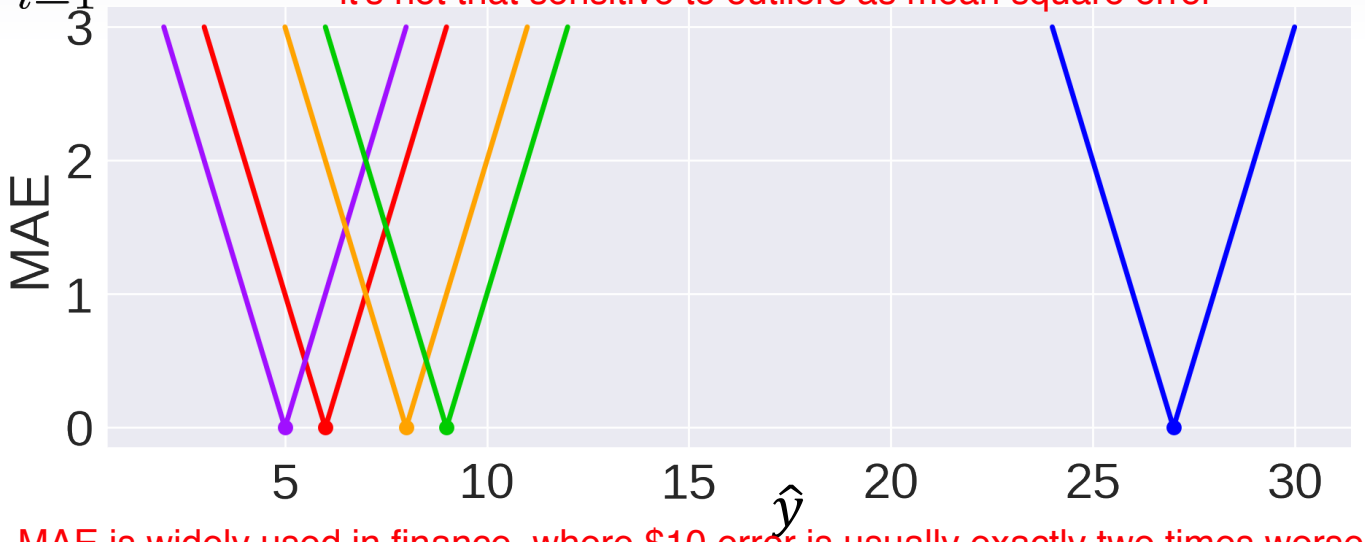
$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

What is important about this metric is that it penalizes huge errors that not as that badly as MSE does

it's not that sensitive to outliers as mean square error

Data:

X	Y
-1	5
1	9
-2	8
3	6
3	27



MAE is widely used in finance, where $10 error is usually exactly two times worse than $5 error. On the other hand, MSE metric thinks that $10 error is four times worse than $5 error.

And if you used RMSE, it would become really hard to explain to your boss how you evaluated your model.

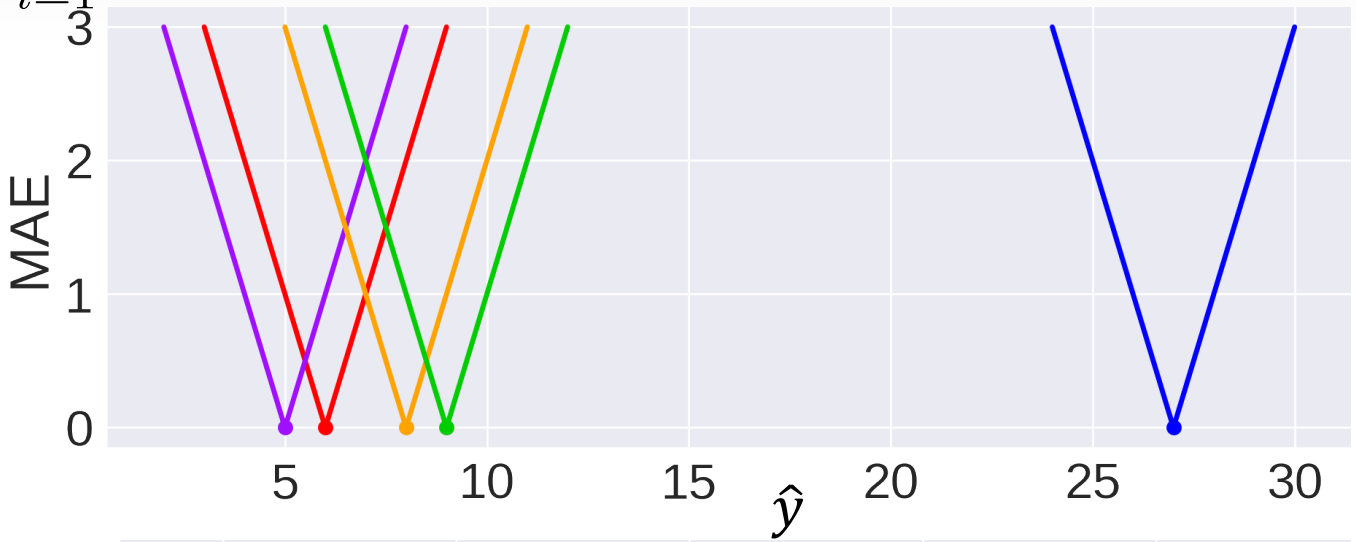
MAE: optimal constant

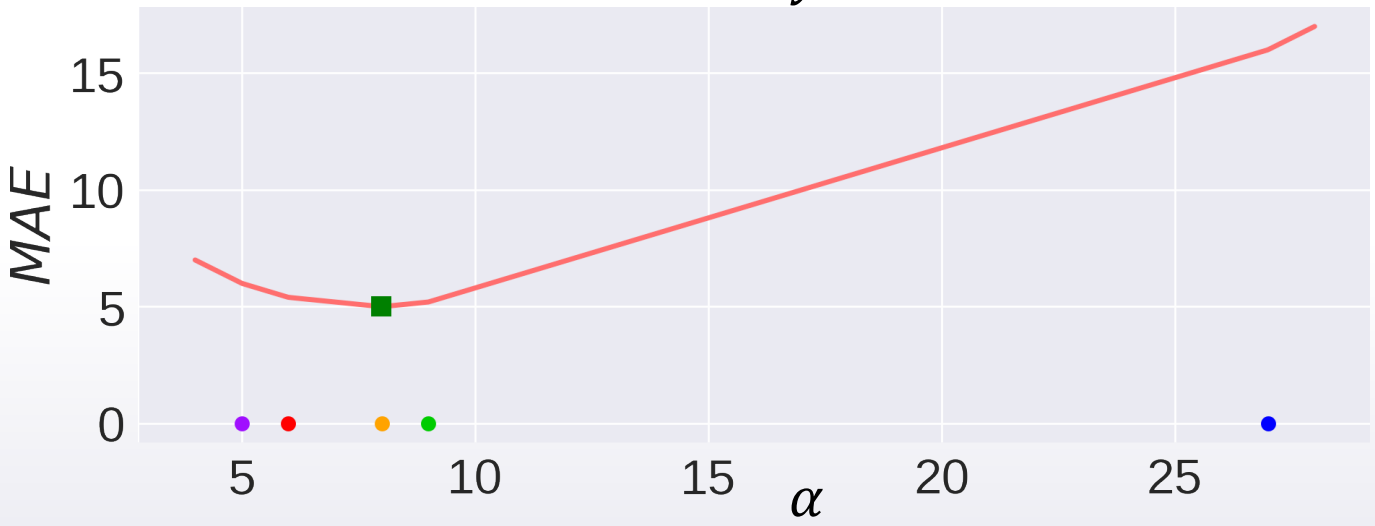
$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \alpha|$$

Best constant: target median

Data:

X	Y
-1	5
1	9
-2	8
3	6
3	27



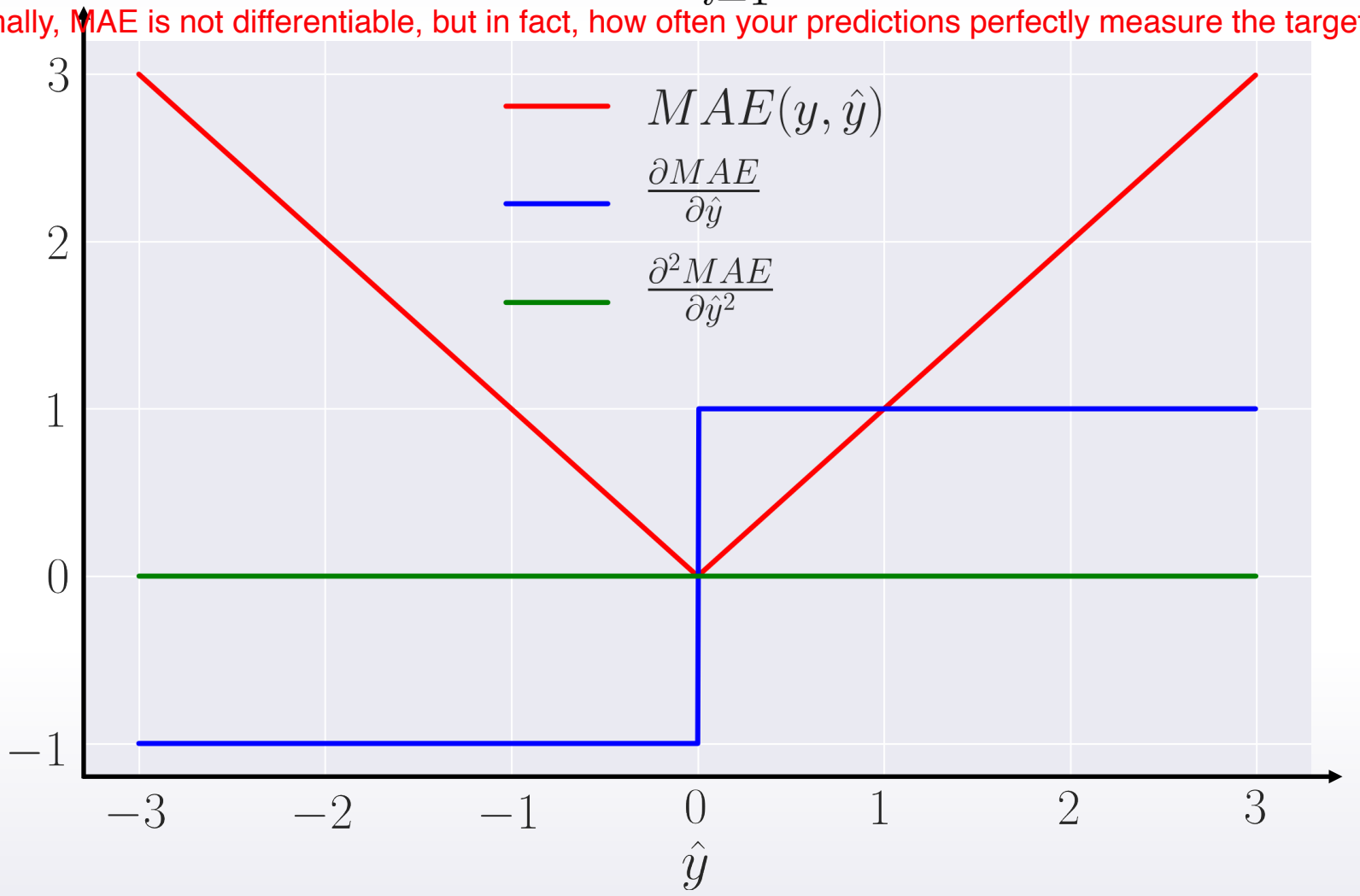


MAE: derivatives

Another important thing about MAE is its gradients with respect to the predictions. The grid end is a step function and it takes -1 when Y_hat is smaller than the target and +1 when it is larger.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

So formally, MAE is not differentiable, but in fact, how often your predictions perfectly measure the target.



MAE vs MSE

MAE is more robust than MSE. That is, it is less sensitive to outliers, but it doesnt mean it is always better to use MAE.

- **Do you have outliers in the data?**
  - Use MAE
- **Are you sure they are outliers?**
  - Use MAE
- **Or they are just unexpected values we should still care about?**
  - Use MSE

Outliers have usually mistakes, measurement errors, and so on, but at the same time, similarly looking objects can be of natural kind. So, if you think these unusual objects are normal in the sense that they're just rare, you should not use a metric which will ignore them. And it is better to use MSE. Otherwise, if you think that they are really outliers, like mistakes, you should use MAE.

Conclusion

- Discussed the following metrics:
  - **MSE, RMSE, R-squared**
    - They are the same from optimization perspective
  - **MAE**
    - Robust to outliers