Linear Regression: Loss function

Fitting an estimator/predictor/model involves solving for the Θ that minimizes the Loss function.

Recall our goal is to make the discrepancy (error) between $\mathbf{y}$ and $\hat{\mathbf{y}}$ "small".

- The discrepancy between $\mathbf{y}^{(\mathbf{i})}$ and $\hat{\mathbf{y}}^{(\mathbf{i})}$ is refered to as the *residual*, usually denoted by ϵ
$$\epsilon^{(\mathbf{i})} = \mathbf{y}^{(\mathbf{i})} - \hat{\mathbf{y}}^{(\mathbf{i})}$$

So

$$\begin{aligned} \mathbf{y} &= \hat{\mathbf{y}} + \epsilon \\ &= \mathbf{X}\Theta + \epsilon \end{aligned}$$

We define the per-example loss to be the residual *squared*

$$\mathcal{L}_{\Theta}^{(\mathbf{i})} = (\mathbf{y}^{(\mathbf{i})} - \hat{\mathbf{y}}^{(\mathbf{i})})^2$$

so that the average loss

$$\begin{aligned} \mathcal{L}_{\Theta} &= \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\Theta}^{(\mathbf{i})} \\ &= \frac{1}{m} \sum_{i=1}^{m} (\mathbf{y}^{(\mathbf{i})} - \hat{\mathbf{y}}^{(\mathbf{i})})^2 \end{aligned}$$

This expression on the right is called the *Mean Squared Error (MSE)*.

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{y}^{(\mathbf{i})} - \hat{\mathbf{y}}^{(\mathbf{i})})^2$$

- You will sometimes see *Root Mean Squared Error (RMSE)* which is the square root of the MSE

Notice that the Performance Metric and Loss Functions are identical in this case.

This will not always be true.

R^2 versus RMSE: Absolute versus relative error

One often sees the term R^2 in the context of Linear Regression.

Whereas RMSE is absolute error (in same units as $\mathbf{y}$), R^2 is a relative error (in units of percent).

The relationship is:

$$
\begin{aligned}
R^2 &= 1 - \left(\frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \bar{y}_i)^2} \right) \\
&= 1 - \left(\frac{m \cdot \mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}})}{\sum_{i=1}^m (y_i - \bar{y}_i)^2} \right) \\
&= 1 - \left(\frac{m \cdot \mathrm{RMSE}(\hat{y}, y)^2}{\sum_{i=1}^m (y_i - \bar{y}_i)^2} \right)
\end{aligned}
$$

In addition to changing the units of error, the R^2 metric has an interesting interpretation.

Consider a naive "baseline" model for prediction

- predict $\bar{\mathbf{y}}$ for every value of $\mathbf{x}$
    - where $\bar{\mathbf{y}}$ is the average (over the training examples) of the target

The loss for the naive model is

$$\mathcal{L}_\text{naive} = \text{MSE}(\mathbf{y}, \bar{\mathbf{y}})$$

Then

$$\begin{array}{lll}
R^2 & = & 1 - \left(\frac{m \cdot \text{MSE}(\mathbf{y}, \hat{\mathbf{y}})}{m \cdot \text{MSE}(\mathbf{y}, \bar{\mathbf{y}})} \right) \\
& = & 1 - \frac{\mathcal{L}}{\mathcal{L}_\text{naive}}
\end{array}$$

Thus, R^2 is the *percent reduction in loss* achieved by our model compared to the naive model that always predicts $\bar{\mathbf{y}}$.

We now know our Loss function.

The "solution" to the Linear Regression task is finding ("fitting") the Θ that minimizes average loss

$$\Theta = \operatorname*{argmin}_{\Theta} \mathcal{L}_{\Theta}$$

which is the Θ that minimizes the MSE.

In [3]:
```
print("Done")
```

```
Done
```