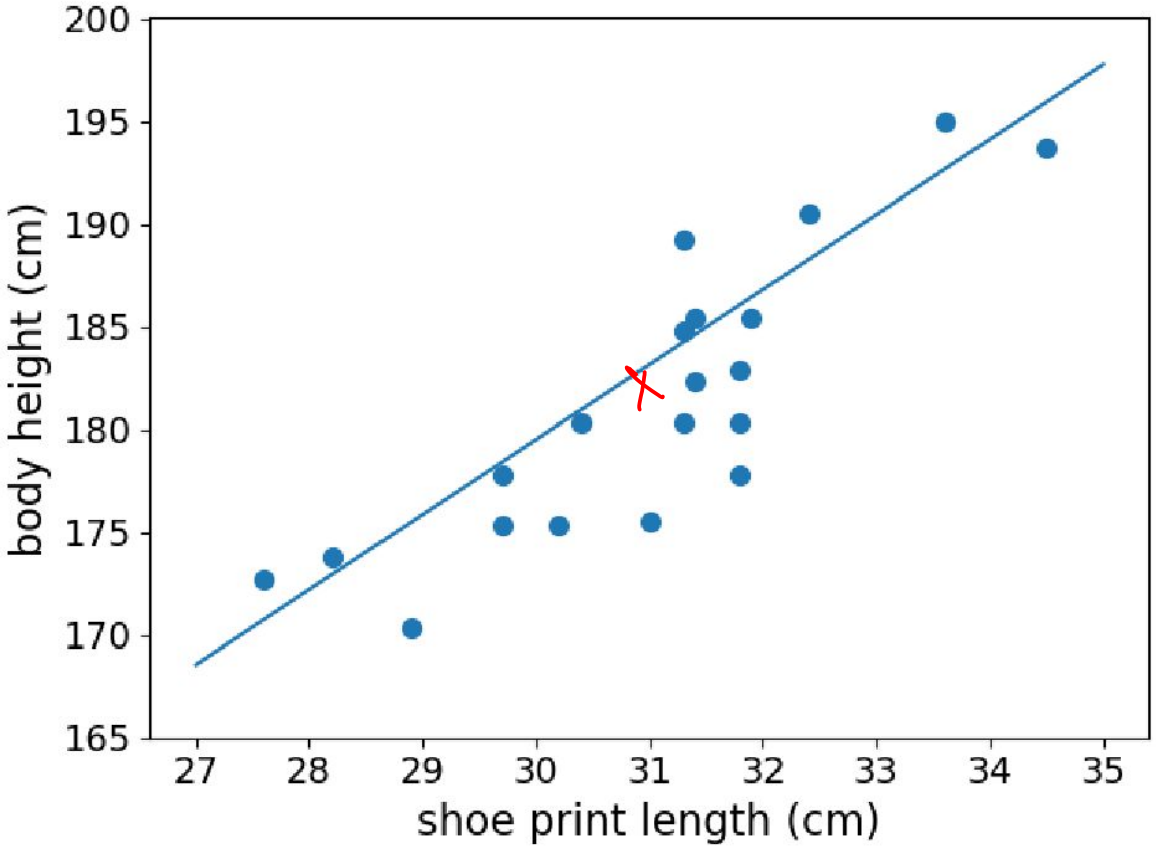
CS5228: Knowledge Discovery and Data Mining

Tutorial 7 — Classification & Regression III (Linear Models)

Question 1

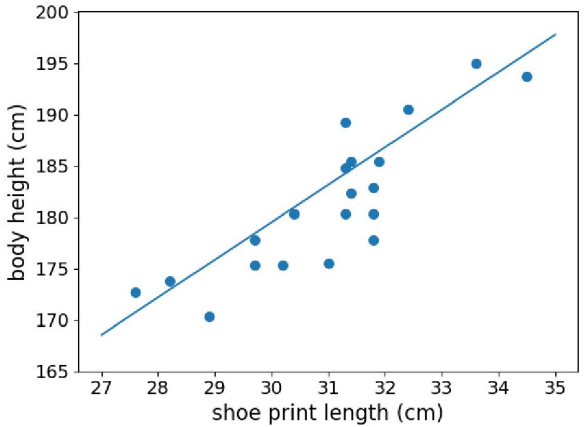

200
195
190
185
180
175
170
165
body height (cm)
27
28
29
30
31
32
33
34
35
shoe print length (cm)

Question 1

(a) The most commonly used loss function L for Linear Regression is the Means Squared Error (MSE):

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

What makes MSE a suitable loss function for Linear Regression, and what part of the formula might cause "issues"?

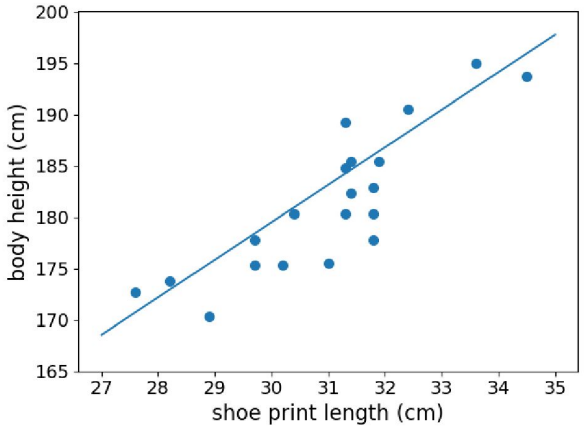


Question 1

(a) The most commonly used loss function L for Linear Regression is the Means Squared Error (MSE):

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

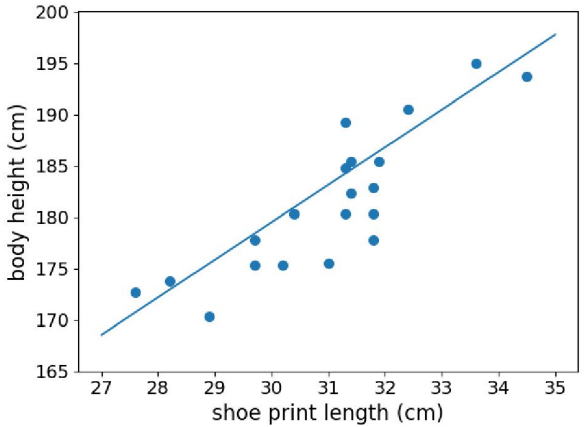
What makes MSE a suitable loss function for Linear Regression, and what part of the formula might cause "issues"?



Solution

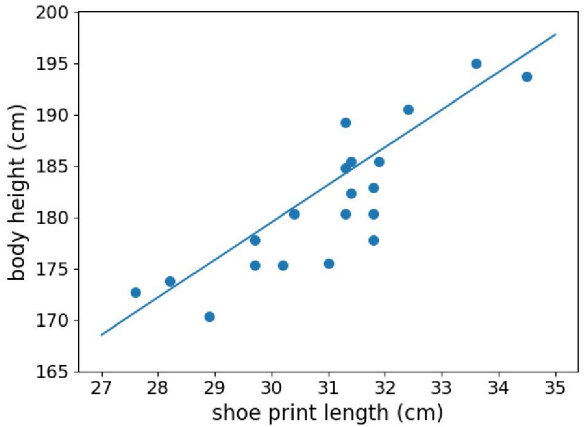
- The MSE loss intuitively captures what makes a good or bad solution – that is, on average, the line should be close to all data points.
- Squaring the residuals ensures that each individual loss is positive, no matter if the the try value y is above or below the predicted value \hat{y}
- The squaring is mathematically convenient when it comes to calculating the derivative of L.
- However, the squaring also potentially over-emphasizes larger residuals making Linear Regression quite sensitive to outliers.

Question 1



(b) After training a Linear Regression model over a dataset, how can we use the result to identify if there is indeed some linear relationship between the input features and the output values.

Question 1



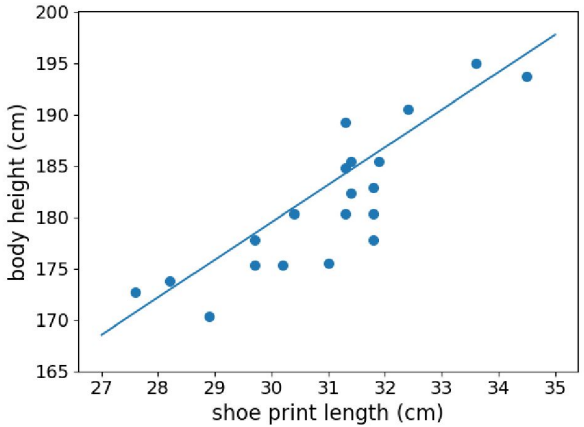
(b) After training a Linear Regression model over a dataset, how can we use the result to identify if there is indeed some linear relationship between the input features and the output values.

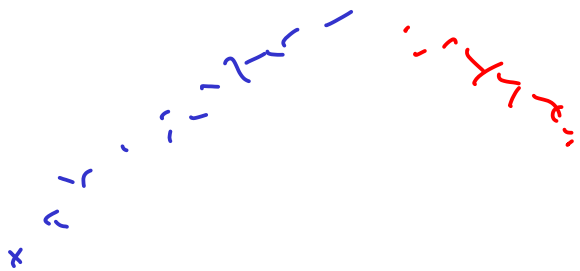
Solution

- Short answer: We can't!

Question 1

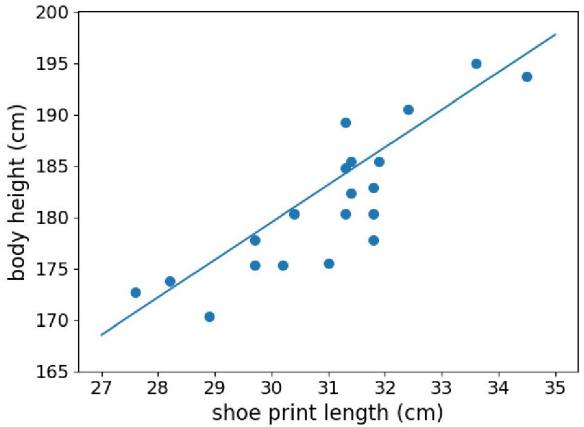
(c) Let's assume our data features strong linear relationships between the input features in output values, and we train a Linear Regression Model. Now we want to predict the values for unseen data points. In which case (i.e., for which type of data points) we might have to very cautious when interpreting the predicted values?





Question 1

(c) Let's assume our data features strong linear relationships between the input features in output values, and we train a Linear Regression Model. Now we want to predict the values for unseen data points. In which case (i.e., for which type of data points) we might have to very cautious when interpreting the predicted values?



Solution

- As long as a data point is in the range of the training data (interpolation) we are generally on the safe side.
- For any data points outside the range of the data (extrapolation), we may not guarantee if the predicted values are meaningful:
  - The data points might not be meaningful. For example, a negative shoe print size does not make sense, but we can give it to the Linear Model and get some result.
  - Outside the range of our training data, the assumption of a linear relationship might no longer hold.
- Note that this result might vary when considering more than this single feature

Questions 1

(d) In the lecture, we only skimmed over the calculation of the derivative of loss function L. Given the matrix notation of L

$$L = \frac{1}{n}\|X\theta - y\|^2$$

find $\frac{\partial L}{\partial \theta}$!

Question 2

This ultimately brought us to the *Cross-Entropy Loss*, the loss function for Logistic Regression (binary classification):

$$
\begin{aligned}
L &= -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \frac{1}{1+e^{\theta^T x_i}} + (1-y_i)\log\left(1-\frac{1}{1+e^{\theta^T x_i}}\right)\right]
\end{aligned}
$$

Well, find $\frac{\partial L}{\partial \theta}$! :)