Linear regression

Evaluation and interpretation

Herman Kamper

http://www.kamperh.com/

Regression evaluation metrics

- Squared loss:

$$J = \sum_{n=1}^{N} \left(y^{(n)} - f(\mathbf{x}^{(n)}; \mathbf{w}) \right)^2$$

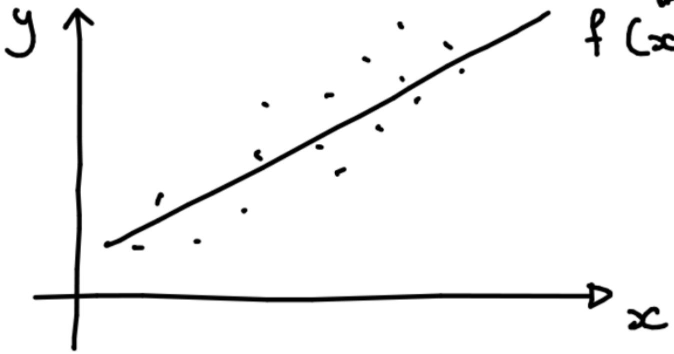
- Mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{N} \sum_{n=1}^{N} \left(y^{(n)} - f(\mathbf{x}^{(n)}; \mathbf{w}) \right)^2$$

- Root-mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y^{(n)} - f(\mathbf{x}^{(n)}; \mathbf{w}))^2}$$

Interpretation of linear regression

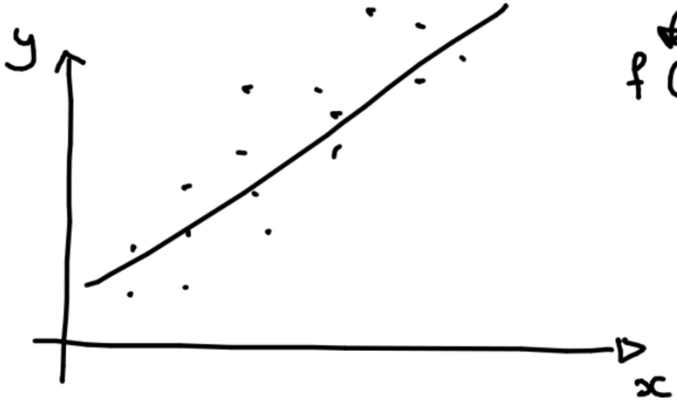


Ice cream sales

$$f(x) = w_0 + w_1 x$$

Shark attacks

What does w_1 tell us in this case?



Ice cream sales

$$f(x) = w_0 + w_1 x$$

Temperature on day

Could potentially solve this problem by using multiple regression:

Ice cream sales

$$f(x) = w_0 + w_1 x_1 + w_2 x_2$$

Temperature

Shark attacks

This might not solve all your problems, and you will still not know whether temperature causes ice cream sales or the other way around.

Interpretation of linear regression

Lasso

Another example: I have this cool (and actually useful) idea of using L_1 regularization to pick the 5 most meaningful features in a problem where $\underline{x} \in \mathbb{R}^{100}$. We vary λ until we have only 5 non-zero w's.
Are these 5 values the things that most "cause" the output y?
No, they are the best 5 values for predicting y, given a number of assumptions (e.g. we are using a linear model).

Takeaway: Be careful about making statements based on linear regression coefficients. Linear models can be very useful since they can be more interpretible. But they won't always give a complete picture — they are often most useful in conjunction with some other model/hypothesis of the real world (domain knowledge).
Sometimes the most you will be able to say is: "These features are the most important/useful for predicting the output given that we use a linear model." (And maybe that is already useful enough!)

Further reading:

The Book of Why - Mackenzie & Pearl

https://www.tylervigen.com/spurious-correlations