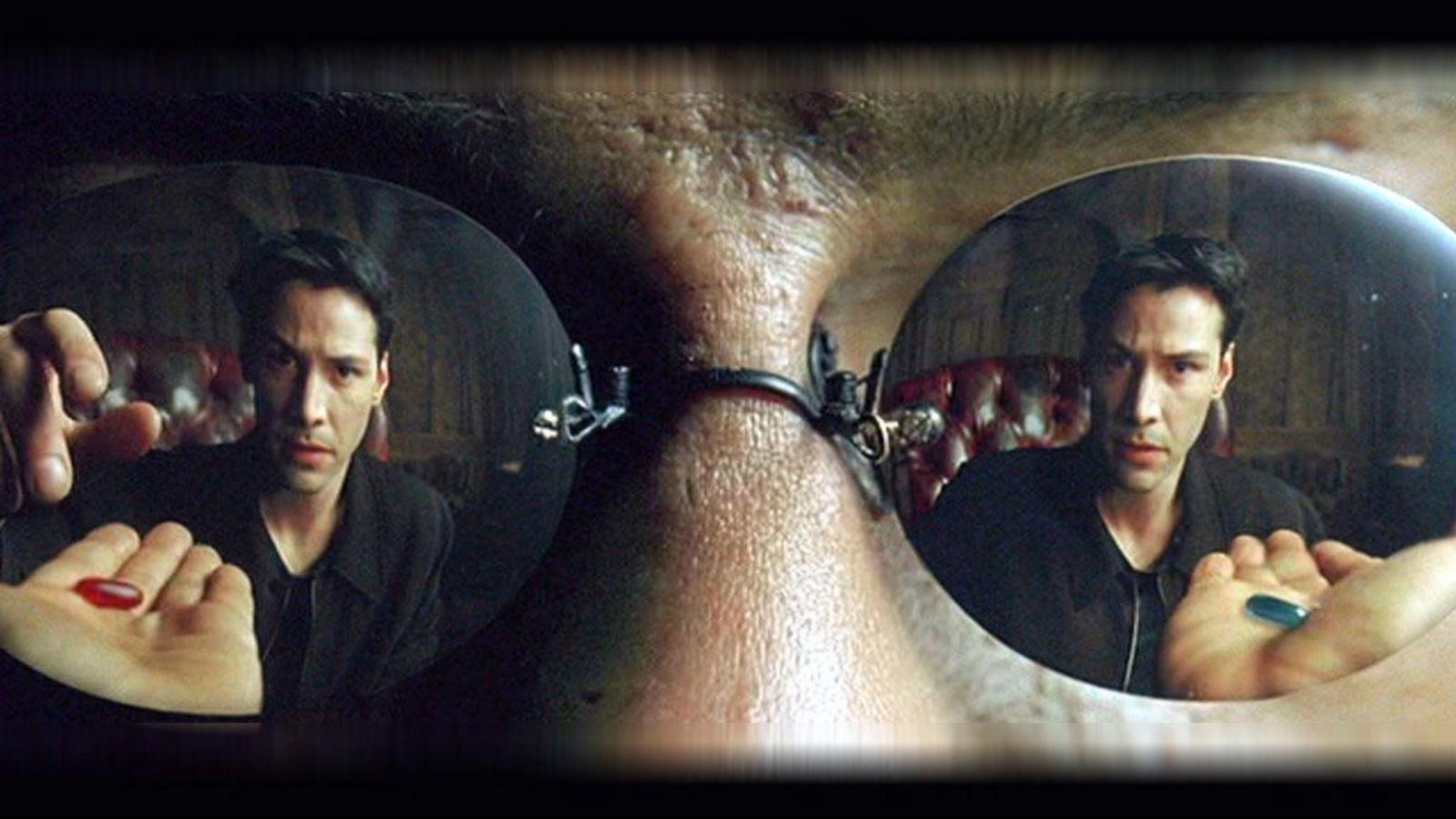
Robust Regression

Week 08 - Day 05

R2 is the best score, right?



The Matrix has you...

Follow the white rabbit...

Knock knock, Neo.

Classification

precision, recall, auc, f1, accuracy,
cohen’s kappa, etc.

Regression

R2, adj R2, mean absolute error, median absolute error, MSE, AIC, BIC, etc.

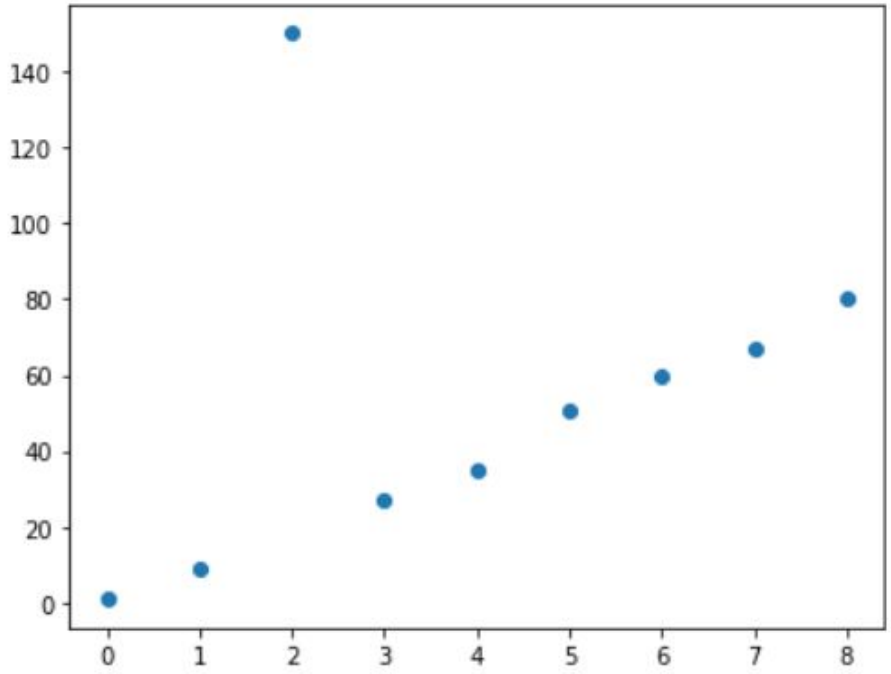
What's wrong with R2?

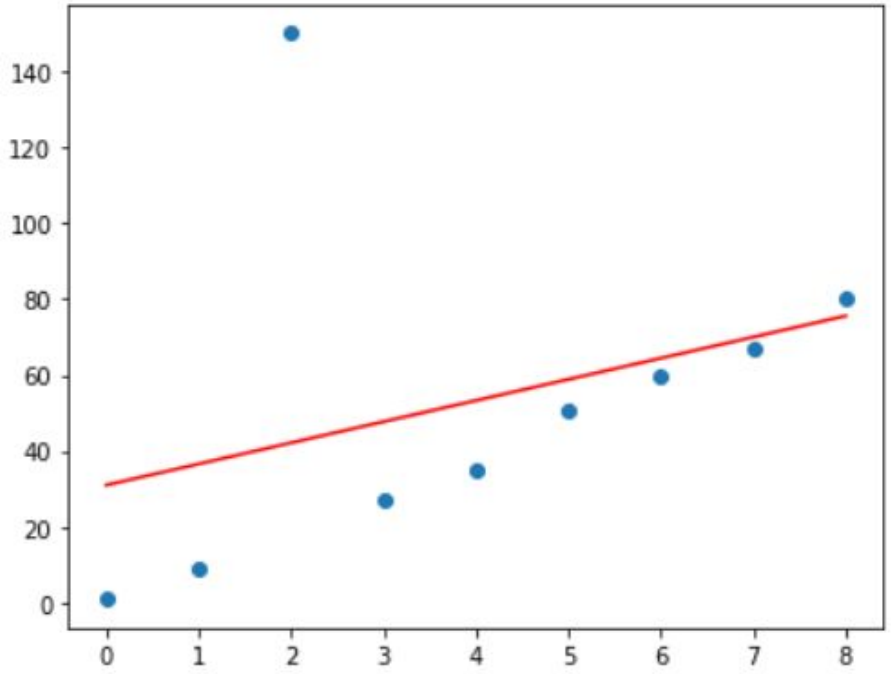
R2 usually increases

when we add other useless predictors

(solution adjusted R2, cross-val)

R2 is sensitive to outliers





How to deal with outliers

Naive manual approach: **Theil-Sen**

1. Pick a pair of points at random.

1. Pick a pair of points at random.
2. Draw a line through them, and remember the gradient of the line.

1. Pick a pair of points at random.
2. Draw a line through them, and remember the gradient of the line.
3. Repeat steps 1 & 2 some number of times (e.g. 25 times).

1. Pick a pair of points at random.
2. Draw a line through them, and remember the gradient of the line.
3. Repeat steps 1 & 2 some number of times (e.g. 25 times).
4. Sort the lines by gradient.

1. Pick a pair of points at random.
2. Draw a line through them, and remember the gradient of the line.
3. Repeat steps 1 & 2 some number of times (e.g. 25 times).
4. Sort the lines by gradient.
5. Choose the line with the median gradient

Better approach: **huber loss**

The error is:

- Squared, if we’re making a small error
- Absolute, if we’re making a large error

We have to define when

an errors is small or large

sklearn.linear_model.HuberRegressor

What metric should I use

Previous models won’t give the best R2!

Simple good metric

to deal with outliers:

Median absolute error

Other possible metrics:

Adjusted R2, BIC, AIC

You can define your own metric!

Example:

Selling lemonades

temperature	rain	# of lemonades
32	yes	47
31	no	35
28	no	33
...	..	..

Production cost: 0.1$

Margin: 0.9$

If I over predict,

how much do I lose?

If I over predict,

how much do I lose?

0.1$ for each extra lemonade!

If I under predict,

how much do I lose?

If I under predict,

how much do I lose?

0.9$ for each lemonade

I haven’t produced!

Error:

- 0.9 * (y_real - y_pred), if y_real>y_pred
- 0.1 * (y_pred - y_real), if y_real<y_pred