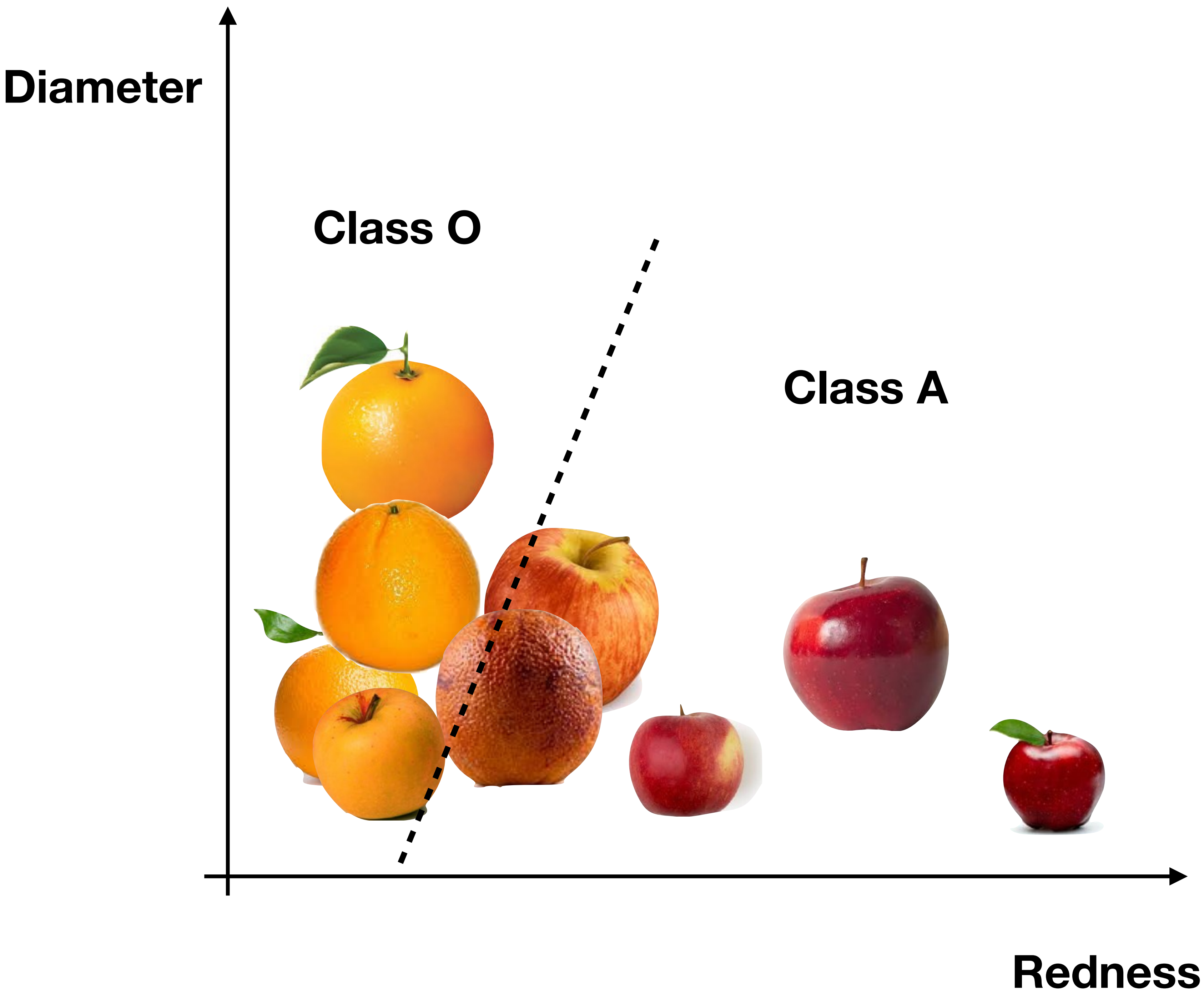
Logistic Regression

What is logistic regression?

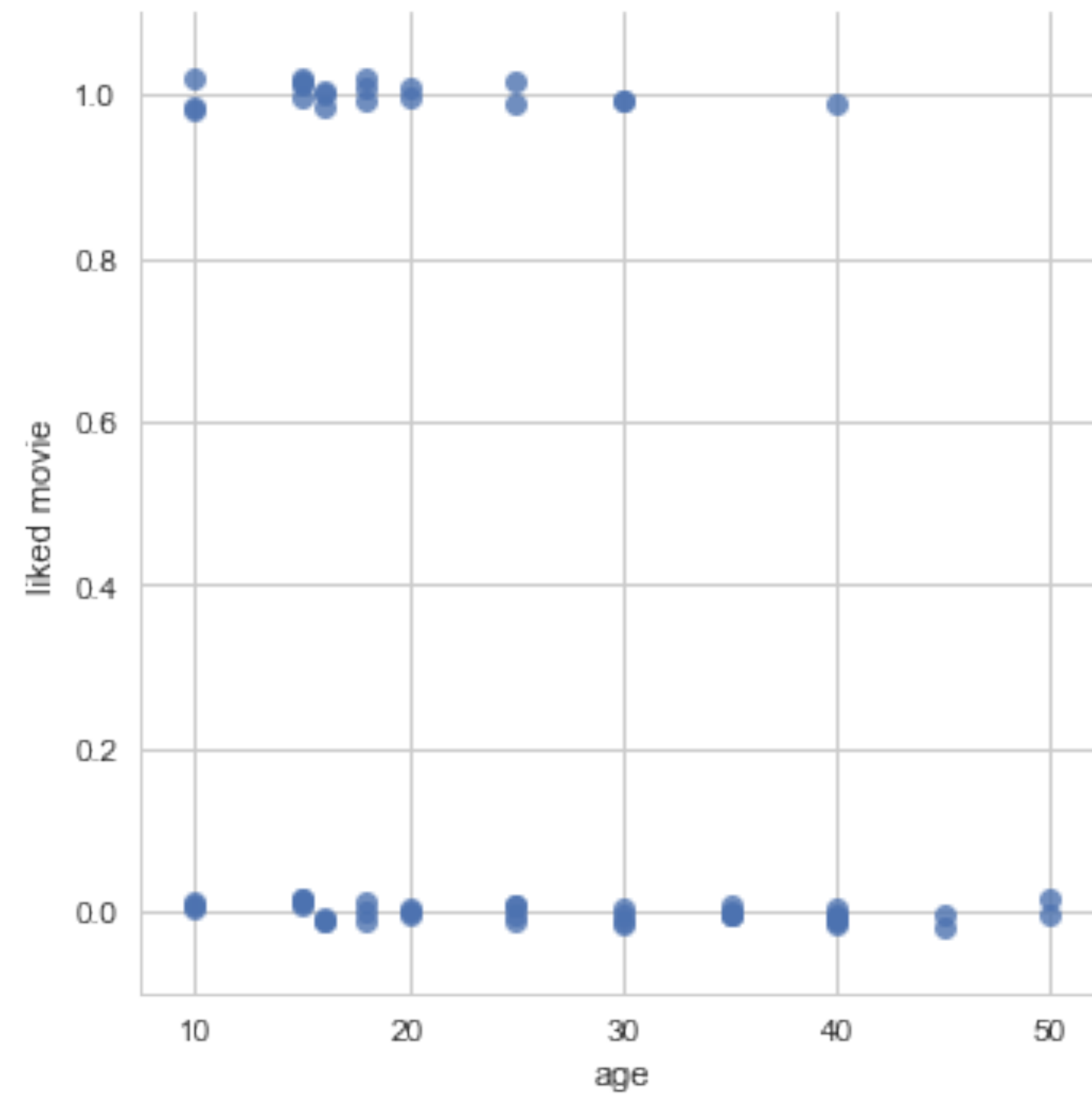
- A model for classification
- Based on a linear model of the features (or covariates)
- Features are used to predict the (log) odds of belonging to (baseline) class 1

Classification



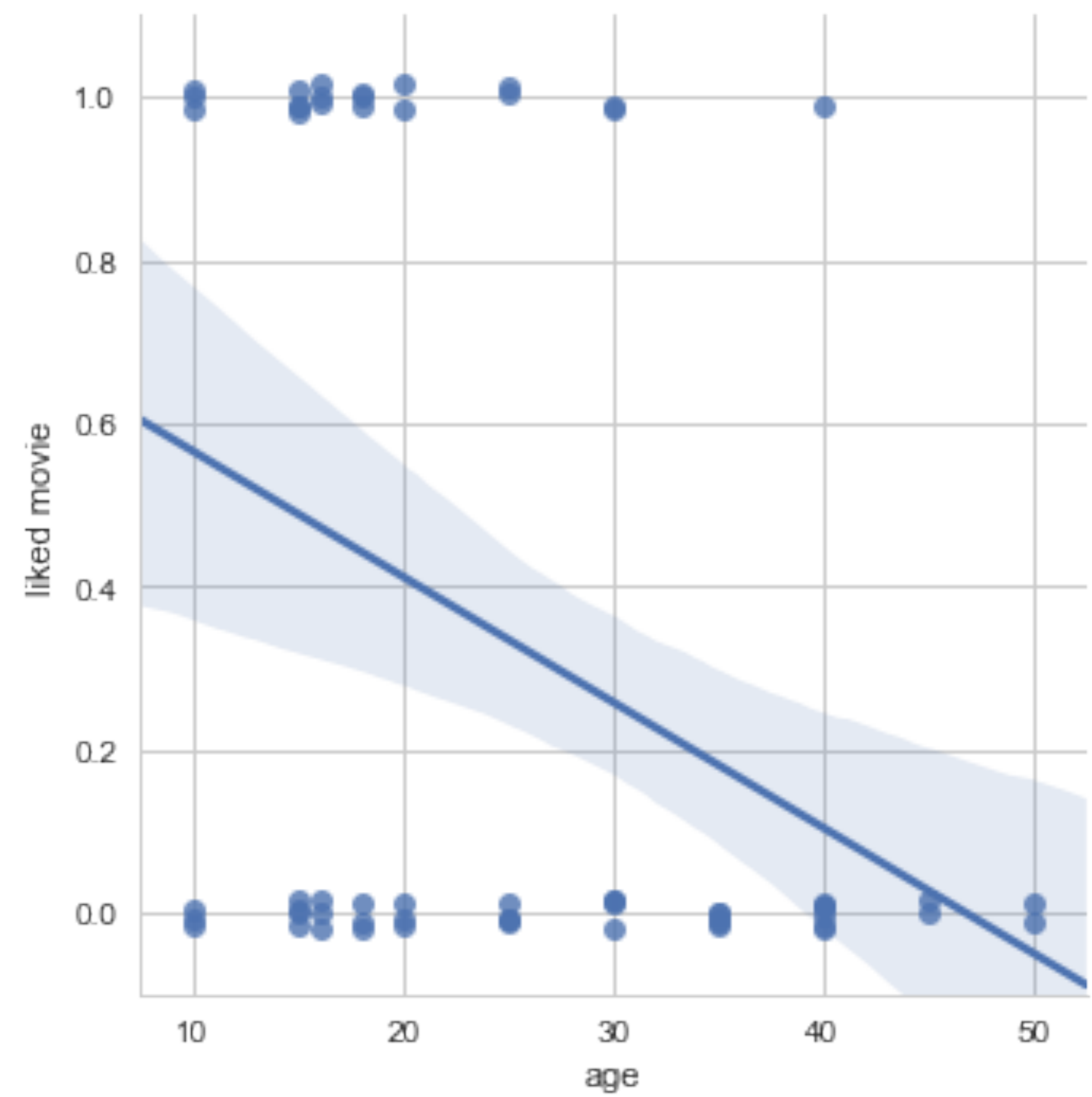
A simple explanation

- A fake dataset:
    - Given the age, predict if someone will like movie X



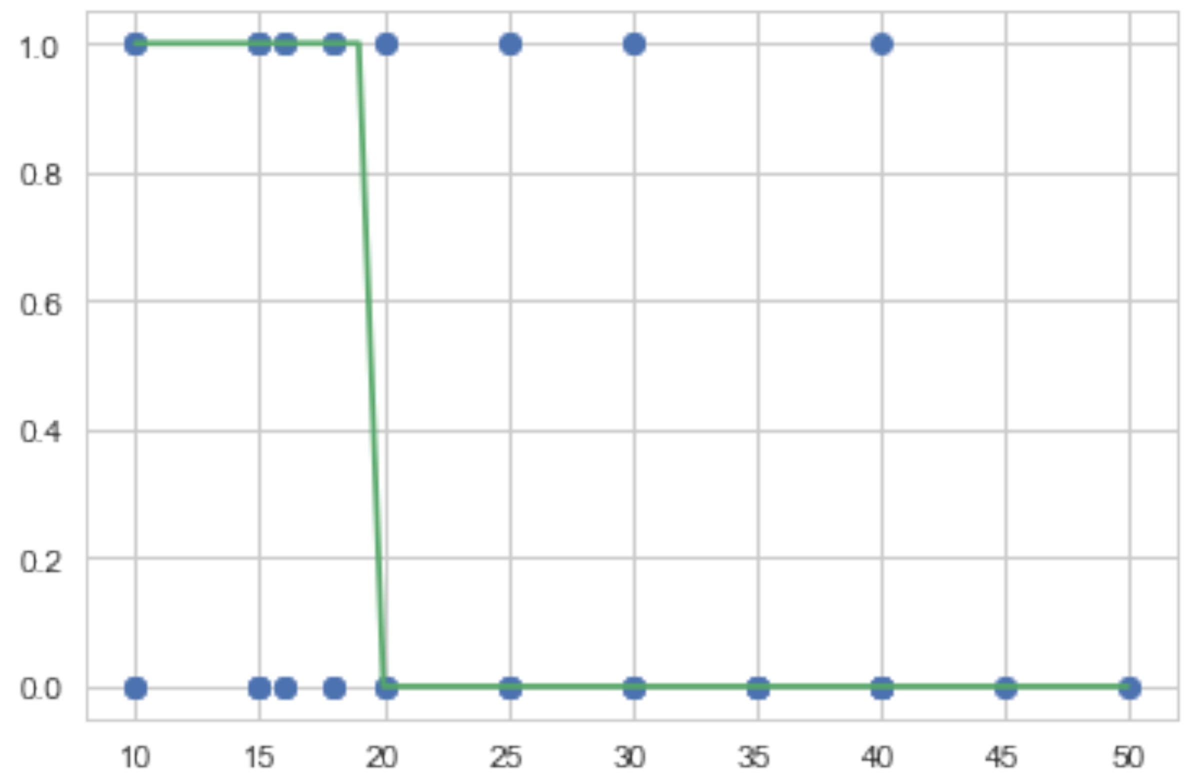
A simple explanation

- A linear regression won't do



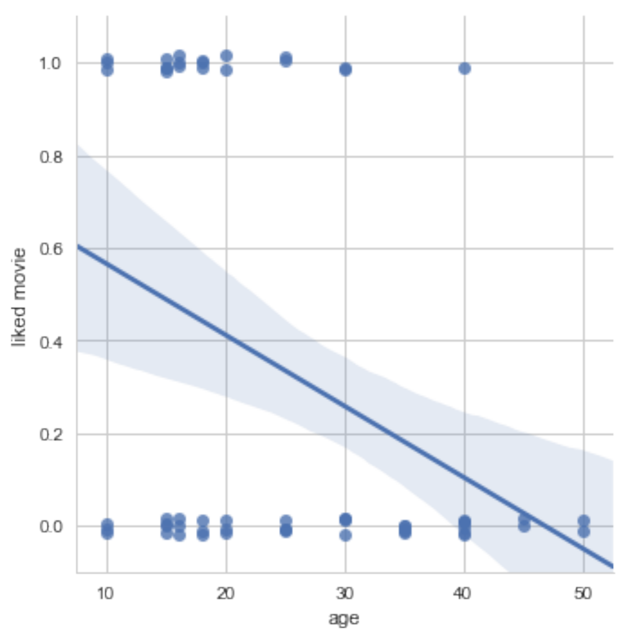
A simple explanation

- We look for something like a threshold

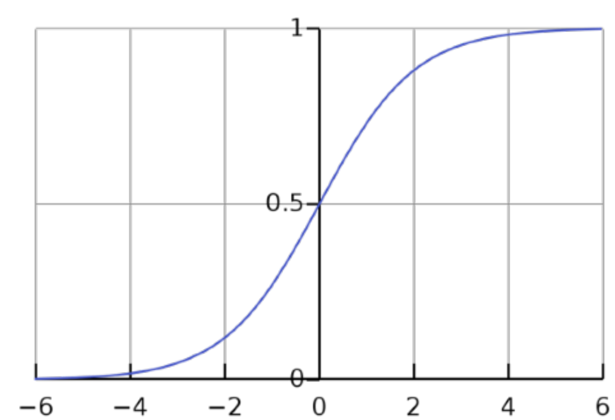


A simple explanation

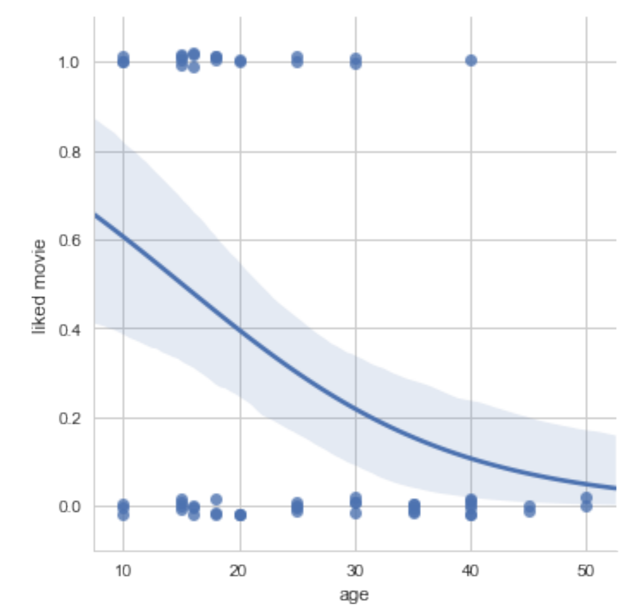
- We can combine linear regression and a nonlinear activation function



+



=



A mathematical explanation

$p(C_1|x)$

A mathematical explanation

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)}$$

Bayes

A mathematical explanation

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)}$$

$$o(x) = \log \frac{p(x|C_1)p(C_1)}{p(x|C_0)p(C_0)}$$

Definition log odds

A mathematical explanation

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)}$$

$$o(x) = \log \frac{p(x|C_1)p(C_1)}{p(x|C_0)p(C_0)}$$

$$= \log \frac{p(x|C_1)p(C_1)}{1 - p(x|C_1)p(C_1)}$$

Probabilities sum to 1

A mathematical explanation

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_0)p(C_0)}$$

$$o(x) = \log \frac{p(x|C_1)p(C_1)}{p(x|C_0)p(C_0)}$$

$$= \log \frac{p(x|C_1)p(C_1)}{1 - p(x|C_1)p(C_1)}$$

$$p(C_1|x) = \frac{1}{1 + exp(-o)}$$

Substitute o into p(C1|x)

A mathematical explanation

$$o(x) = \log \frac{p(x|C_1)p(C_1)}{p(x|C_0)p(C_0)}$$

$$= \log \frac{p(x|C_1)p(C_1)}{1 - p(x|C_1)p(C_1)}$$

We will estimate the log odds with a linear model

$$p(C_1|x) = \frac{1}{1 + exp(-o)}$$

And use the sigmoid function to compute the probability for C1

A mathematical explanation

$$= \log \frac{p(x|C_1)p(C_1)}{1 - p(x|C_1)p(C_1)}$$

$$p(C_1|x) = \frac{1}{1 + exp(-o)}$$

Data → linear model of log odds o → Class probability = sigmoid of odds

x

o = w_1 x + w_0

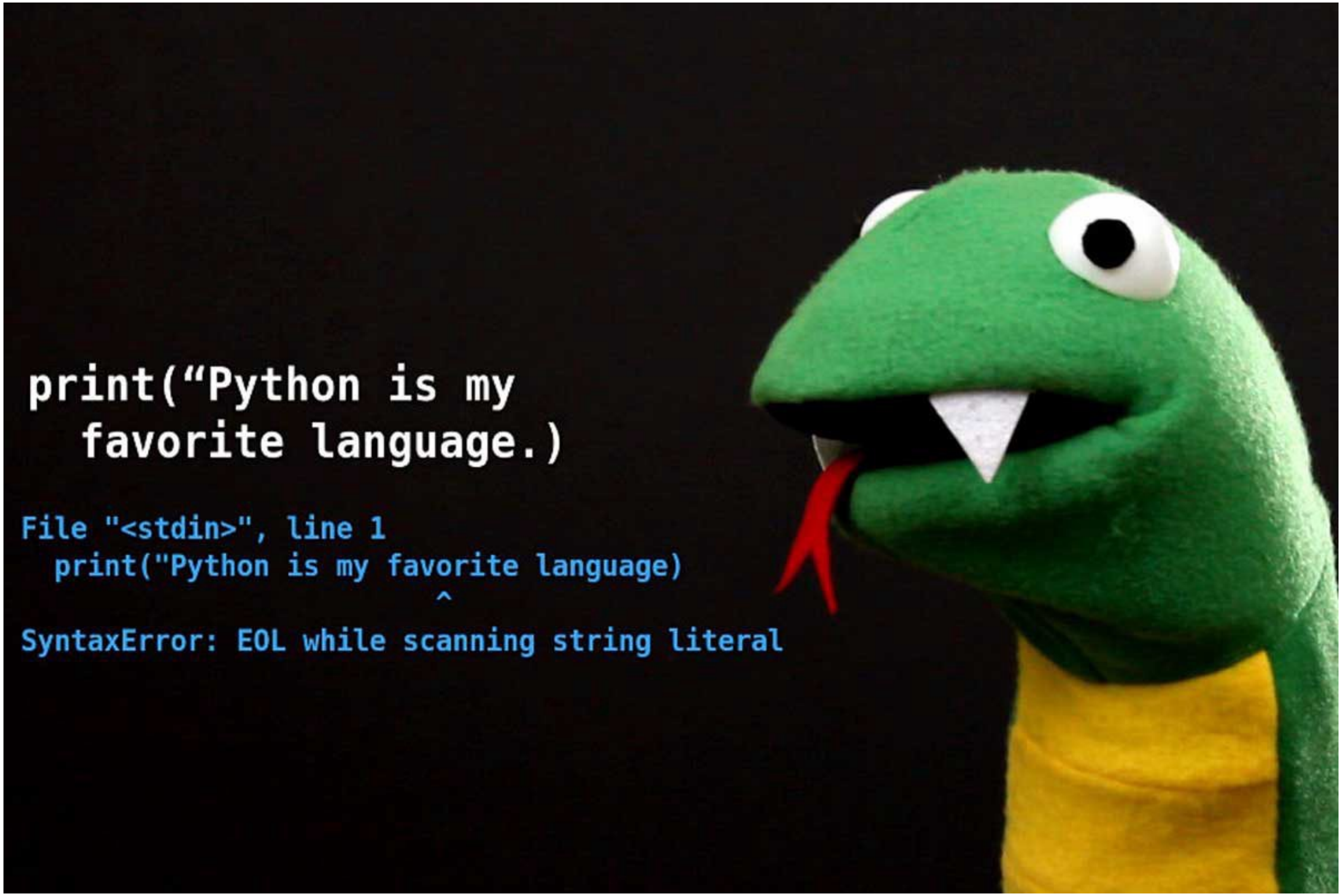
P(C_1|x)=Sigmoid(o)

Assumptions

- Linear dependence of features to log odds
- Measurements are independent
- Features are not (too) correlated (no or few multi-collinearity)
- Typically a large sample size
  - at minimum of 10 cases with the least frequent outcome for each independent variable
  - 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10*5 / .10)
  - from http://www.statisticssolutions.com/assumptions-of-logistic-regression/

Model fitting

- In principle:
  - Due to the sigmoid, the model is not linear,
    i.e. no closed form solution anymore
  - But the objective function is convex
    - i.e. easy to optimise numerically



https://medium.com/level-up-web/best-python-books-in-2017-b064dfac287

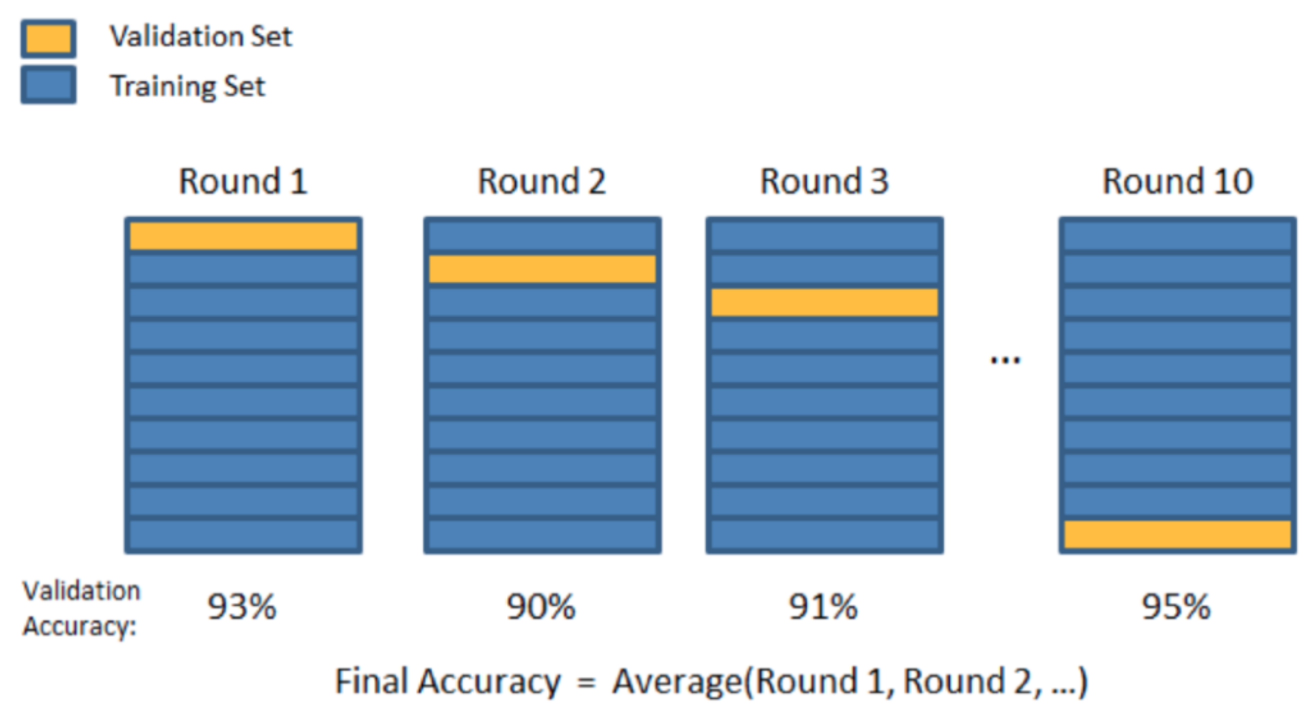
- Don’t worry, we will use Python (scikit-learn)

Model fitting

- But we still need to select ‘good’ features

Model validation

- Training and test set
- **Cross validation**
- Accuracy
- Precision, recall, f-score
- ROC
- Confusion matrix



https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Model validation

- Training and test set
- Cross validation
- **Accuracy** **TP + TN / (TP + FP + TN + FN)**
- Precision, recall, f-score
- ROC
- Confusion matrix

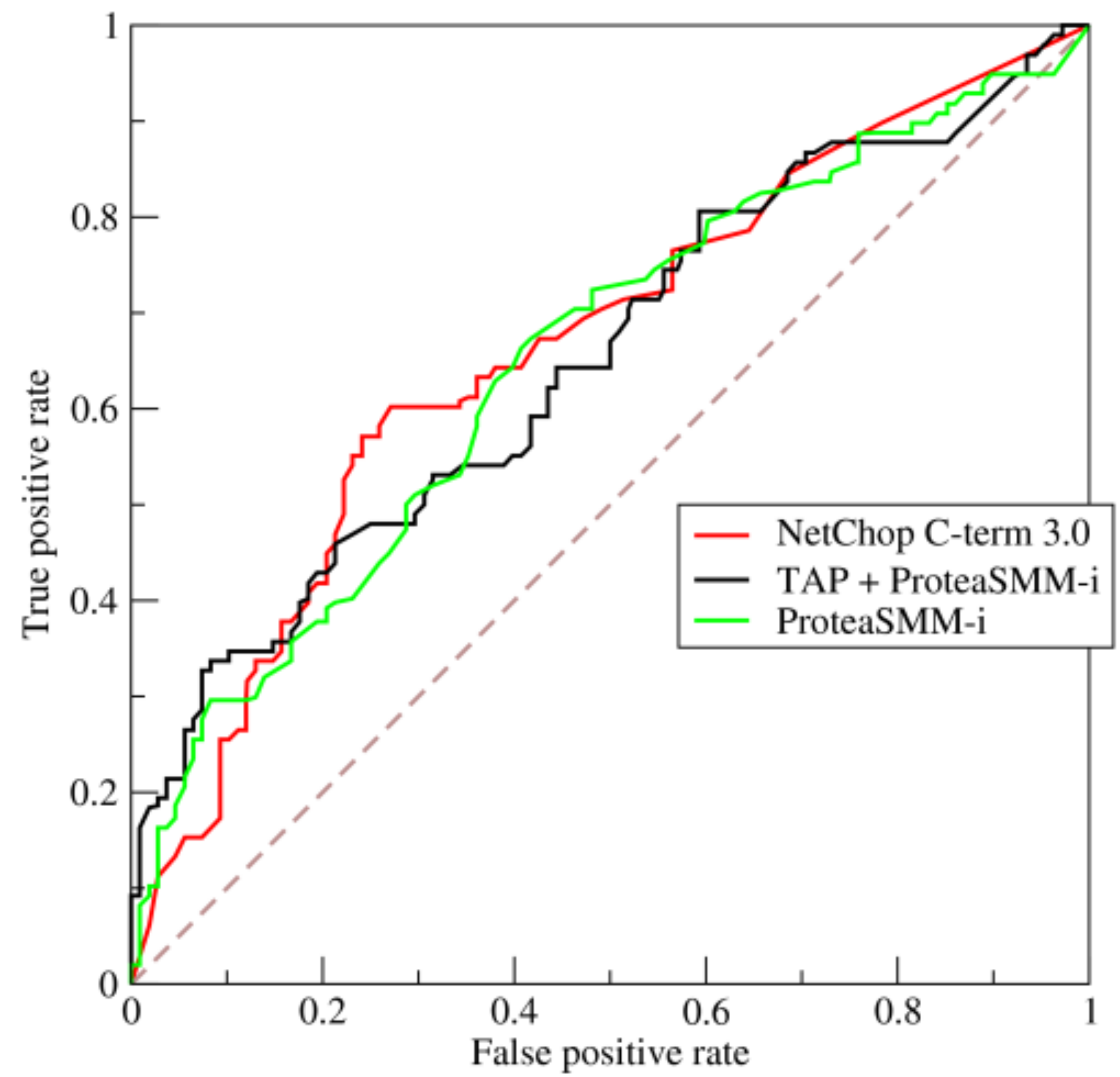
Model validation

- Training and test set
- Cross validation
- Accuracy
- **Precision, recall, f-score**
- ROC
- Confusion matrix

Precision (PPV) = TP/(TP + FP)
Recall (TPR) = TP/(TP + FN)

Model validation

- Training and test set
- Cross validation
- Accuracy
- Precision, recall, f-score
- **ROC**
- Confusion matrix



Model validation

- Training and test set
- Cross validation
- Accuracy
- Precision, recall, f-score
- ROC
- **Confusion matrix**

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

Model interpretation

- You can read how a unit change in your feature changes the log odds.
  - (keeping the other features fixed)
- This can be transformed back into probabilities

Summary

+

- Easy to model and compute
- Interpretable model

-

- Requires feature selection
- Linear decision boundary between the classes