The Augmented Social Scientist

Using Sequential Transfer Learning to Annotated Millions of Texts with Human-Level Accuracy (Do, Ollion & Shen, *SMR*, 2022)

II. Some Machine Learning Notions

More at www.css.cnrs.fr (click “Tutorial”)

Étienne Ollion (CNRS/ École polytechnique)
Rubing Shen (Sciences Po / École polytechnique)

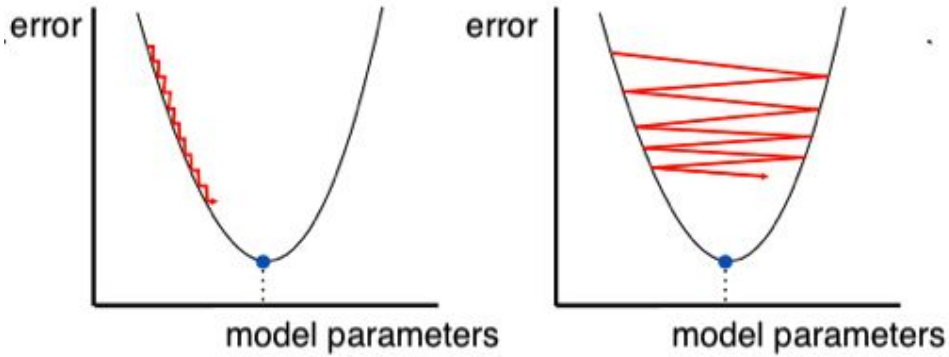
Some Machine Learning Notions

- Training parameters
  - Learning rate
  - Number of epochs

- Evaluation metrics
  - Precision
  - Recall
  - F1-Score

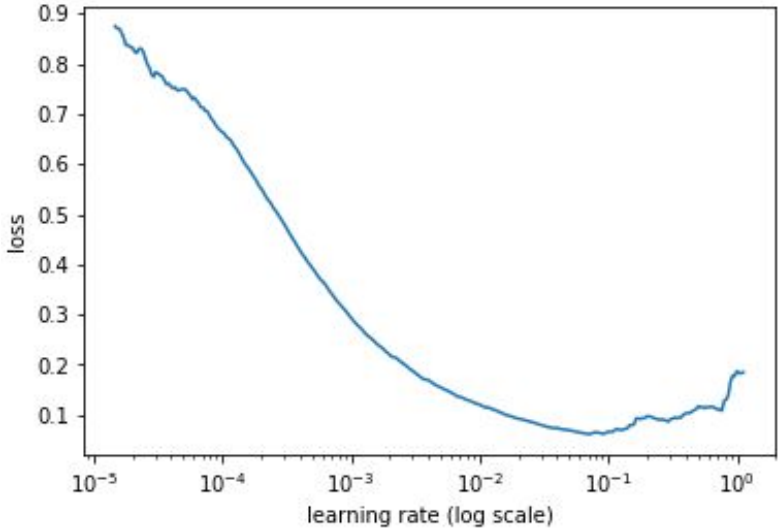
Training Parameters: Learning Rate

Learning rate: “step” of gradient descent



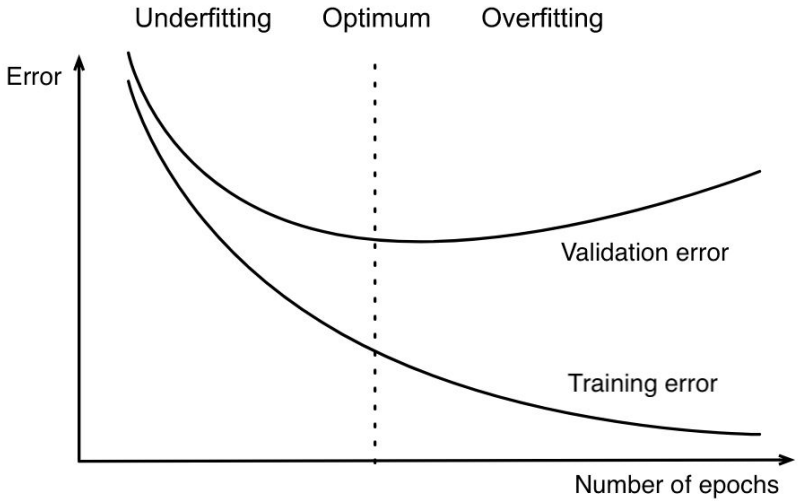
Learning rate too small

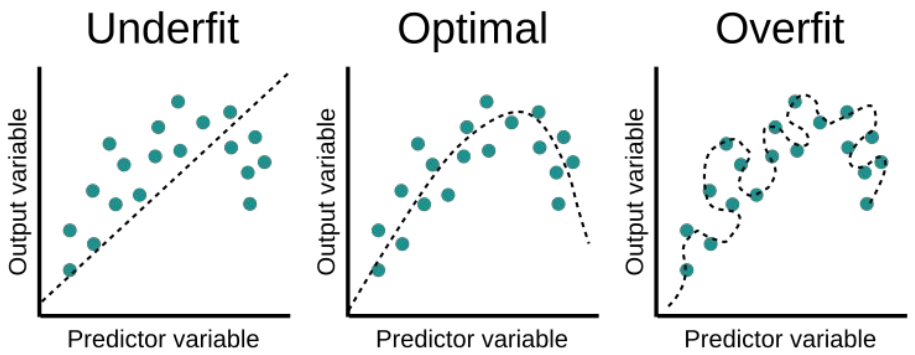
Learning rate too big



Training Parameters: Number of Epochs

Number of epochs: number of complete passes through the training dataset



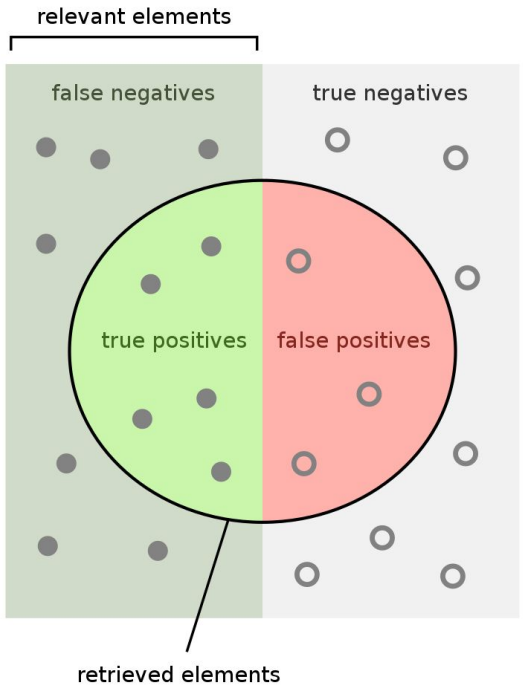


Evaluation Metrics

- Naive metrics
  - Accuracy = % correct answers

- Problem: unbalanced dataset
  - If there are 3% of positive rate, a model that always predicts ‘negative’ has an *accuracy* of 97%

Evaluation Metrics: Precision, Recall, F1-Score



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{retrieved elements}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

F1 = (harmonic) mean of precision and recall