

# Python Data Products

Course 3: Making Meaningful Predictions from Data

Lecture: Guidelines on the implementation of  
predictive pipelines

# Learning objectives

In this lecture we will...

- Suggest practical guidelines for model selection
- Show how our “theorems” about model selection can be applied in practice
- Demonstrate other cases where the validation set can be used, besides selection regularization parameters

# Choosing among several different models

**1.** As well as selecting model hyperparameters (like  $\lambda$ ) the validation set can also be used to select among model alternatives. E.g.:

- Should I use an SVM or a logistic regressor?
  - How deep should my decision tree be?
- How many layers should my neural network have?

# Choosing among several different models

**2.** When using iterative models (like gradient descent/ascent), it is not necessary (or desirable) to train the model to convergence

Rather, we should periodically compute the validation error, and **stop once the validation error is no longer improving**

# Choosing among several different models

**3.** The validation set can also be used to guide feature engineering choices. E.g.:

- How many words should we include in our dictionary?  
In our example we used 1000 but we could make a better choice using our validation set
- Should we remove punctuation, capitalization, etc.?

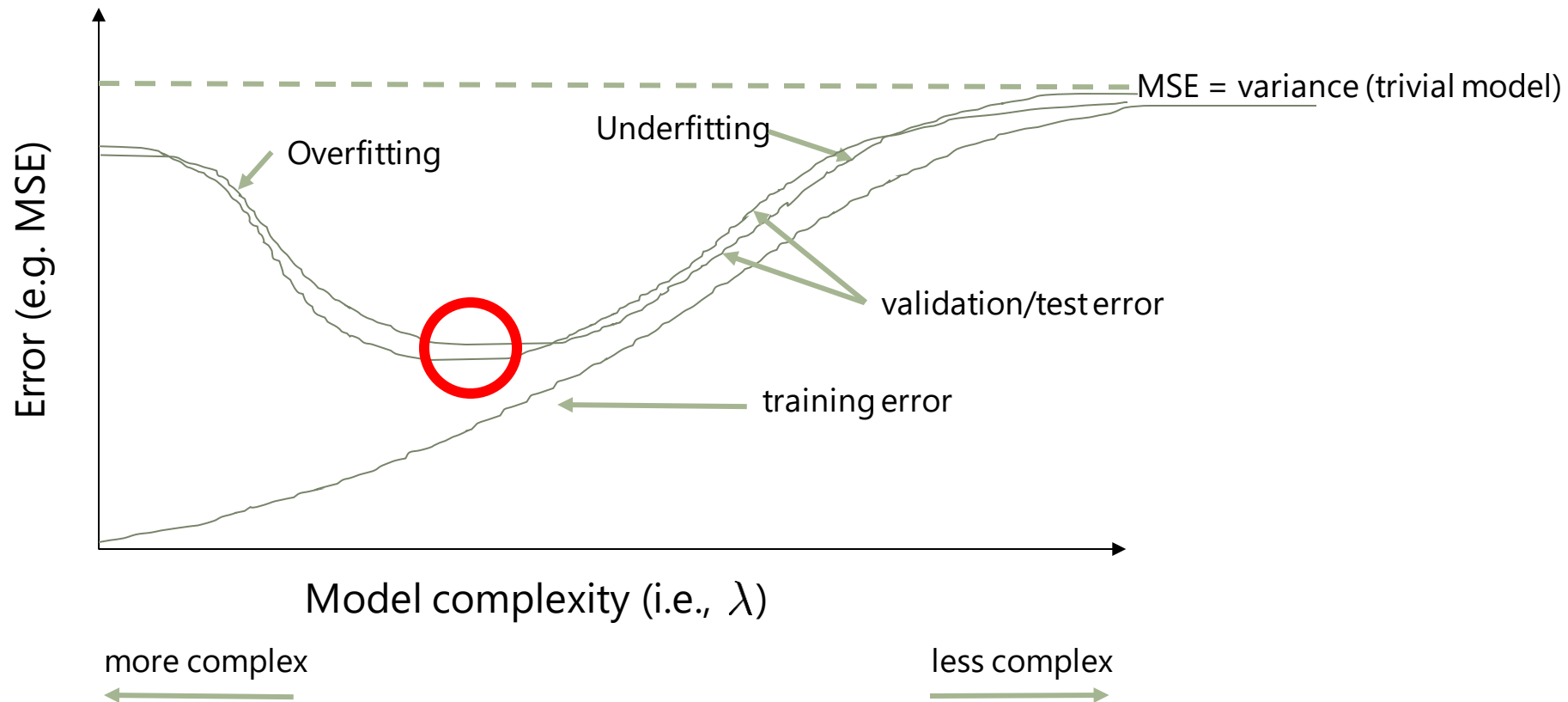
# Choosing among several different models

## 4. What values of lambda should we choose

- Our validation "theorems" can guide us to good choices of values
- E.g. given two values of lambda ( $a$  and  $b > a$ ), if the **validation error** is higher for  $a$ , then we should try **larger** values than  $b$ ; if the validation error is higher for  $b$ , we should try **smaller** values than  $a$

# Choosing among several different models

## 4. What values of lambda should we choose



# Summary of concepts

- Gave practical advice as to how to select regularization parameters
- Showed how the same concepts can be used to guide other modeling decisions (e.g. different feature representation options)

On your own...

- Use these guidelines to optimize various model parameters, e.g. the dictionary size, text-processing options, etc., as well as the hyperparameter lambda