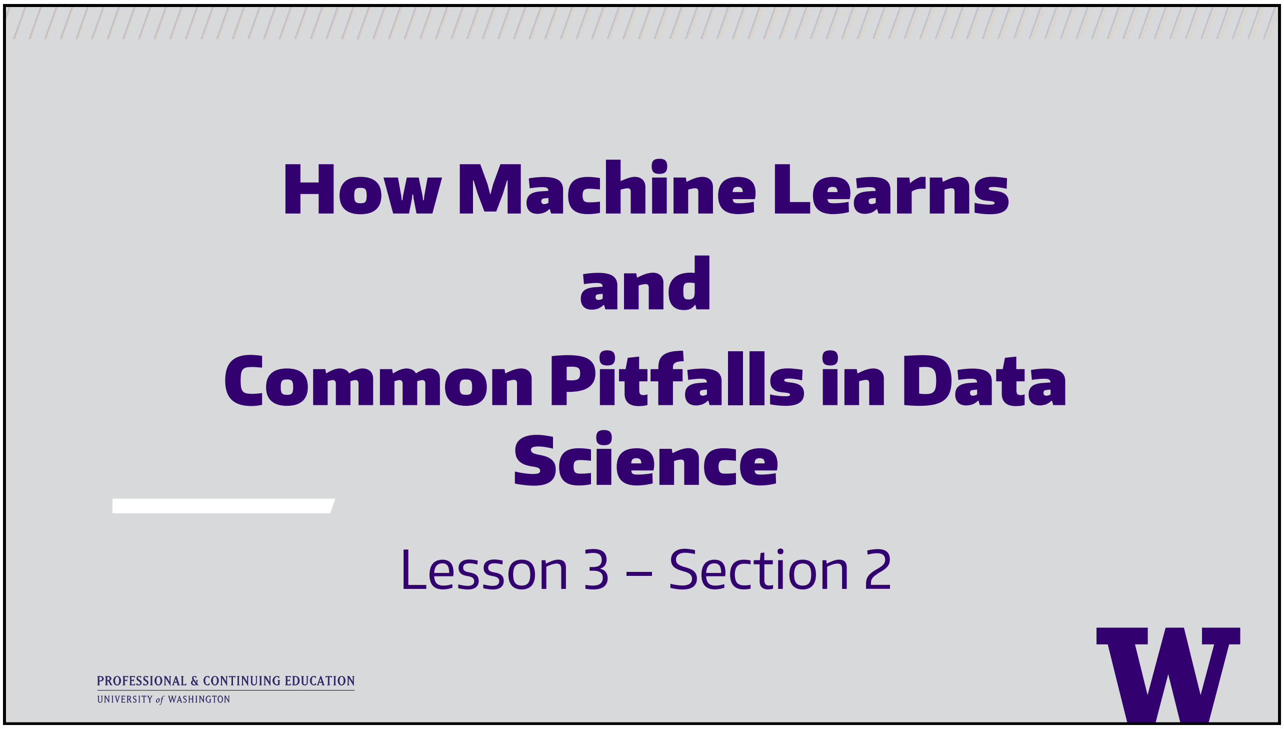
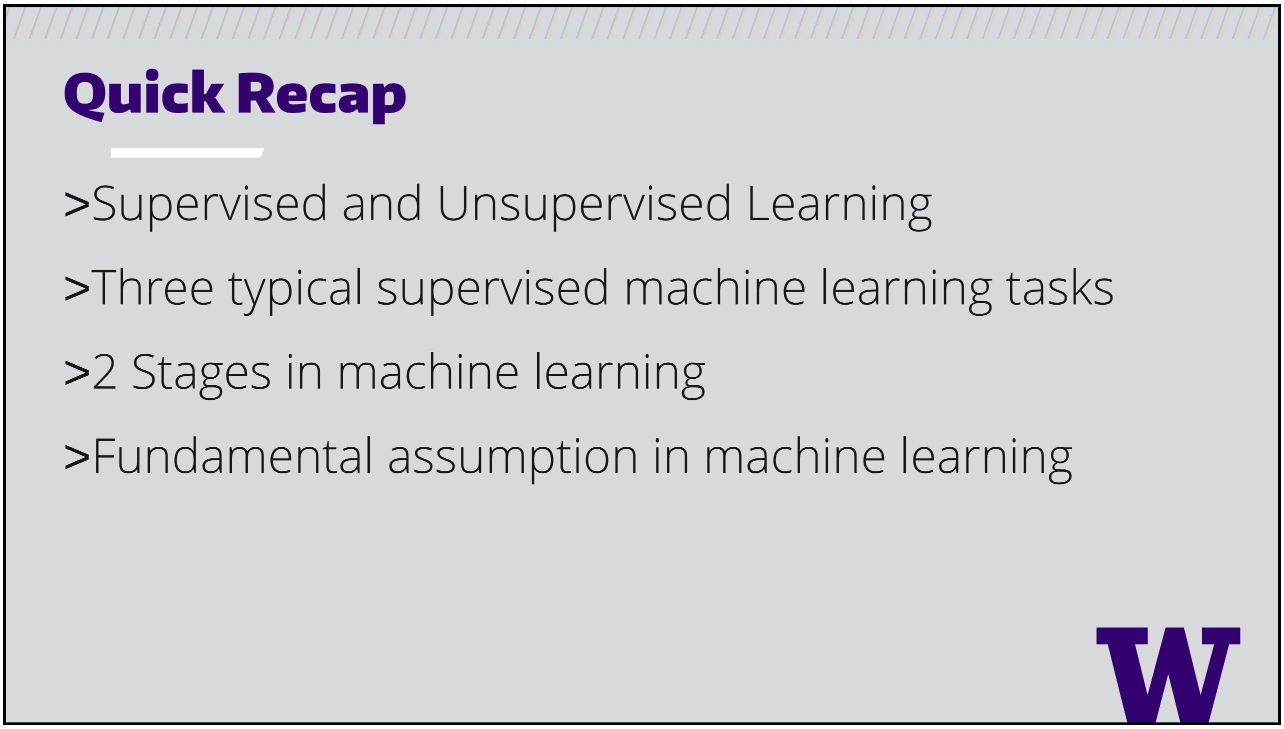


How Machine Learns and Common Pitfalls in Data Science

Lesson 3 – Section 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Quick Recap

>Supervised and Unsupervised Learning

>Three typical supervised machine learning tasks

>2 Stages in machine learning

>Fundamental assumption in machine learning

Overview

Mathematical Framework of Machine Learning.

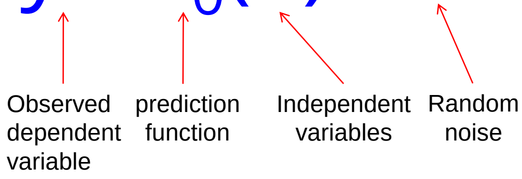
How Machine Learns from Training Data?

- Stochastic Gradient Descent

Three Common Pitfalls in Machine Learning.

The machine learning framework

$$y = f_\theta(\mathbf{x}) + \varepsilon$$

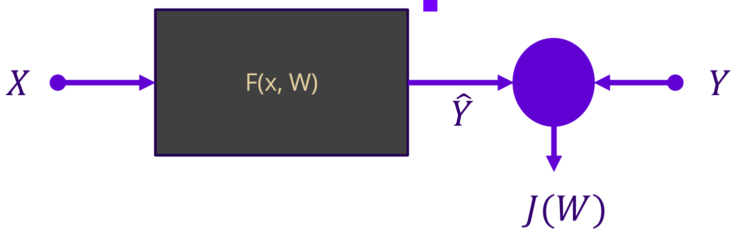


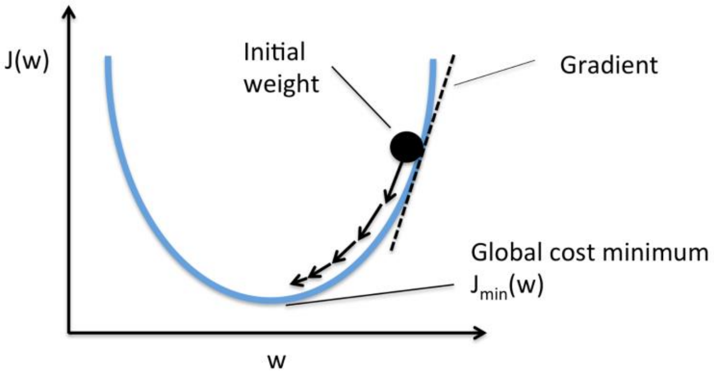
- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f and parameters θ which minimizes the prediction error on the training set

$$E_\theta(Y,X) = \sum_{i=1}^{N} \left(y_i - \widehat{f}_\theta(x_i)\right)^2$$

- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

How to learn model parameters θ?





Many classifiers to choose from

- SVM
- Neural networks
- Naïve Bayes
- Bayesian network
- Logistic regression
- Randomized Forests
- Boosted Decision Trees
- K-nearest neighbor
- Etc.

Which is the best one?

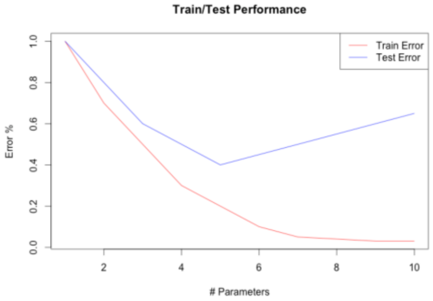
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Generalization

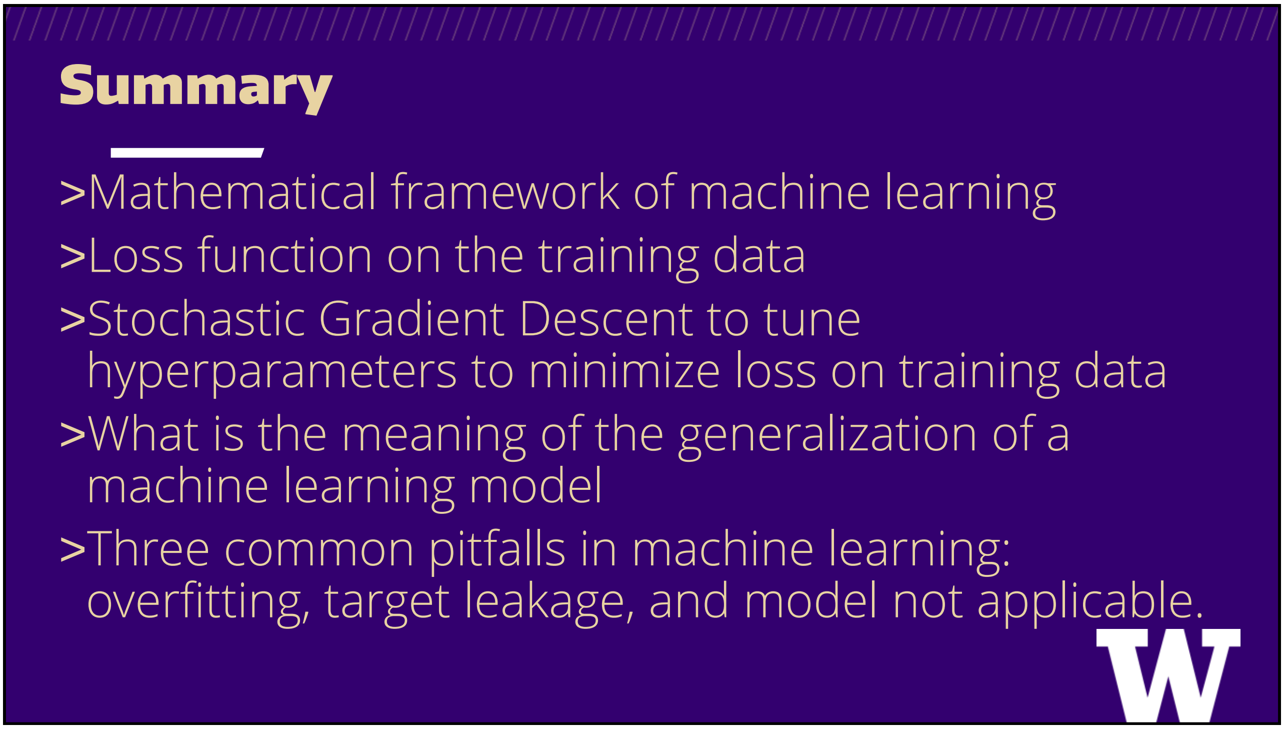
- What does the model generalization mean?
  - –We say a model generalizes well, meaning the model achieve similar performance on the training and validation data
  - –We need to split the original dataset into training and validation, in order to test the generalization of models. Usually 70-80% in training, and remainder in validation.
- **Underfitting:** model is too "simple" to represent all the relevant class characteristics
  - –High training error and high validation error
- **Overfitting:** model is too "complex" and fits irrelevant characteristics (noise) in the data
  - –Low training error and high validation error (big gap between training and validation performance)

Common Pitfalls in Machine Learning

- Overfitting



- Target leaking
- Model has good performance on validation, but not applicable
  - –Have to think about when the model is in production, whether you have data available for the variables of this model when prediction is made



Summary

- >Mathematical framework of machine learning
- >Loss function on the training data
- >Stochastic Gradient Descent to tune hyperparameters to minimize loss on training data
- >What is the meaning of the generalization of a machine learning model
- >Three common pitfalls in machine learning: overfitting, target leakage, and model not applicable.