Class 7 Supervised Learning Basics

Dr Wei Miao

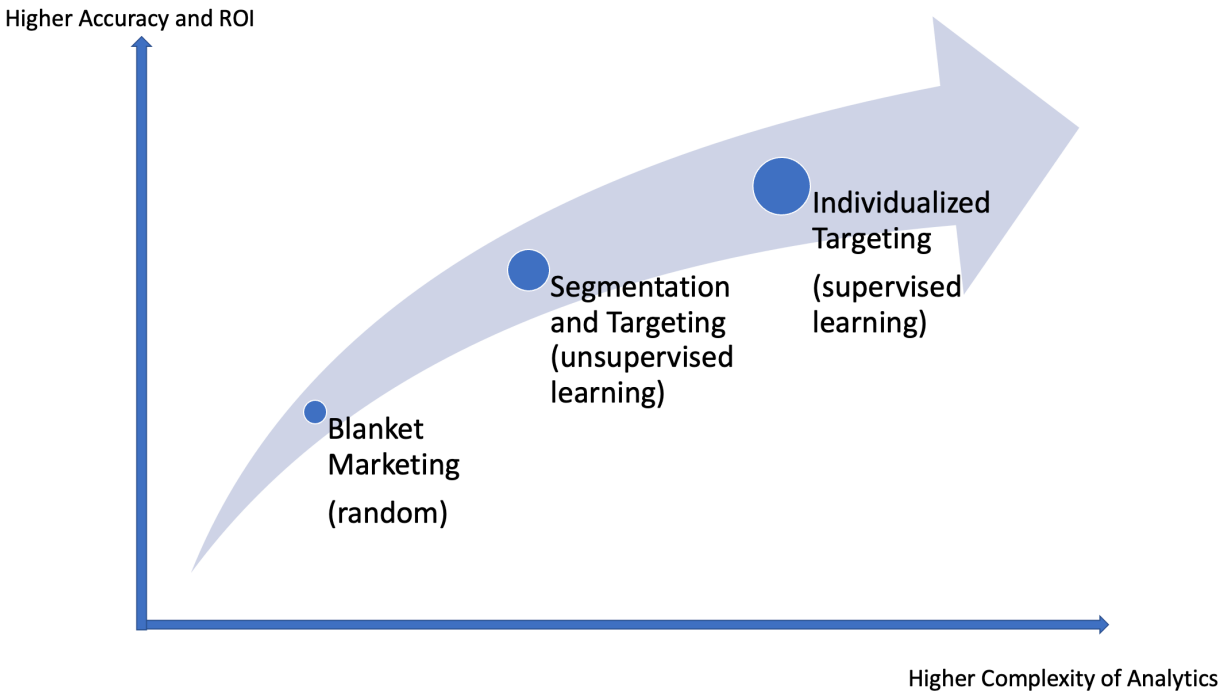
UCL School of Management

October 25, 2023

Section 1

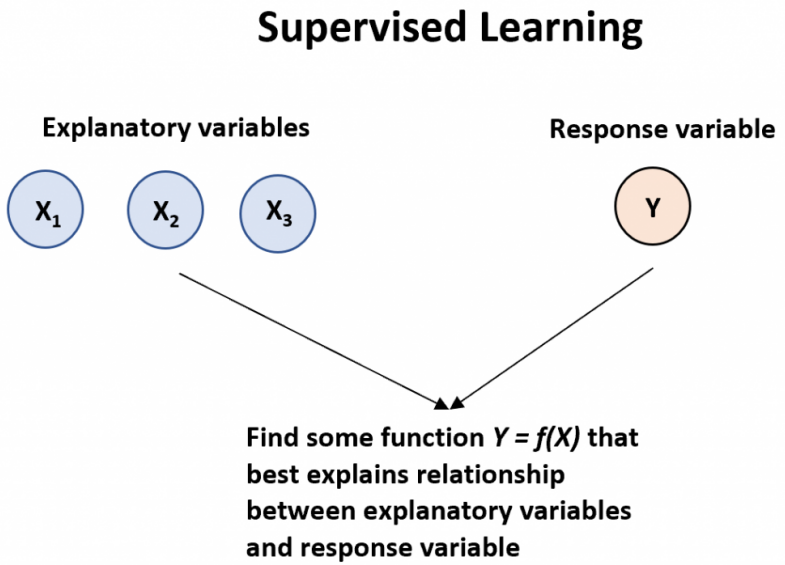
Supervised Learning

Motivation: Why Supervised Learning for Business?



Supervised Learning

- A **supervised learning model** is used when we have one or more **explanatory variables** AND **a response variable** and we would like to learn the relationship (data generating process, or DGP) between the **explanatory** variables and the **response** variable as accurately as possible.



Data Generating Process

$$Y = f(X; \theta) + \epsilon$$

- f is the function that characterizes the true relationship between X and Y, or DGP, which is never known to us[1]
- Y the **response or outcome** variable to be predicted, e.g., if a customer responded to our marketing offer.
- $X = (X_1, X_2, ..., X_p)$ are a set of **explanatory variables,** sometimes called features or predictors, e.g.,
  - (1) customers' past purchase history (e.g., spending in each category, frequency of purchase, recency of purchase)
  - (2) demographic variables (e.g., gender, income, age, kids, etc.)
- θ represents the set of **function parameters** to be trained
- ϵ is the **error term**

[1] "All model are wrong, but some are useful" – George Box. As business analysts, we need to use the "wrong models" correctly.

Types of Supervised Learning Algorithms

Depending on the type of the **response variable**, supervised learning tasks can be divided into two groups:

- **Classification tasks** if the outcome is **categorical**
  - Whether a customer responds to marketing offers
  - Whether a customer churns
  - Which product a customer purchases
- **Regression tasks** if the outcome is **continuous**
  - Customer total spending in each period
  - Demand forecasting such as the daily sales of a product

Difference between Supervised and Unsupervised Learning

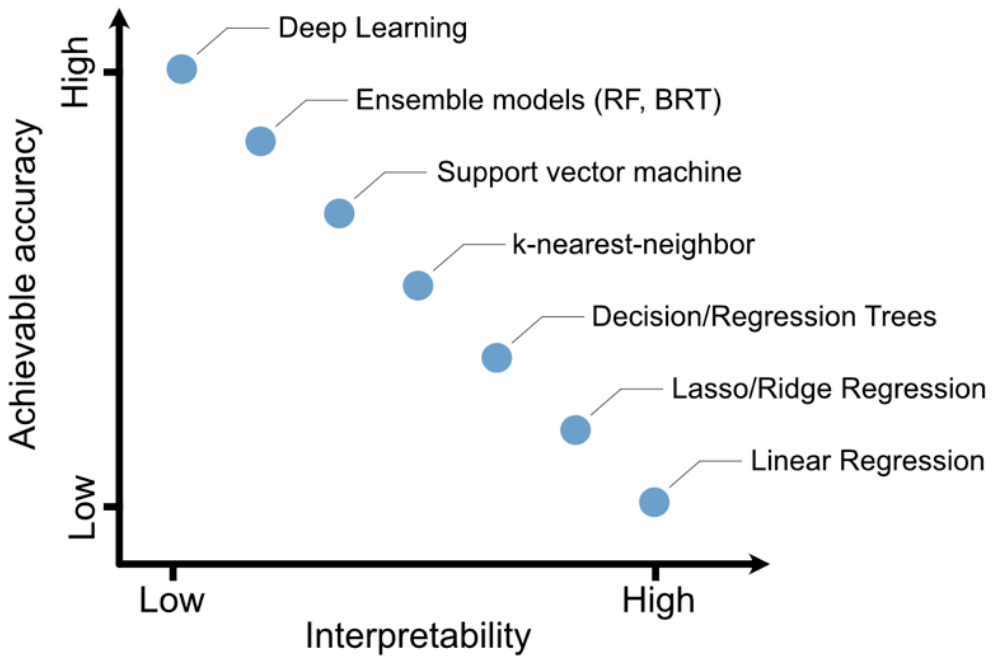
	Supervised Learning	Unsupervised Learning
Description	Involves building a model to estimate or predict an output based on one or more inputs.	Involves finding structure and relationships from inputs. There is no "supervising" output.
Variables	Explanatory and Response variables	Explanatory variables only
End goal	Develop model to **(1)** predict new values or **(2)** understand existing relationship between explanatory and response variables	Develop model to **(1)** place observations from a dataset into a specific cluster or to **(2)** create rules to identify associations between variables.
Types of algorithms	**(1)** Regression and **(2)** Classification	**(1)** Clustering and **(2)** Association

Section 2

Fundamental Tradeoffs

Accuracy-Interpretability Tradeoff

- Simpler models are easier to interpret but gives lower accuracy
- Complicated models can give better prediction accuracy but results are hard to interpret

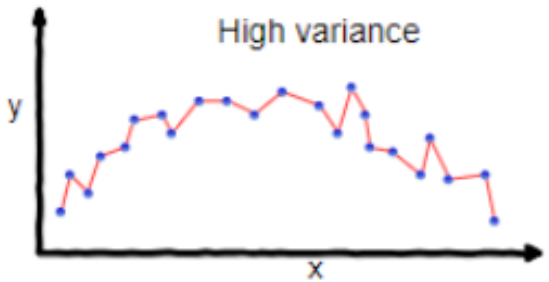


Comparision of Classic Supervised Learning Models

- Linear regression class models (easy to interpret, low accuracy)
  - Linear regression coefficients have economic interpretations but prediction accuracy is low
- **Tree-based Models (balance between interpretability and accuracy)**
  - Decision tree, random forest
- Neural-network based models (hard to interpret, high accuracy)
  - Deep learning only give estimated weights that have no direct business interpretations

Bias-Variance Tradeoff

- After we have trained a machine learning model, **bias** is the prediction error of the model on the historical data; **variance** is the prediction error of the model on unseen, new data.
- If a predictive model **fits historical data too well**, then it may not be flexible enough to accommodate future data and thus have a higher chance of failing to make predictions for new data accurately. This problem is called **overfitting**.
- Overfitting leads to **low bias** but **high variance**. Hence the name bias-variance tradeoff or bias-variance dilemma.



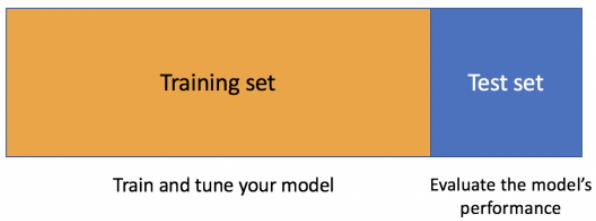
overfitting

Section 3

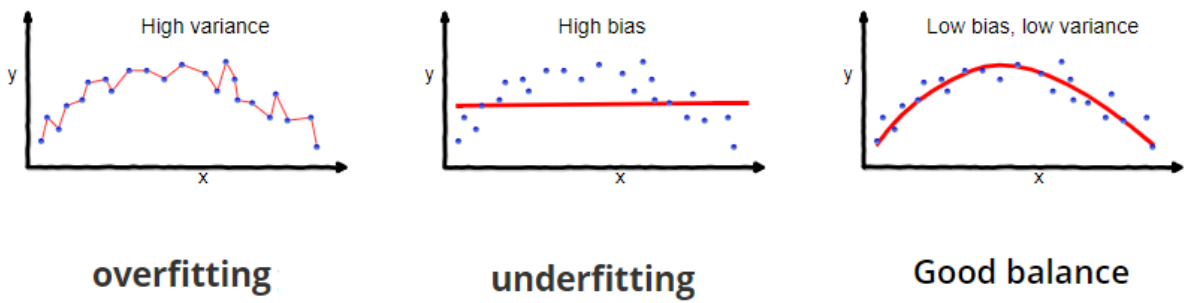
Overfitting and Underfitting

How to Mitigate Overfitting

- To mitigate the overfitting problem, when training predictive models, we need to use the **cross-validation** technique by splitting the full **historical data** into a **training set** and a **test set**.
  - **A training set** (70% - 80% of labelled data)**:** we train the predictive model based on the training set.
  - **A test set** (20% - 30% of labelled data)**:** we pretend that we don't know the outcomes for the test set and make predictions from the predictive model. However, in fact, we do observe the actual outcomes for the test set, so that we can evaluate the prediction accuracy by comparing the predicted outcomes versus the actual outcomes.



- For more complicated models with hyper-parameters such as deep learning models, we may even need to split our data into 3 sets (training, validation, and test sets).

Underfitting

- **Underfitting** occurs when a predictive model cannot sufficiently capture the DGP on both historical data and new data.
- Underfitting leads to **high bias** as well as **high variance**. Thus, underfitting is the worst case, which should be avoided by all means.
- To mitigate the underfitting problem, we need to select more suitable models.