

Introduction to Machine Learning

Shaobo Li

University of Kansas

Fall 2026

Recommended Textbook

- An Introduction to Statistical Learning with Application in Python (or R)
 - [Book website](#)
 - [Online course videos](#)
 - [Notes from a book club from the R4DS online learning community](#)
- Elements of Statistical Learning
 - [Book website](#)
 - Recommended for advanced students

- **Python** – Used in this class
 - Google Colab (no need to download and install)
 - VSCode
 - Cursor
- **R** – Optional
 - Download R, and install
 - Download RStudio, and install
- You can get start with Google Colab by watching and following this [video](#).

■ Data

- Textbook data (James, et al.): install R package [ISLR](#)
- [UCI Machine Learning Repository](#)
- [Kaggle](#)
- [KDD Nuggets](#)

■ Lectures and other tutorials

- [Videos of textbook \(ISLR\)](#)
- [DataCamp](#)
- [Coursera](#)

■ Your AI friends

- [ChatGPT](#), [Claude AI](#)
- many more...

■ Course website

- [a website](#) maintained by [Prof. Li](#)
- [Canvas](#) (for assignment and other work)

You should take this course if...

- You are passionate about AI and data science
- You know that data are assets for businesses and you like exploring data
- You want to pursue this career
- You want to pursue advanced degrees in data science and AI

An [article](#) on who should pursue a master's degree in BA.

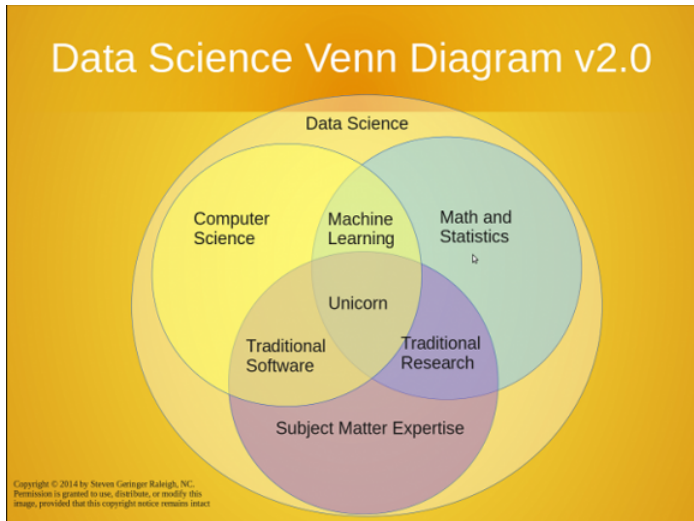
How much salary can a data scientist earn? ([a report from indeed](#))

Machine Learning in Different Fields

Learning from data is essential in different scientific disciplines

- Predict stock returns in next six months based on historical data;
- Predict the probability of a loan default based on customer's information and historical records;
- Identify certain diseases based on medical image;
- Identify handwritten digits from image;
- Facial recognition;
- Natural language processing;
- Cluster customers based on their purchase behavior and other information
-

In practice, you need...



Examples

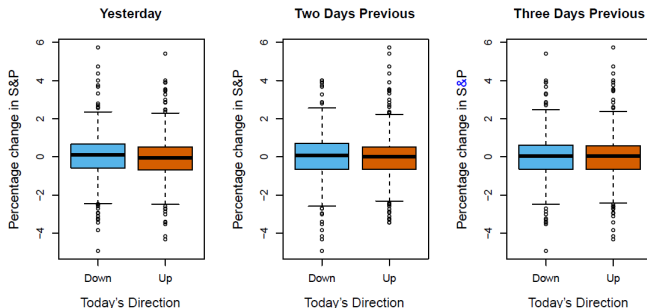
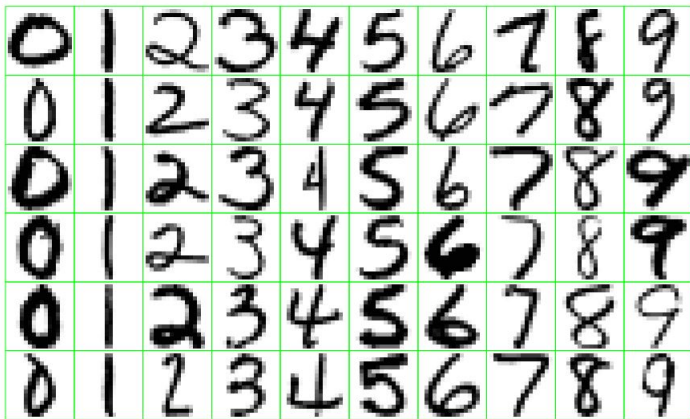


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the Smarket data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

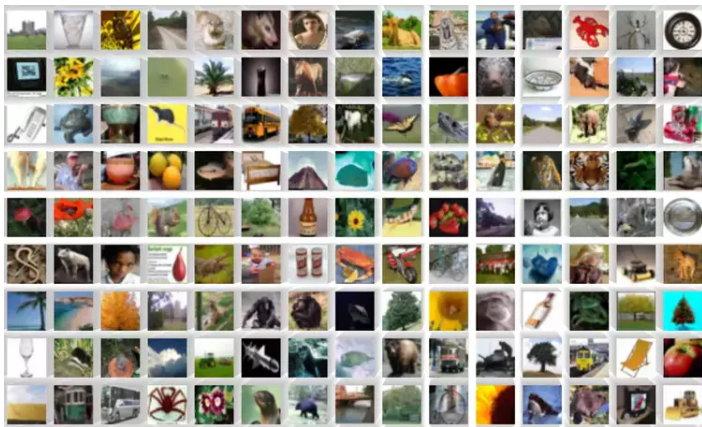
- This example is from ISLR page 3 (stock market data).
- Can you describe these figures?
- Can you define a business problem that ML may help?

Examples



- This is part of the famous [MNIST](#) dataset for image recognition.
- Can you define a business problem that ML may help?

Examples



- This is part of the famous **CIFAR-100** dataset for image recognition.
- There are 100 classes labeled manually by researchers.
- Can you define a business problem that ML may help?

- **Supervised learning**
 - There is specific response you need to predict
- **Unsupervised learning**
 - No response, instead, you need to create response based on some patterns
- **Semi-supervised learning**
 - Mixture of both

- **Supervised learning**
 - There is specific response you need to predict
- **Unsupervised learning**
 - No response, instead, you need to create response based on some patterns
- **Semi-supervised learning**
 - Mixture of both
- What type of learning is for large language model?

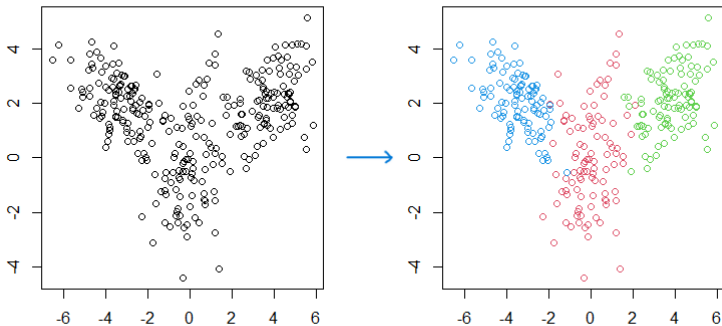
Unsupervised Learning

- Data is **unlabeled** (no Y 's)
- Uncover patterns, groups among X 's
- Subjective, no simple goal such as prediction
- Examples:
 - Marketing segmentation – clustering
 - Dimension reduction – principle component analysis (PCA)
 - Market basket analysis – association rules mining

A simple ML algorithm – k-means clustering

- Clustering: group similar observations together
- k-means: one out of many algorithms for clustering
 - k: how many groups
 - means: the center of the group
- This is an iterative algorithm
 - making a little progress many times
- Similarity measures, or distances, are the key component behind the algorithm

An illustration



Supervised Learning

- Labeled data
- The goal is to predict or explain certain outcome
- Type of problem:
 - Regression: outcome is continuous
 - Classification: outcome is categorical
- Popular ML algorithms:
 - Least square, nearest neighbor, CART, gradient boosting, neural network, deep learning

Supervised Learning

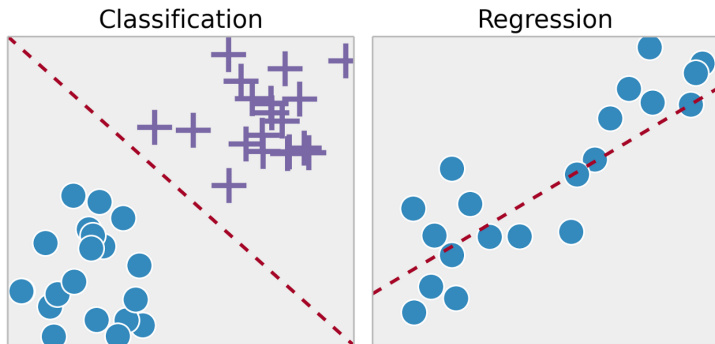
- Suppose we observe data Y_i and $\mathbf{x}_i = (X_{i1}, \dots, X_{ip})$ for $i = 1, \dots, n$.
- Y is the outcome (or response, dependent variable, target), and $\mathbf{x}$ is predictor (or independent variables, covariates, features, inputs)
- The learning problem can be modeled as

$$Y_i = f(\mathbf{x}_i) + \epsilon_i$$

where $f(\cdot)$ is unknown function, and ϵ is random error.

- A general workflow: Learn from the **training data** and test the trained model on a holdout sample (**testing data**).

Regression and Classification



- Response variable is continuous
 - e.g., stock return, housing price, temperature
- Goal: prediction
- Learning methods:
 - Parametric methods, e.g., linear regression
 - Nonparametric methods, e.g., decision tree, neural network
- Commonly used evaluation metrics:
 - Mean squared error (MSE), mean absolute error (MAE), R-square

An example for regression

Can we predict Sales using ad media spending?

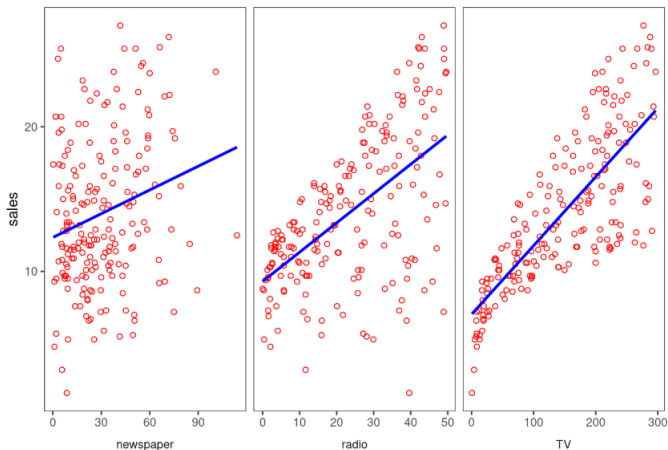
- Three channels of media: TV, radio, and newspaper
 - Spending on these channels are predictors, often denoted by X
- The outcome variable is sales, often denoted by Y
- Because our intuition is higher ads spending leads to higher sales. Thus, a linear regression can be developed to predict sales:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \epsilon$$

- Here the unknown function is a linear function, that is
 - $f(x) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3$
- The model needs to “learn” from the data to figure out the parameters: $\beta_0, \beta_1, \beta_2, \beta_3$. How?

Visualizing the data

Exploratory data analysis (EDA) is always important and a necessary step before building any models



Prediction accuracy for regression problem

- How good is the model?
 - model fitting (in-sample performance)
 - prediction (out-of-sample performance)

Prediction accuracy for regression problem

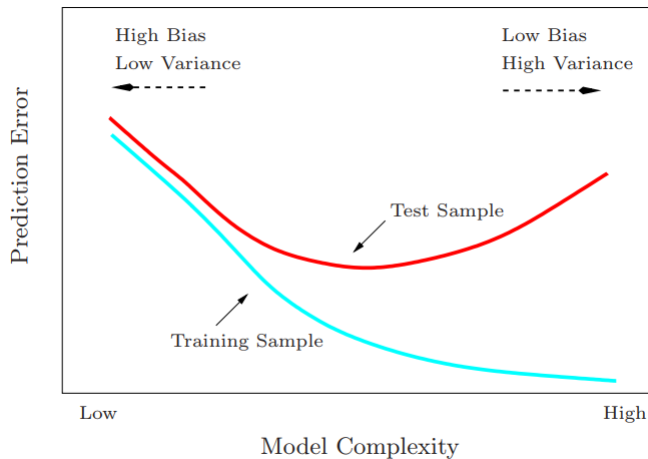
- How good is the model?
 - model fitting (in-sample performance)
 - prediction (out-of-sample performance)
- One of the most common metrics is the *mean squared error*
- Denote training set $Tr = \{x_i, y_i\}_1^N$, and testing set $Te = \{x_i, y_i\}_1^M$

$$MSE_{Tr} = \frac{1}{N} \sum_{i \in Tr} (y_i - \hat{f}(x_i))^2$$

$$MSE_{Te} = \frac{1}{M} \sum_{i \in Te} (y_i - \hat{f}(x_i))^2$$

- Training error, MSE_{Tr} , may be biased due to overfitting
- In this course, we denote *MSE* as training error, and *MSPE* as testing error

Prediction error – a tradeoff

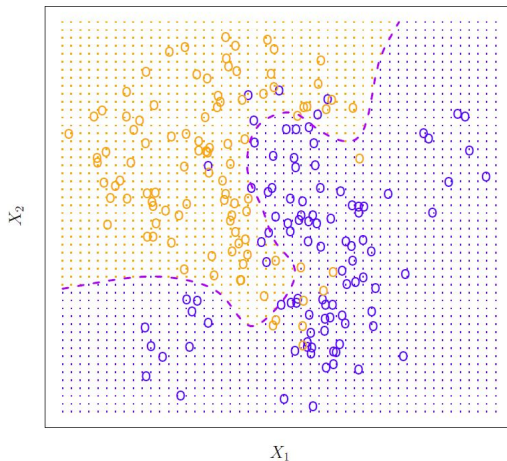




Classification

- Response variable is categorical
 - e.g., {A, B, C}, {dog, cat}, {0, 1}
- Goal: prediction (classification)
- Learning methods:
 - Parametric methods, e.g., logistic regression
 - Nonparametric methods, e.g., decision tree, neural network
- an example of neural networks: [\[link\]](#)
- Commonly used evaluation metrics:
 - misclassification error, false positive rate (FPR), false negative rate (FNR), AUC, F1 score, pseudo R-square, etc.

Classification – an illustration¹



¹Source: ISLR pp.38, Figure 2.13

A simple classifier: K-nearest neighbor

- Model:

$$\hat{Y} = \hat{f}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i$$

where $N_k(\mathbf{x})$ is the neighborhood of $\mathbf{x}$ defined by the k closest points $\mathbf{x}_i$.

- k : size of the neighbor
 - smaller or bigger neighbor is better?
- How to find the neighbor of any point $\mathbf{x}$?

Prediction accuracy for classification problems

- Same question: How good is the model?
 - model fitting (in-sample performance)
 - prediction (out-of-sample performance)

Prediction accuracy for classification problems

- Same question: How good is the model?
 - model fitting (in-sample performance)
 - prediction (out-of-sample performance)
- Response variable is **qualitative**, which has no numeric meaning
- A classifier $\hat{C}(x)$ either correctly predict x or not
- A commonly used prediction error: Misclassification rate (MR)

$$MR_{Te} = \frac{1}{M} \sum_{i \in Te} \mathbb{I}[y_i \neq \hat{C}(x_i)]$$

Prediction accuracy for classification problems

- Same question: How good is the model?
 - model fitting (in-sample performance)
 - prediction (out-of-sample performance)
- Response variable is **qualitative**, which has no numeric meaning
- A classifier $\hat{C}(x)$ either correctly predict x or not
- A commonly used prediction error: Misclassification rate (MR)

$$MR_{Te} = \frac{1}{M} \sum_{i \in Te} \mathbb{I}[y_i \neq \hat{C}(x_i)]$$

- Most classifier $\hat{C}(x)$ involves certain decision rule
 - As a learning outcome, the machine may say “a flower with characteristics x is more likely to be setosa than other types.”

K-fold cross-validation

- Instead of doing train and test once, we can do it K times

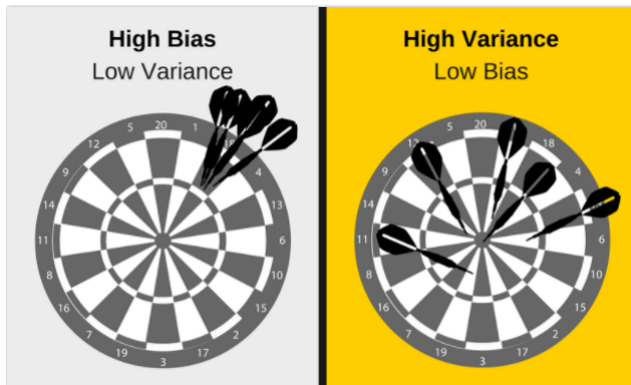
Folder 1	Folder 2	Folder 3	Folder 4	Folder 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

- Use 2,3,4,5 as training and 1 as testing
- Use 1,3,4,5 as training and 2 as testing
- Keep doing this loop...
- Average 5 testing errors, that is CV score
- In practice, $K = 5$ or 10 is recommended

Bias-Variance Tradeoff

- This is a very important tradeoff that governs the choice of statistical learning methods.
- **Bias:** how far the estimated model $\hat{f}(x)$ is to the true model $f(x)$.
 - Unbiased estimate is defined as: $\mathbb{E}\hat{f}(x) = f(x)$
 - Usually, we calculate the squared bias: $(\mathbb{E}\hat{f}(x) - f(x))^2$
- **Variance:** the variation of estimated model $\hat{f}(x)$ based on different training set.

Bias-Variance Tradeoff



Source: [link](#)

Bias-Variance Tradeoff

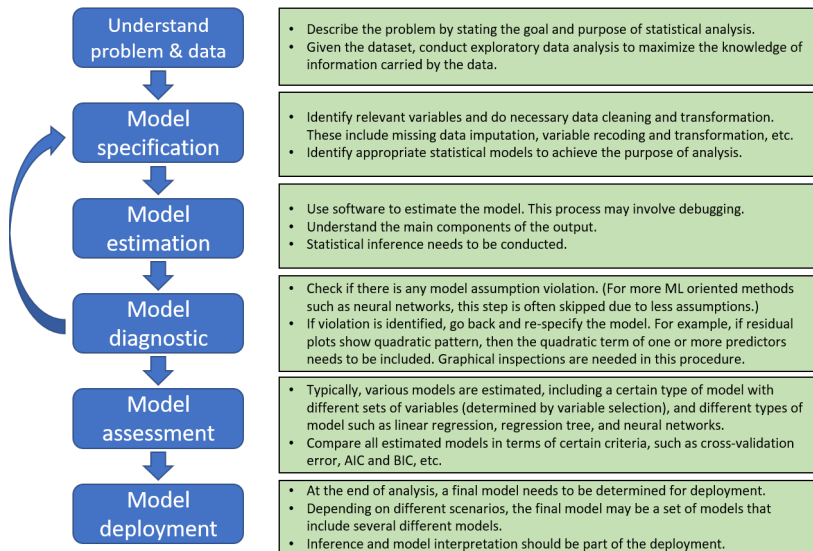
- Suppose the data arise from a model $Y = f(x) + \epsilon$, with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$.
- Suppose $\hat{f}(x)$ is trained based on some training data, and let (x_0, y_0) be a test observation from the same population.
- The *expected prediction error* can be decomposed to:

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \sigma^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0))$$

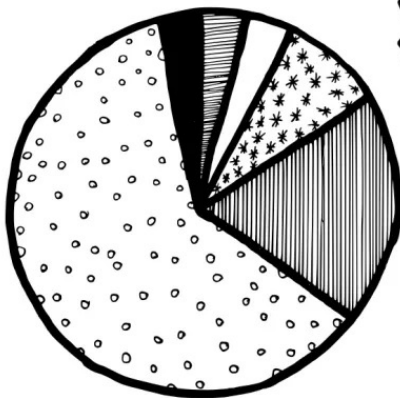
(Challenge yourself: Show it.)

- Typically as the **flexibility** of $\hat{f}$ increases, its variance increases, and its bias decreases.

Model building process



The reality



WHAT DATA SCIENTISTS SPEND THEIR TIME DOING



知乎 @李沐