MACHINE LEARNING - I

UNIT # I

1

TODAY'S AGENDA

- Course Overview
- Applications of ML
- Introduction to Regression Analysis
- Parametric vs Non-Parametric Models
- Model Assessment and Overfitting vs Underfitting
- Variance-Bias Tradeoff
- Introduction to Python

2

COURSE OUTLINE

- Regression and Classification Methods
- Model Evaluation
- Feature Selection, Pipelines, Grid Search
- Time Series Methods
- Neural Networks and Deep Learning
- Recent Advancement: Interpretable Machine Learning and AutoML

3

LEARNING OBJECTIVES

- Students are expected to:
  - understand and apply mathematical concepts and algorithms used in supervised machine learning, including classification, regression, and time-series problems
  - evaluate and select the most appropriate machine learning method for a given problem, taking into account factors such as the bias-variance tradeoff, overfitting and underfitting, feature reduction, and loss functions
  - participate in data analytics competitions held at Kaggle.com and other sites and have proficiency in doing machine learning tasks using the Python language
  - understand the complete life cycle of building machine learning systems that include data pre-preparation, experimental design, and model selection, evaluation and interpretation

4

REFERENCE BOOKS

- An Introduction to Statistical Learning (2nd Edition) by James et al. (2020)
- Introduction to Machine Learning with Python by Muller and Guido (2017)
- Data Mining for Business Analytics in Python by Shmueli et al. (2020)
- Interpretable Machine Learning by Molnar (2019)

(TENTATIVE) MARKS DISTRIBUTION

Midterms	25%
Final	40%
Project	15%
Assignments	14%
Quizzes	6%

SOFTWARE AND KAGGLE COMPETITIONS (DATA REPOSITORY)

- https://www.kaggle.com/competitions (Kaggle)
- https://www.anaconda.com/products/individual (Python)
- https://code.visualstudio.com/ (Python IDE – Optional)
- https://www.knime.com/downloads (KNIME - Visual Programming – Optional)

MEETING HOURS

- Office (Tabba Building, Room 211) Hours:
  - Monday/Friday: 11:30 AM – 1:00 PM
  - or by appointment (by e-mailing me at sahaider@iba.edu.pk).
- Note: I **DO NOT** entertain SMS/WhatsApp messages. E-mail is the official medium of correspondence.

MACHINE LEARNING IN DAILY USE

- E-mails (Gmail, Outlook)

- YouTube

- Social Networks (Facebook, Amazon)

9

APPLICATIONS OF MACHINE LEARNING

- Traffic Predictions
  - Google Maps
- Online Transportation Networks
  - Uber/Careem for price prediction
- Video Surveillance
  - Crime detection
- Fraud Detection
  - Financial institutions

10

APPLICATIONS OF MACHINE LEARNING (CONT'D)

- Social Media Services
  - Face recognition by Facebook
  - Hate speech detection by Facebook/Twitter
  - Inappropriate content by YouTube
- Emails
- Product Recommendation
  - Amazon, YouTube, and others
- Machine Translation
- Autonomous Vehicles

HISTORY

- At the beginning of the nineteenth century, the method of least squares was developed, implementing the earliest form of what is now known as linear regression.
- In the 1940s, various authors put forth an alternative approach, logistic regression.
- By the end of the 1970s, many more techniques for learning from data were available. However, they were almost exclusively linear methods because fitting non-linear relationships was computationally difficult at the time.
- By the 1980s, computing technology had finally improved sufficiently that non-linear methods were no longer computationally prohibitive. In the mid 1980s, classification and regression trees were developed.
- Neural networks gained popularity in the 1980s, and support vector machines arose in the 1990s.

DATA SCIENCE AND THE ART OF PERSUASION (HBR 2019)

- Despite the success stories, many companies aren't getting the value they could from data science.
- Four of the top seven "barriers faced at work":
  - lack of management/financial support
  - lack of clear questions to answer
  - results not used by decision makers and
  - explaining data science to others

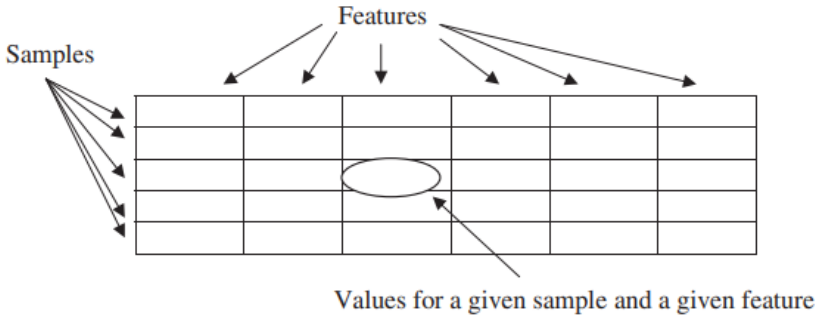
DO YOUR DATA SCIENTISTS KNOW THE 'WHY' BEHIND THEIR WORK? (HBR 2019)

- Data science, broadly defined, has been around for a long time. But the failure rates of big data projects in general and AI projects in particular remain disturbingly high.
- The following were found to be the two most important reasons:
  - Many data scientists are much more interested in pursuing their crafts — namely, finding interesting nuggets buried in data — than they are in solving business problems.
  - From the company's perspective, the talent is rare and protecting data scientists from the chaos of everyday work just makes sense. But doing so increases the distance between data scientists and the company's most important problems and opportunities.

STATISTICAL LEARNING VS MACHINE LEARNING (CHATGPT)

- Machine learning is a subfield of artificial intelligence that focuses on building systems that can learn from data and make predictions or decisions without being explicitly programmed. It includes a wide range of algorithms and techniques, such as decision trees, random forests, neural networks, and deep learning.
- Statistical learning, on the other hand, is a subfield of statistics that focuses on using statistical methods to understand and make predictions from data. It includes techniques such as linear and logistic regression, Lasso and Ridge regression, and principal component analysis (PCA).

SUPERVISED VS UNSUPERVISED LEARNING



17

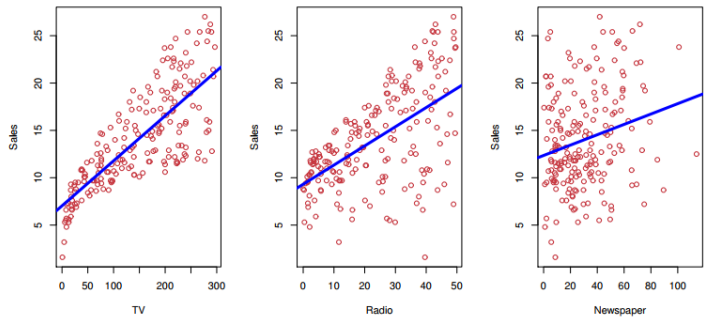
SAMPLE DATA: ADVERTISING

- 200 Rows
- 4 Columns
- Input variables or predictors or features or independent variables or variables
- Output variable or dependent variable or response

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7

18

SAMPLE DATA: ADVERTISING (CONT'D)

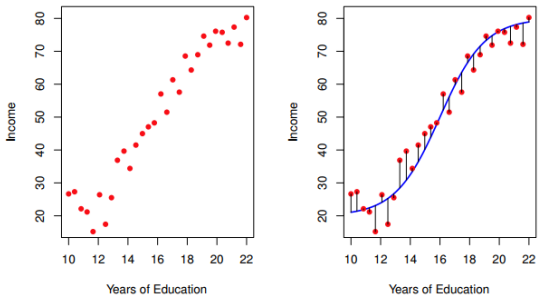


INFERENCE TASK

- Which predictors are associated with the response?
  - Identifying the few important predictors among a large set of possible variables can be extremely useful, depending on the application.
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?
- Consider the Advertising data, one may be interested in answering questions such as:
  - Which media are associated with sales?
  - Which media generate the biggest boost in sales? Or
  - How large of an increase in sales is associated with a given increase in TV advertising?

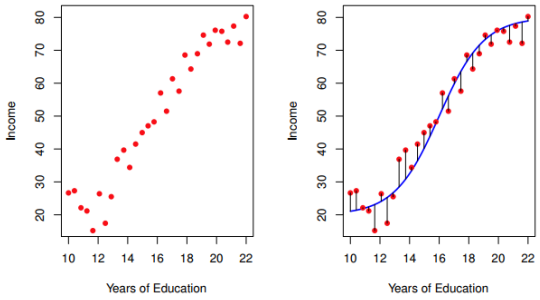
SAMPLE DATA: INCOME

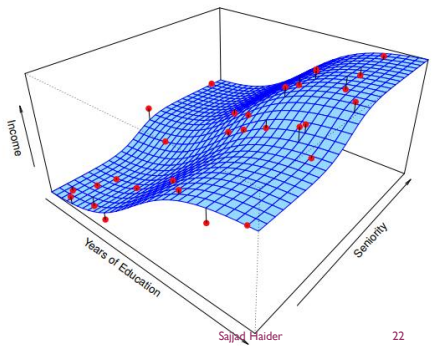
Education	Seniority	Income
21.5862069	113.103448	99.91717261
18.2758621	119.310345	92.57913486
12.0689655	100.689655	34.67872715
17.0344828	187.586207	78.70280624
19.9310345	20	68.00992165
18.2758621	26.2068966	71.50448538
19.9310345	150.344828	87.97046699
21.1724138	82.0689655	79.81102983
20.3448276	88.2758621	90.00632711
10	113.103448	45.6555295
13.7241379	51.0344828	31.91380794
18.6896552	144.137931	96.2829968
11.6551724	20	27.9825049
16.6206897	94.4827586	66.60179242
10	187.586207	41.53199242



SAMPLE DATA: INCOME (CONT'D)

- The blue curve represents the true underlying relationship between income and years of education, which is generally unknown (but is known in this case because the data were simulated).
- The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.





REDUCIBLE ERROR

$$\begin{aligned} \mathrm{E}(Y-\hat{Y})^2 &= \mathrm{E}[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

- The accuracy of Y^ depends on two quantities: reducible and irreducible error.
- f^ will not be a perfect estimate for f. This error is reducible because we can potentially improve the accuracy of f^ by using the most appropriate statistical learning technique.
- However, even if it were possible to form a perfect estimate for f, our prediction would still have some error in it! This is because Y is also a function of ε, which, by definition, cannot be predicted using X.

IRREDUCIBLE ERROR

- Why is the irreducible error larger than zero?
  - The quantity may contain unmeasured variables that are useful in predicting Y : since we don't measure them, f cannot use them for its prediction.
  - The quantity may also contain unmeasurable variation. For example, the risk of an adverse reaction might vary for a given patient on a given day, depending on manufacturing variation in the drug itself or the patient's general feeling of well-being on that day.
- The focus of this course is on techniques for estimating f with the aim of minimizing the reducible error.

PREDICTION EXAMPLE

- Consider a company that is interested in conducting a direct-marketing campaign.
- The goal is to identify individuals who are likely to respond positively to a mailing, based on observations of demographic variables. In this case, the demographic variables serve as predictors, and response to the marketing campaign (either positive or negative) serves as the outcome.
- The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants to accurately predict the response using the predictors. This is an example of modeling for prediction.

INFERENCE OR PREDICTION?

- Brand of a product that a customer might purchase based on variables such as price, store location, discount levels, competition price, and so forth.
  - Inference Problem: To what extent is the product's price associated with sales?
- In a real estate setting, one may seek to relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, and so forth.
  - Inference Problem: how much extra will a house be worth if it has a view of the river? This is an inference problem.
  - Prediction Problem: Alternatively, one may simply be interested in predicting the value of a home given its characteristics: is this house under- or overvalued?

TRADE-OFF

- Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate.
- For example, linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches.
- In contrast, some of the highly non-linear approaches that we discuss in the later chapters of this book can potentially provide quite accurate predictions for Y , but this comes at the expense of a less interpretable model for which inference is more challenging.

ESTIMATING F

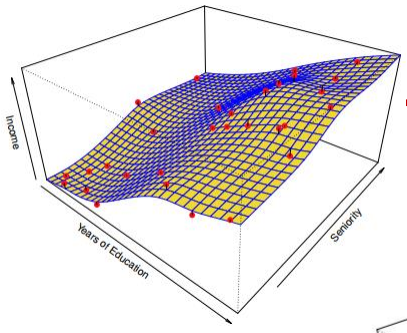
- While estimating f, we always assume that we have observed a set of n different data points. These observations are called the training data because we will use these observations to train, or teach, our method how to estimate f.
- Let xij represent the value of the jth predictor, or input, for observation i, where i = 1, 2, ..., n and j = 1, 2, ..., p. Correspondingly, let yi represent the response variable for the ith observation. Then our training data consist of {(x1; y1), (x2; y2), . . ., (xn; yn)}.
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f.

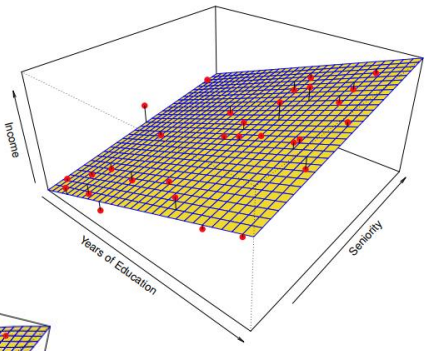
PARAMETRIC METHODS

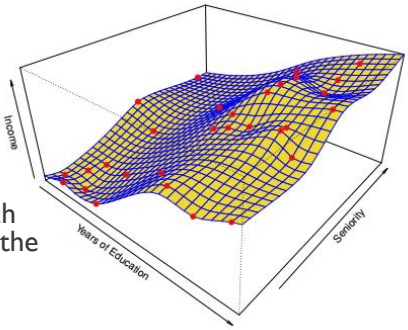
- Broadly speaking, most statistical learning methods for this task can be characterized as either parametric or non-parametric.
- Parametric methods involve a two-step model-based approach. The first assumption is about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X: $f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp$
- After a model has been selected, we need a procedure that uses the training data to fit or train the model. In the case of the linear model, we need to estimate the parameters $\beta 0, \beta 1, \beta 2, \ldots, \beta p$.
- The most common approach to fitting the model is referred to as (ordinary) least squares. However, least squares is one of many possible ways to fit the linear model.

UNDERFITTING VS OVERFITTING

- The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
- If the chosen model is too far from the true f, then our estimate will be poor.
- We can try to address this problem by choosing *flexible* models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters.
- These more complex models can lead to a phenomenon known as *overfitting* the data, which essentially means they follow the errors, or noise, too closely.







A rough spline fit which makes zero errors on the training data.

the true f has some curvature that is not captured in the linear fit.

31

NON-PARAMETRIC METHODS

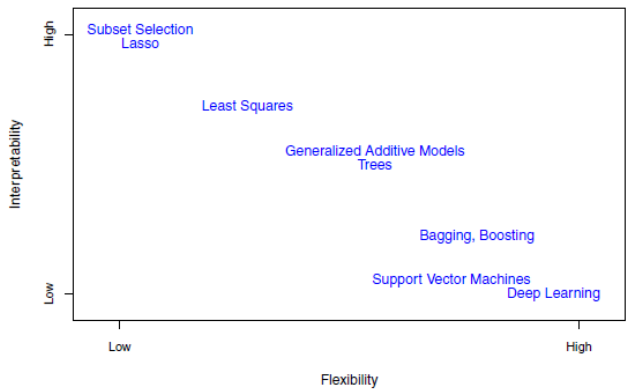
- Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible.
- By avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
- But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

32

TRADE-OFF BETWEEN ACCURACY AND INTERPRETABILITY

- Why would we ever choose to use a more restrictive method instead of a very flexible approach?
- If we are mainly interested in inference, then restrictive models are much more interpretable. For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and X1, X2, . . ., Xp.
- In contrast, very flexible approaches, can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

ACCURACY AND INTERPRETABILITY (CONT'D)



MEASURING THE QUALITY OF FIT

- Why is it necessary to learn different statistical learning approaches, rather than just a single best method?
- There is no free lunch in statistics: no one method dominates all others over all possible data sets.
- On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set. Hence it is an important task to decide for any given set of data which method produces the best results.
- Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

MEAN SQUARED ERROR

- To evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions match the observed data.
- The most commonly-used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

- The MSE will be small if the predicted responses are very close to the true responses and will be large if for some observations, the predicted and true responses differ substantially.
- In general, we do not really care how well the method works on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

TRAINING VS TESTING MSE

- When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don't exist in the test data.
- Regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE.

THE BIAS-VARIANCE TRADEOFF

- Variance refers to the amount by which ^f would change if we estimated it using a different training data set. Different training data sets will result in a different ^f.
- But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in ^f.
- In general, more flexible statistical methods have higher variance.

THE BIAS-VARIANCE TRADEOFF (CONT'D)

- Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- For example, linear regression assumes that there is a linear relationship between Y and X1, X2, . . ., Xp. It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f.
- Generally, more flexible methods result in less bias.

RECAP OF TRADEOFF

- Overfitting means a model fits the existing observations too well but fails to predict future new observations.
- This can occur when we're over extracting too much information from the training sets and making our model just work well with them, which is called **low bias** in machine learning. The model, as a result, will perform poorly on datasets that weren't seen before. We call this situation **high variance** in machine learning.
- The opposite scenario is underfitting. When a model is underfit, it doesn't perform well on the training sets and won't do so on the testing sets, which means it fails to capture the underlying trend of the data. We call any of these situations a **high bias** in machine learning; although its **variance is low** as the performance in training and test sets is pretty consistent, in a bad way.

PYTHON DEMO

PANDASGUI AND MITOSHEET