Predictive Analytics, Machine Learning, and Statistical Learning

AFM 346- FOUNDATIONS OF MACHINE LEARNING (PREDICTIVE ANALYTICS)

JESÚS CALDERÓN

Learning Objectives

By the end of this week, students will be able to:

- Define predictive analytics, machine learning, and statistical learning, as well as explain their differences and commonalities.
- Explain and provide examples of applications of predictive analytics in business and government. Discuss the implications and risks for business, government, and society.
- Explain the modelling process and its components.
- Explain the different classifications of models
  - Classification by model use: descriptive, inferential, and predictive.
  - Classification by mathematical qualities: parametric and non-parametric.
  - Classification by objective: supervised (regression/classification) and unsupervised learning.
- Define error in the context of supervised learning, as well as the variance-bias trade-off.
- Explain the interpretability-flexibility trade-off.

Introduction

What Does It All Mean?

Different terms mean different things and imply distinct perspective:

- **Predictive Analytics:** a business term, implies the production and use of actionable predictions.
- **Machine Learning:** a more technical term, a computer program that autonomously learns from examples how to perform a task and it gets better at it as it sees more examples.
- **Statistical Learning:** a subset of statistics that deals with the optimal extraction of statistical results from a data set. In many cases, it overlaps or runs parallel to machine learning, but not all machine learning models are statistical learning models.

Predictive Analytics

"Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision-making for candidate transactions."

(Wikipedia, 2020)

Predictive Analytics

- Uses data science techniques to make actionable predictions.
- Actionable predictions enhance organizations' ability to pursue their objectives.
- "Predictive Analytics" mostly used in business literature.
- Broad term, many times used to discuss implications on strategy and operations.

Machine Learning

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

(Mitchell, 1997)

Machine Learning (ML) is a collection of methods that allow a computer to:

- **Learn autonomously** to perform a task based on patterns in a set of examples and without being explicitly programmed.
- **Gain from experience** such that the method performs better in the measure that it observes more examples.
- **Results generalize** beyond the data that is used for training the method.

Machine Learning

- Technical and well-defined term.
- Includes other types of learning problems or objectives that are not predictive in the sense described before (e.g., reinforcement learning).
- Blends concepts from many disciplines, including computer science, statistics, information theory, and game theory.

Statistical Learning

"Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning. The field encompasses many methods such as the lasso and sparse regression, classification and regression trees, and boosting and support vector machines."

(James et al, 2013)

- Part of statistics concerned with optimizing a model selection process based on statistical methods that may be parametric or non-parametric.
- Solid statistical foundation that, under some circumstances, may provide inference about our results.
- Some concepts overlap with Machine Learning, but not all Machine Learning is Statistical Learning.

What Machine Learing is Not

- ML is not big data, cloud computing, or business intelligence.
- ML is not Artificial General Intelligence
  - ML has no autonomous intention, it has to be run for a specific purpose.
  - ML methods focus on specific objective (e.g., classify an observation), and not a general purpose (e.g., examine a bank and write a report).
- ML is not necessarily reliant on large data sets:
  - It is possible to apply some ML methods successfully with small data sets
- ML is not necessarily:
  - A "black box" model.
  - An explainable model.
- ML is not necessarily unbiased:
  - If training data contains bias, then trained model will be biased.
  - Even if data does not contain bias, the modeling choices may embed bias.
- ML is not a method to determine causality.
- ML is not a substitute for human cognition.
- ML is not a silver bullet for any business problem.

Why?

Why Machine Learning?

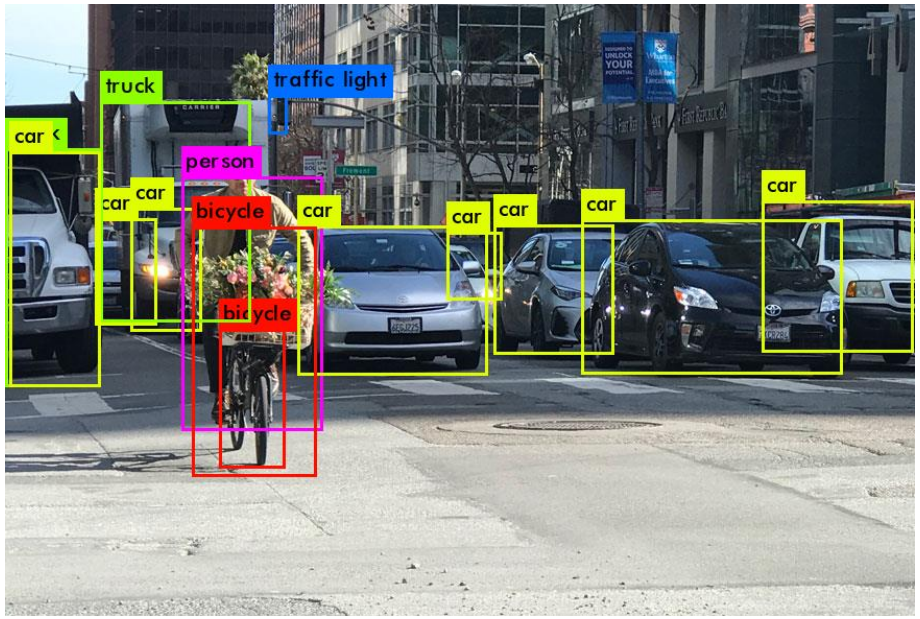
- Why Use Models?
- Why Use Machine Learning?
- Why Now?

Models Make Us Smarter and More Productive

- Understand complex phenomena
  - We have a limited ability to think about a large number of dimensions
  - Understand non-evident relationships
- Reason better
  - Reduce the likelihood of logical fallacies
  - Discard spurious relations
- Work with available data:
  - Granular and high-dimensional
  - Large volume and high velocity
  - Variable: structured/unstructured
  - Value must be extracted and not directly available
- Automation

Why Use Machine Learning?

- ML is used when a task is complex or impractical to program
- Machine learning can:
  - **Expand** capacity: through automation and economies of scale.
  - **Enhance** capabilities by reducing complexity of task or highlighting relevant details.
  - **Augment** capabilities by performing new tasks.



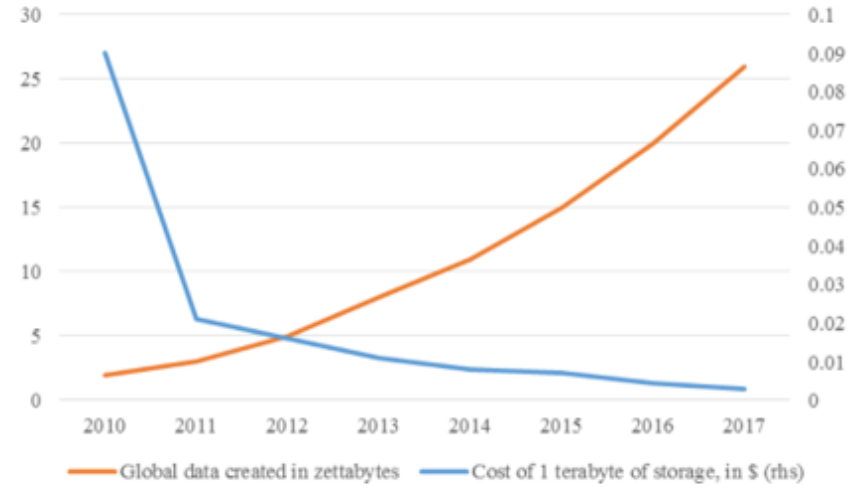
Object recognition is too complex to program directly

Why Now?

- Data and software have been a major driving force in the development and adoption of predictive analytics in the last 20 years
- Models have been around for a long time
- Computation capacity has been available for some time

(Ayers, 2008)

Figure 3: Costs of storage and global data availability, 2009-2017



Source: Reinsel, Gantz and Rydning (2017); Klein (2017). One zettabyte is equal to one billion terabytes.

Data storage is highly available (FSB, 2017)

A Brief History Review (1/4)

- **Early 19th Century**
  - **Models:** formalization of Bayes' Theorem (Laplace), normal distribution (Gauss), least squares method (Legendre)
- **Late 19th Century**
  - **Models:** "Wisdom of the Crowds" (Galton), standard deviation (Pearson)
- **1910s**
  - **Models:** Markov chains
  - **Business and society:** IBM was founded (Hollerith)
- **1920s**
  - **Data and code:** method of storing information magnetically (Pfleumer)
- **1930s**
  - **Models:** design of experiments, statistical significance, Linear Discriminant Analysis (Fisher)
- **1940s**
  - **Models:** Cox's Theorem, Information Theory (Sharon)

A Brief History Review (2/4)

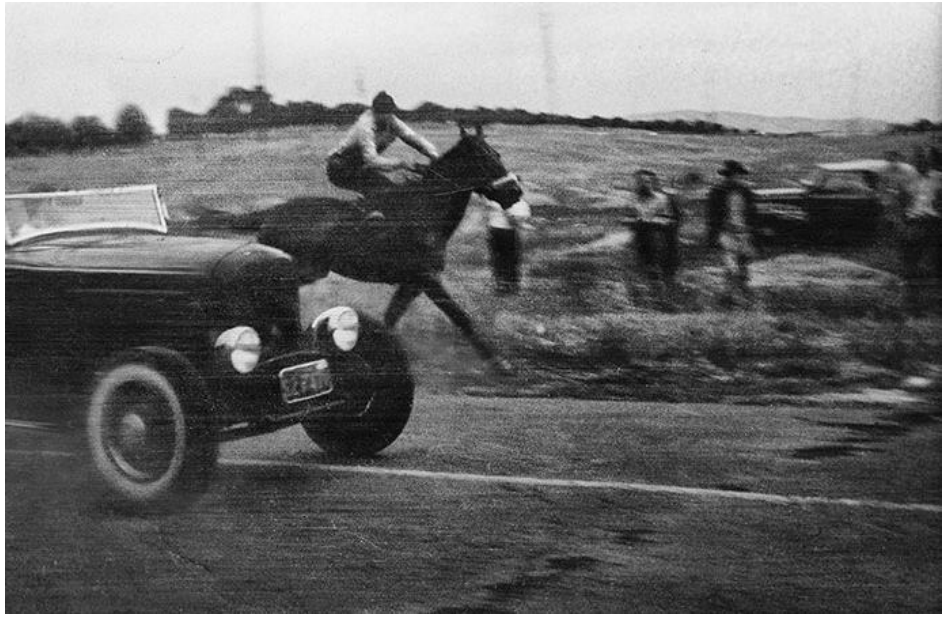
- **1950s**
  - **Models:** Turing's test, first neural network computer (Minsky and Edmonds), computer learns checkers (Samuel), perceptron for pattern and shape recognition (Rosenblatt)
  - **Business and society:** business intelligence (Luhn)
- **1960s**
  - **Models:** Naive Bayes (Maron), reinforcement learning model plays tic-tac-toe(Michie), nearest neighbor algorithm (Cover and Hart), Perceptrons (Minsky and Papert), automatic differentiation (Linnainmaa)
- **1970s**
  - **Data and code:** Relational Algebra (Codd), SQL (Chamberlin and Boyce)
  - **Business and society:** Lighthill Report to Parliament: First AI winter
- **1980s**
  - **Models:** explanation-based cognition (Dejong), Hopfield network (Hopfield), decision tree classifier (Quinlan), back propagation (Hinton et al.), Q-learning (Watkins)

A Brief History Review (3/4)

- **1990s**
  - **Models:** Random Forest (Tin Kam Ho), Support Vector Machines (Cortes and Vapnik), LSTM (Horchreiter and Schmidhuber)
  - **Data and code:** Internet/WWW (Nerners-Lee), R statistical programming language
  - **Business and society:** Deep Blue beats Kasparov, "Big Data" appears in print
- **2000s**
  - **Models:** MINST data set (LeCun), ImageNet (Fei-Fei Li)
  - **Data and code:** Torch library, Hadoop, digital information surpasses non-digital information
  - **Business and society:** Netflix Prize (1 MM for + 10% performance)

A Brief History Review (4/4)

- **2010s**
  - **Models:** AlexNet won ImageNet competition by large margin, ResNet (a major advance in ConvNet)
  - **Data and code:** Kaggle, Tensorflow (Google), rise of mobile devices
  - **Business and society:** Watson wins 'Jeopardy!', GoogleBrain - a large NN that recognizes cats from unlabeled YouTube images, DeepFace (Facebook), AlphaGo beats Lee Sedol at Go



Are we playing this game?

Types of Models

Different Classifications

- Classification by Model Use
- Types of Predictive Models
- Classification of Statistical/Machine Learning Methods

Classification by model use

- **Descriptive**
  - Describe or illustrate characteristics of some data.
  - Analysis might have no other purpose than to emphasize some trend or artifact of the data visually.
  - Example: locally estimated scatterplot smoothing model (LOESS).
- **Inferential**
  - Produce a decision for a research question or to test a specific hypothesis.
  - Objective: make a statement of truth regarding a predefined conjecture or idea. In many cases, a qualitative statement is produced ("statistically significant").
  - May make assumptions about the data and the underlying processes that generated the data (for example, iirv).
- **Predictive**
  - Produce the most accurate prediction possible for new data.
  - Objective: predicted values have the highest possible fidelity to the true value of the new data.
  - Problem is of *estimation* and not *inference*.
  - Depending on the context, the application may not even require a reason why a prediction was made.
  - Prediction can include measures of uncertainty.

Types of Predictive Models

Mechanistic, parametric, or structural model

- Derived using first principles to produce a model equation that is dependent on assumptions.
- Data are used to estimate unknown parameters of the equation so that predictions can be generated.
- Easy to make data-driven statements about how well the model performs based on how well it predicts the existing data.
- Require a two-step process for development:
  - Make an assumption about the functional form or shape of 'f()'.
  - After the model has been selected, use the training data to *fit* or *train* the model.

Empirically-driven or non-parametric models

- Seek an estimate of *f()* that gets as close to the data points as possible without being overly rough or wiggly.
- Makes more vague assumptions or does not make explicit assumptions about the functional form of *f()*.
- More flexible and can fit a wider range of possible shapes for *f()*.
- Given flexibility, requires more data than parametric models.
- Tend to fall into the machine learning category.
- May only be defined by the structure of the prediction and no theoretical or probabilistic assumptions are made about the variables.

(Khun and Silge, 2021; James et al, 2013)

Classification of Statistical/Machine Learning Methods

- **Supervised Learning**
  - Supervised learning problems are those in which there is an associated response measurement.
  - Many times the response measurement is called a "label" and the data set is called a labelled data set.
  - For example, a data set is provided with atmospheric measures like temperature, wind direction and speed, atmospheric temperature, relative humidity, and an indicator of the amount of rainfall in the next 24 hours.
  - A supervised learning problem can be formulated as:
    - **Regression:** predict a numeric variable. For example, predict the amount of rainfall in the next 24 hours.
    - **Classification:** predict a categorical variable. For example, predict if there will be more than 1 mm of rainfall in the next 24 hours.

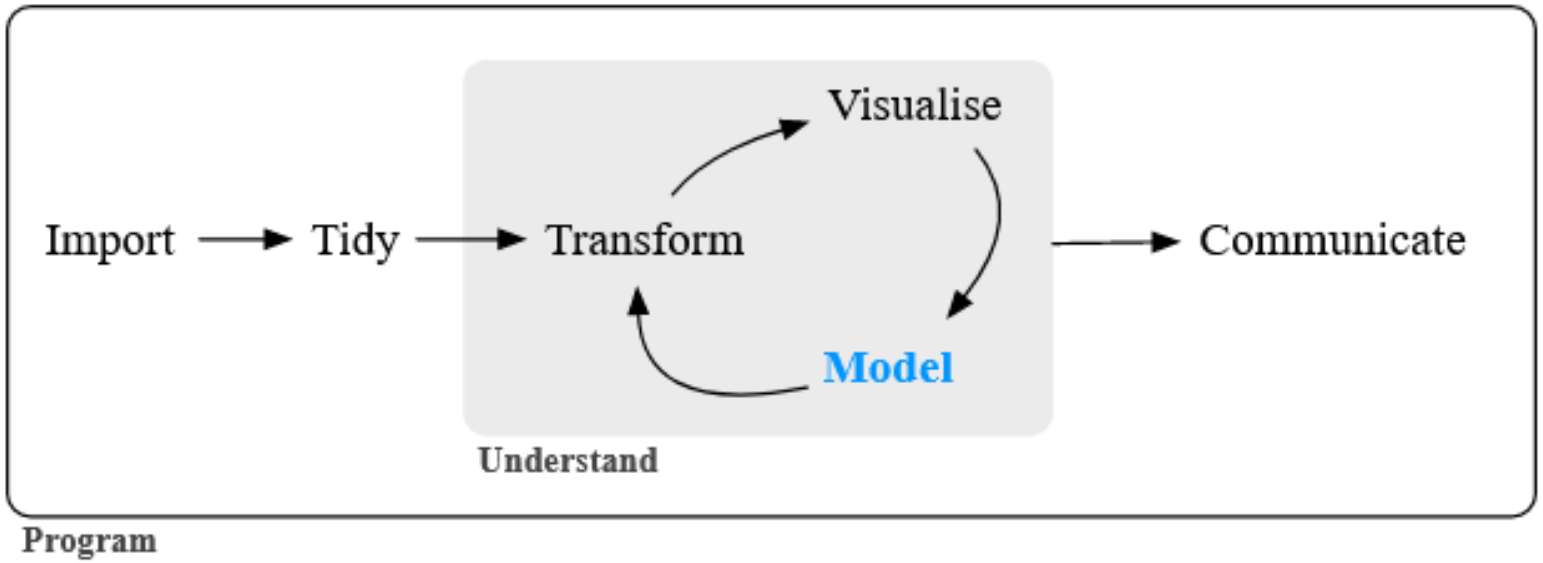
- **Unsupervised Learning**
  - Unsupervised learning problems are those in which our data does not contain labels but we want to extract patterns from our data.
  - Some unsupervised learning tasks:
    - **Dimensionality reduction:** compress the data by reducing the numbers of dimensions needed to express it. Some methods by which we can achieve this are Principal Components Analysis or Diffusion Maps.
    - **Clustering analysis:** group observations that are similar.
    - **Anomaly detection:** detect observations that are so dissimilar that we suspect that they were generated by a process that is different from the process that generates normal data.

An Overview of the Modelling Process

The Modelling Process

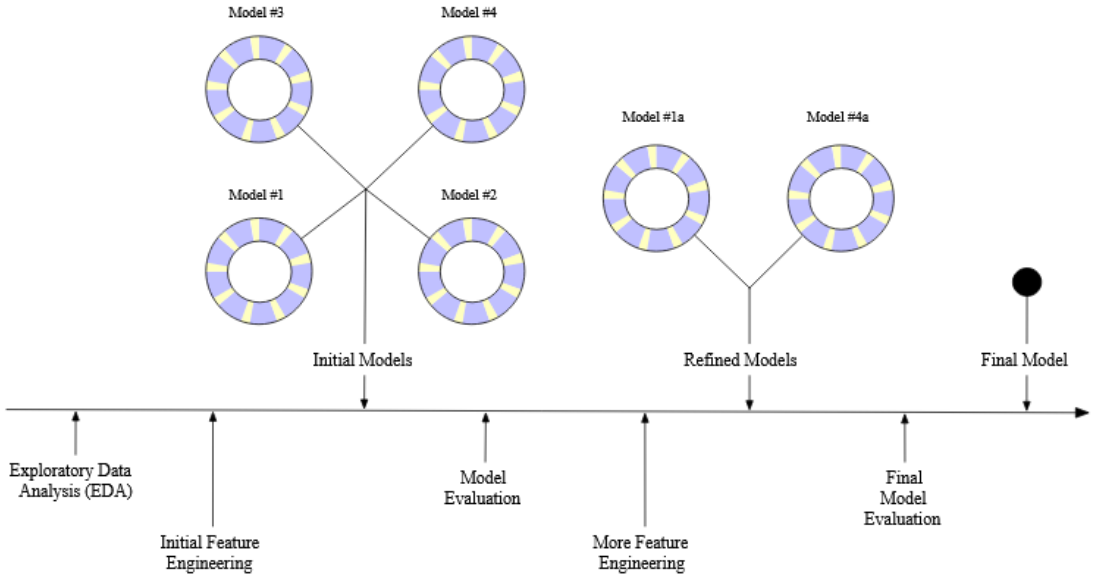
- The Data Science Process
- The Modelling Process
- A Reference Architecture of Model Deployment
- Why Try Many Models?
- Flexibility, Complexity, and Interpretability

The Data Science Process



The data science process (Wickham and Grolemund, 2020)

The Modelling Process (1/2)



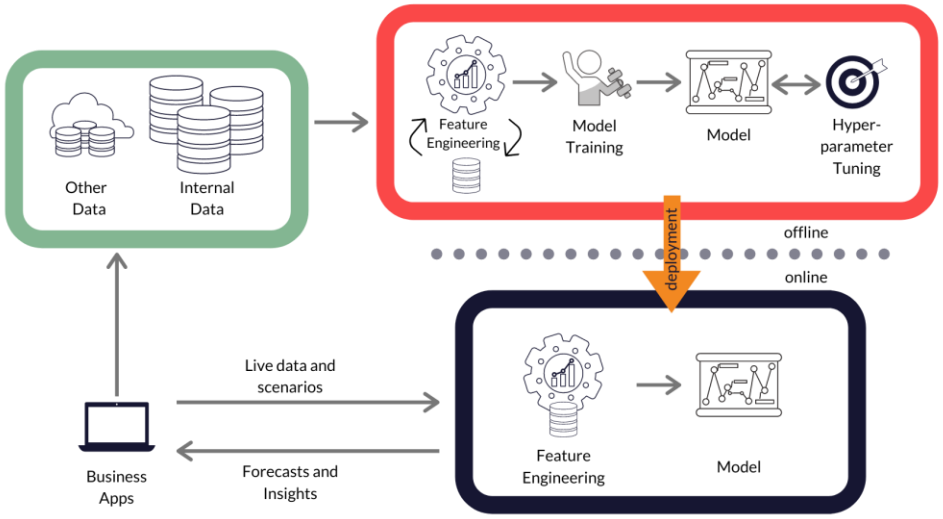
The modeling process (Khun and Silge, 2021)

The Modelling Process (2/2)

- **Exploratory data analysis (EDA):**
  - Initial back-and-forth between numerical analysis and visualization.
  - Different discoveries lead to more questions.
  - Data analysis "side-quests" to gain more understanding.
- **Feature engineering:**
  - Understanding gained from EDA results in the creation of specific model terms that make it easier to model the observed data accurately.
  - Includes complex methodologies (e.g., PCA) or simpler features (using the ratio of two predictors).
- **Model tuning and selection:**
  - A variety of models are generated, and their performance is compared.
  - Some models require parameter tuning where some structural parameters are needed to be specified or optimized.
  - Resampling/cross-validation is applied.
- **Model evaluation:**
  - During this phase of model development, we assess the model's performance metrics, examine residual plots, and conduct other EDA-like analyses to understand how well the models work.

(Khun and Silge, 2021)

Reference Architecture



MS Reference Architecture (Agrawal et al, 2020)

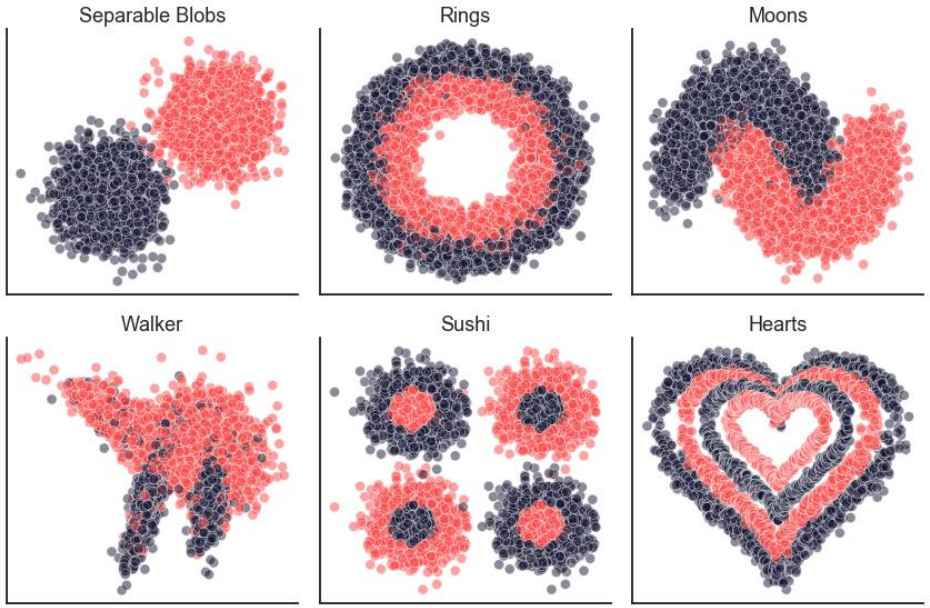
Why Try Many Models?

"A classifier must be represented in some formal language that the computer can handle. Conversely, choosing a representation for a learner is tantamount to choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space of the learner. If a classifier is not in the hypothesis space, it cannot be learned."

(Domingos, 2012)

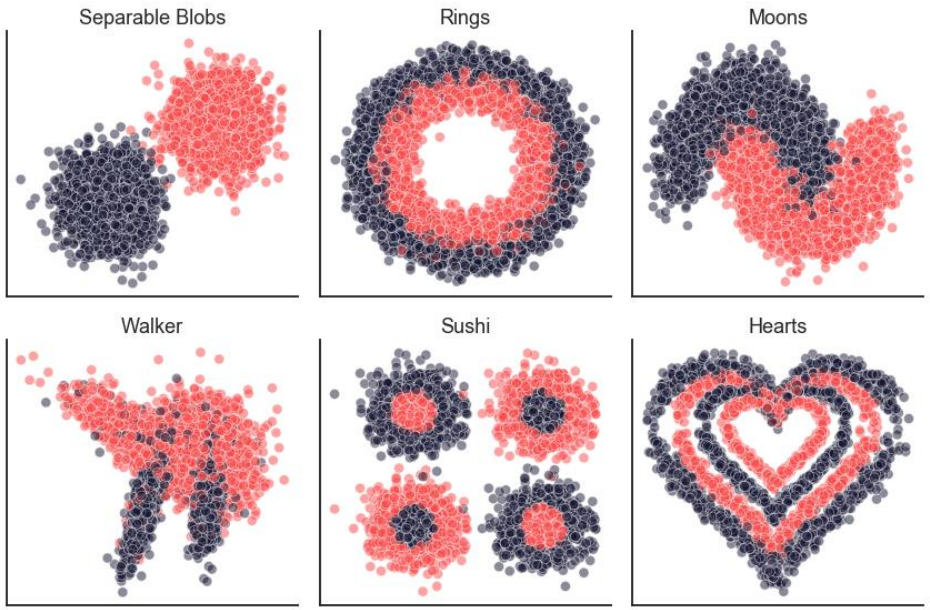
- The fundamental goal of ML is to generalize beyond examples in training data.
  - Doing well on the training set is easy; in some cases, trivial.
  - Unlikely, we will see the same training examples again, particularly in the world of Big Data.
- A test data set can be used to evaluate the model's performance without leaking information into training.
  - Contamination of classifier by test data can occur if it is used for parameter tuning.

Synthetic Examples

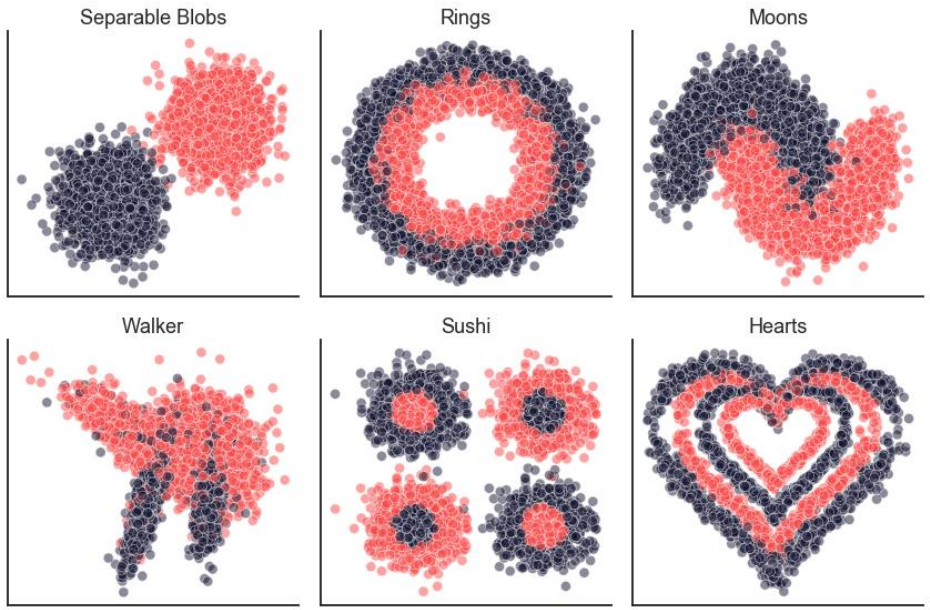


Synthetic Examples

Synthetic Examples

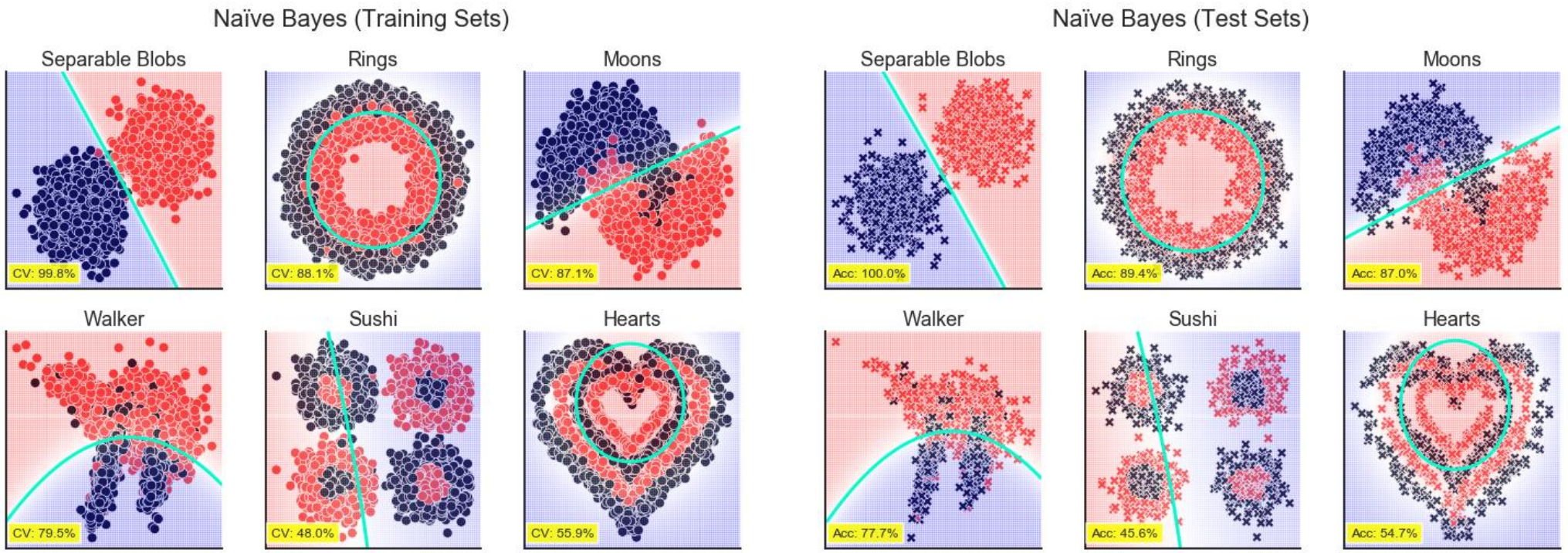


Training data (n = 4,000)

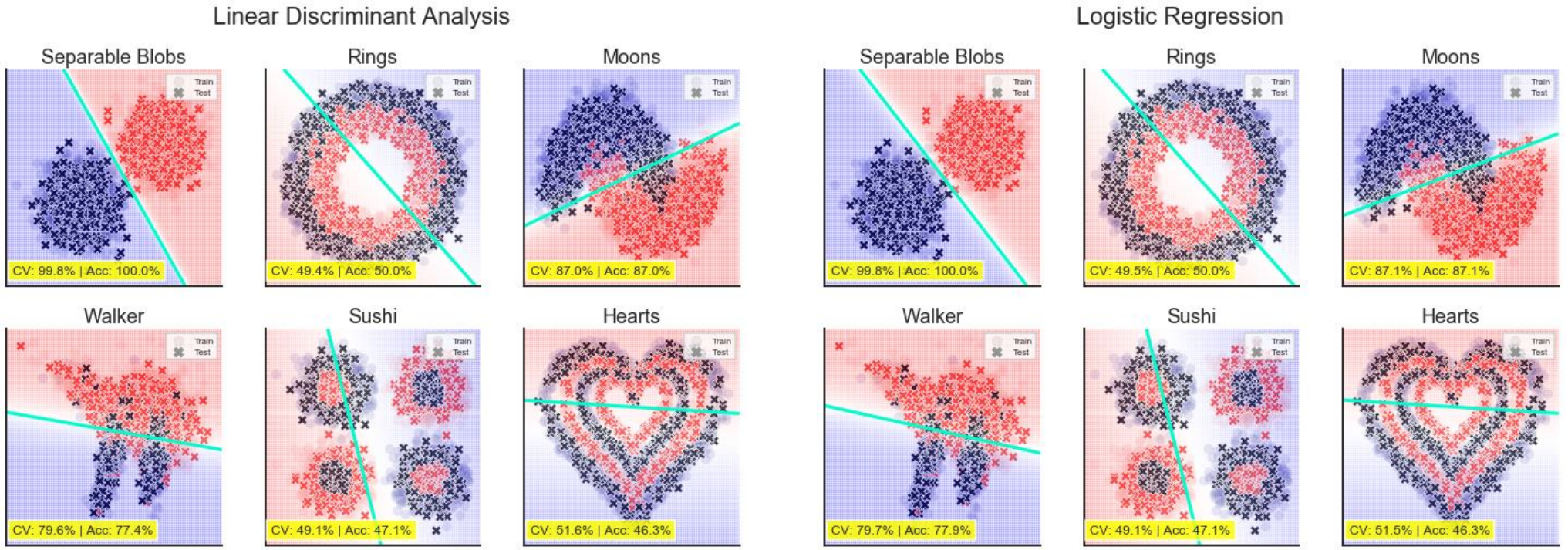


Test data (n = 1,000)

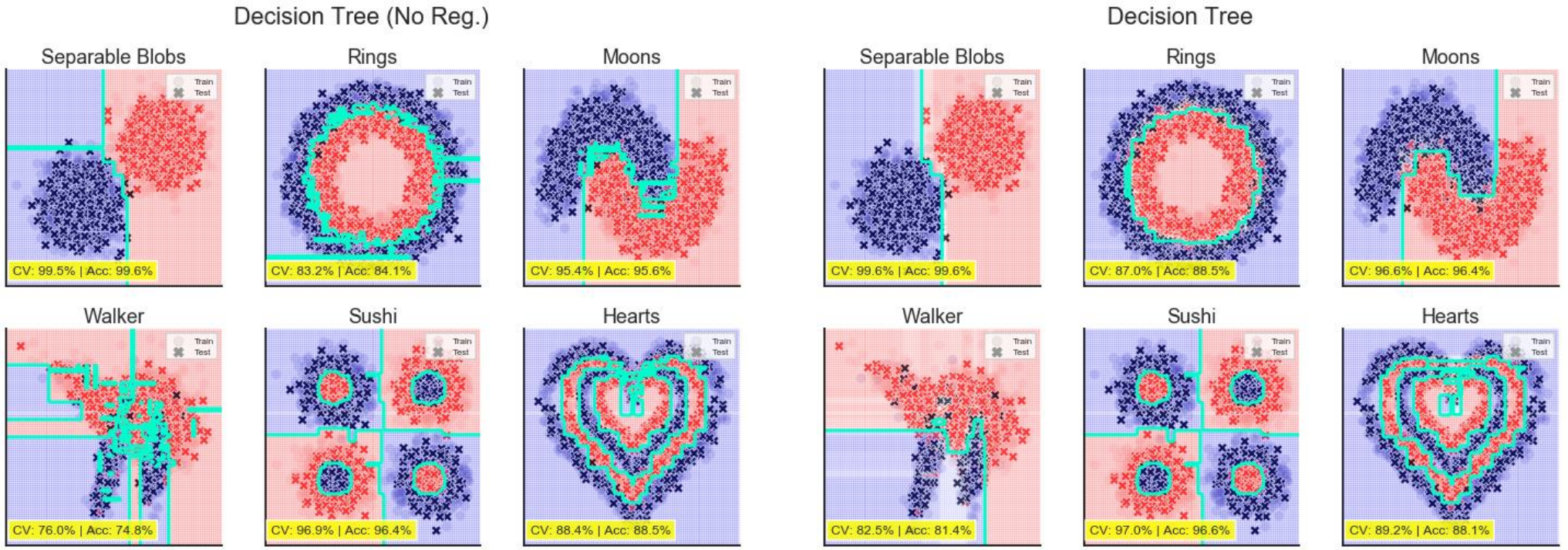
Naïve Bayes


Naïve Bayes (Training Sets)
Separable Blobs
CV: 99.8%
Rings
CV: 88.1%
Moons
CV: 87.1%
Walker
CV: 79.5%
Sushi
CV: 48.0%
Hearts
CV: 55.9%
Naïve Bayes (Test Sets)
Separable Blobs
Acc: 100.0%
Rings
Acc: 89.4%
Moons
Acc: 87.0%
Walker
Acc: 77.7%
Sushi
Acc: 45.6%
Hearts
Acc: 54.7%

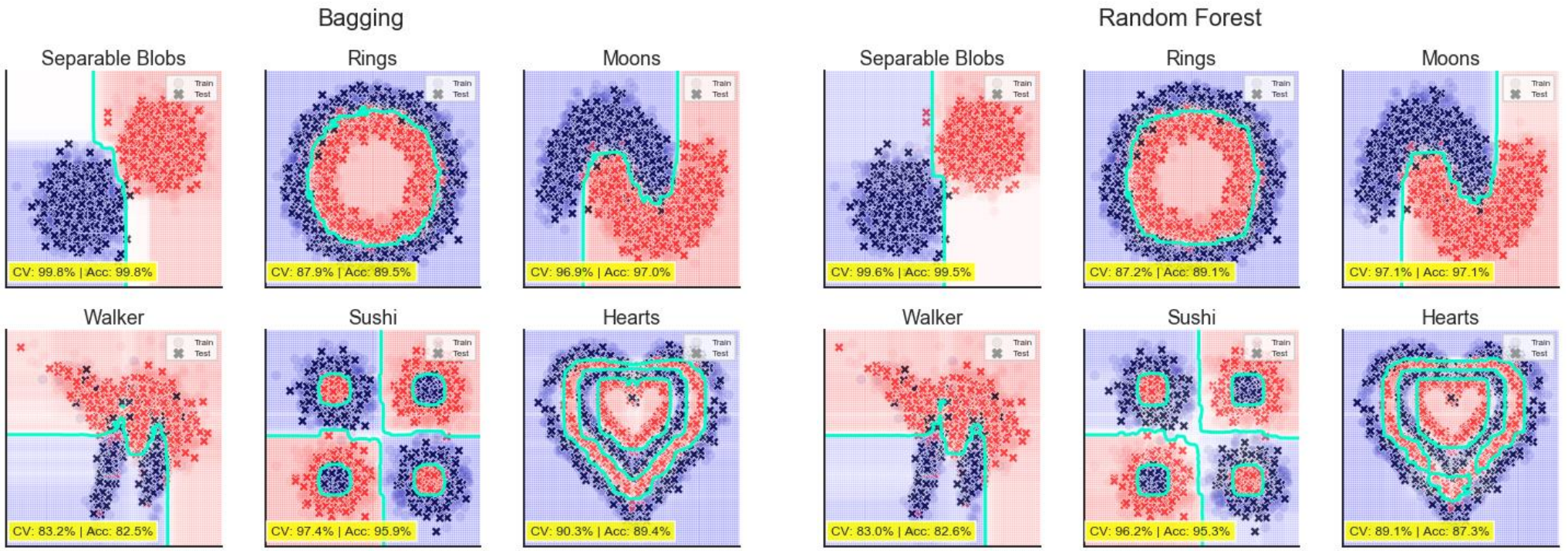
LDA and Logistic Regression



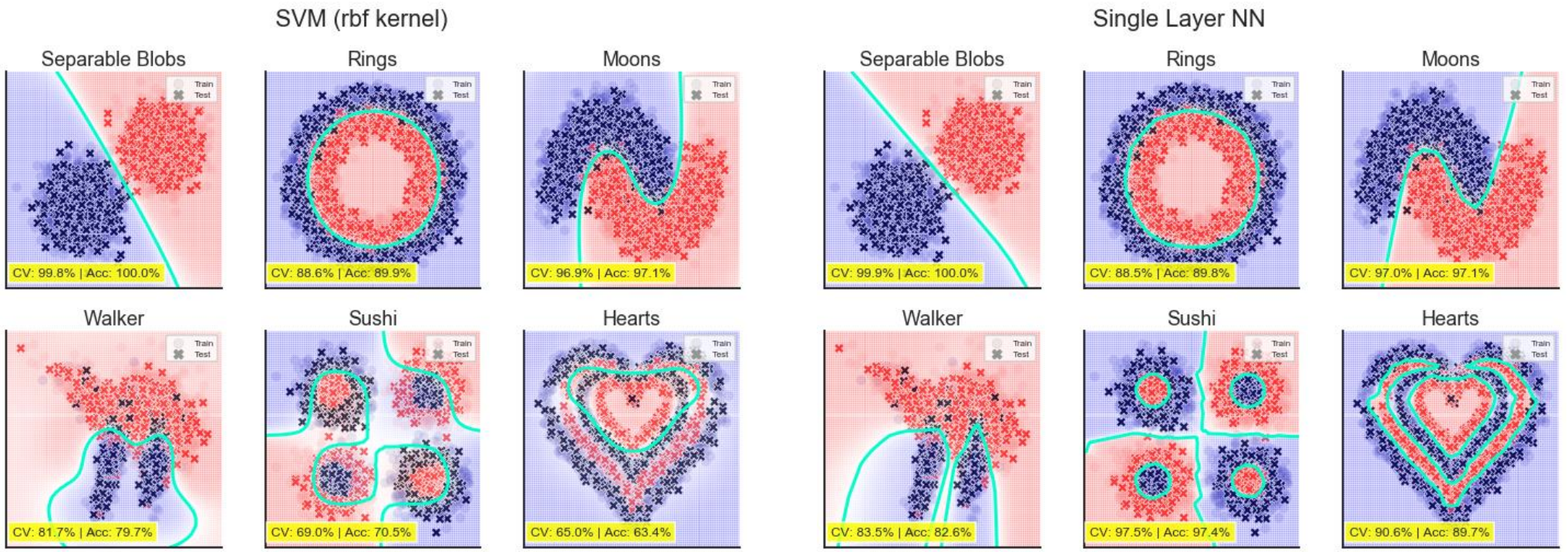
Decision Trees and Regularization



Ensemble Methods

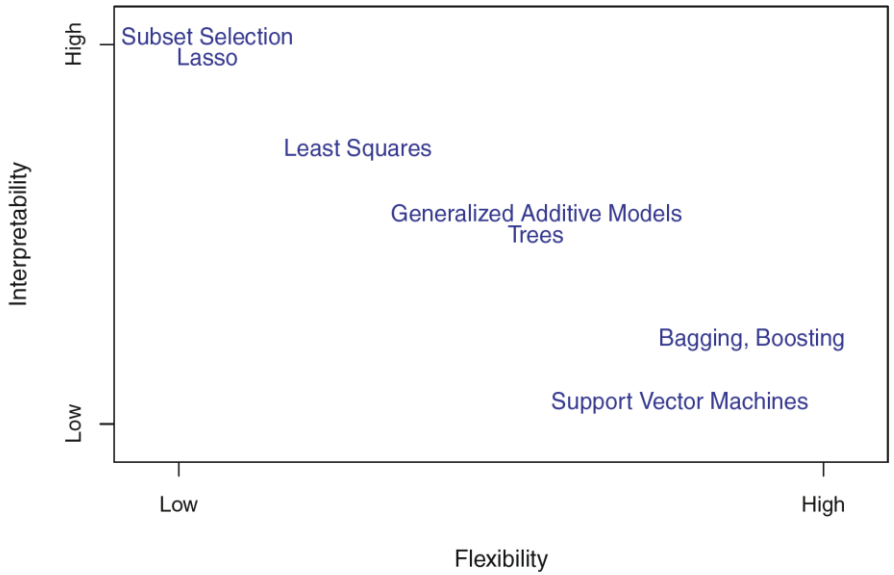


SVM and Neural Nets



Flexibility, Complexity, and Interpretability

- More flexible models can perform better than simpler models, but they also tend to be more complex.
  - A more complex model has more parameters to fit.
  - Requires more data to train.
- Why choose a more restrictive method instead of a more flexible approach?
  - Inference affords interpretability.
  - Avoid overfitting.
  - Detect, correct, or avoid bias.
  - Trust, relinquish control, explainability.



Interpretability and Flexibility of Different Models (James et al, 2013)

Assessing Model Accuracy

Assessing Model Accuracy

- Prediction
- Measuring Quality of Fit
- An Optimization Problem
- Overfitting the Data
- Variance-Bias Trade-Off

Prediction

- We believe that some process exists that relates the variables *X* with some output value *Y*.
- We think that the model is $Y = f(X) + \epsilon$ but we do not know the function $f()$.
- Prediction problem: given some data about some variables *X*, find a function *f()* that relates a quantitative value of interest *Y*.
- The result is a 'model', $\hat{f}()$ that has been *trained* on *X*.
- A prediction is the result of applying the trained model to a data set: $\hat{Y} = \hat{f}(X)$.

Measuring Quality of Fit

- No free lunch: there is no single method that dominates others in all circumstances.
- We need some way to measure how well a model's predictions actually match the observed data.
- Quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

- In regression analysis, we usually calculate mean squared error (MSE):

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

- A typical result requires MSE to be calculated on the training data. We call this result the *training* MSE.
- We are more interested in the accuracy of predictions on data that the model has not yet seen, the *test* data.

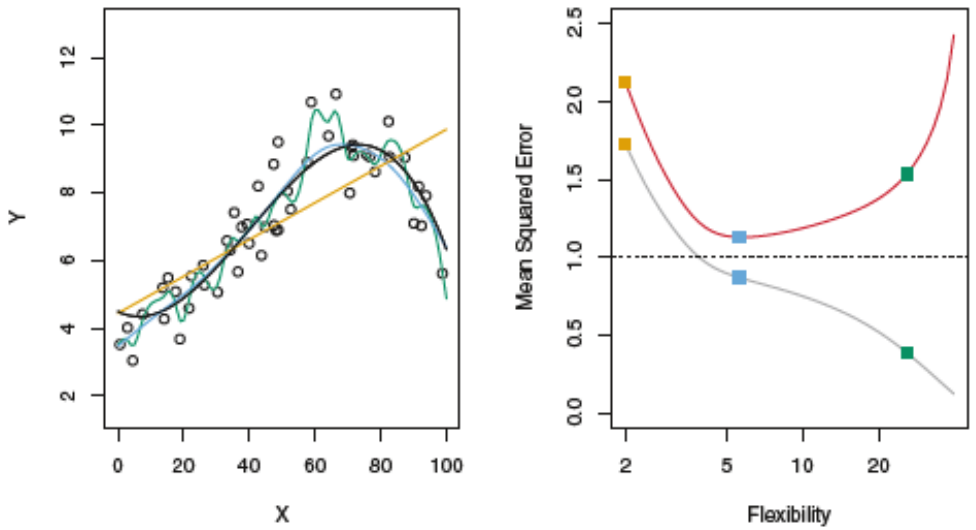
(James et al, 2013)

An Optimization Problem

- Represent unseen observations with (x_0, y_0).
- We are looking for a method that gives the lowest *test* MSE, as opposed to the lowest *training* MSE.
- If we had a large number of (x_0, y_0) we could caluclate the average squared prediction error for these test observations:

$$Ave(y_0 - \hat{f}(x_0))^2$$

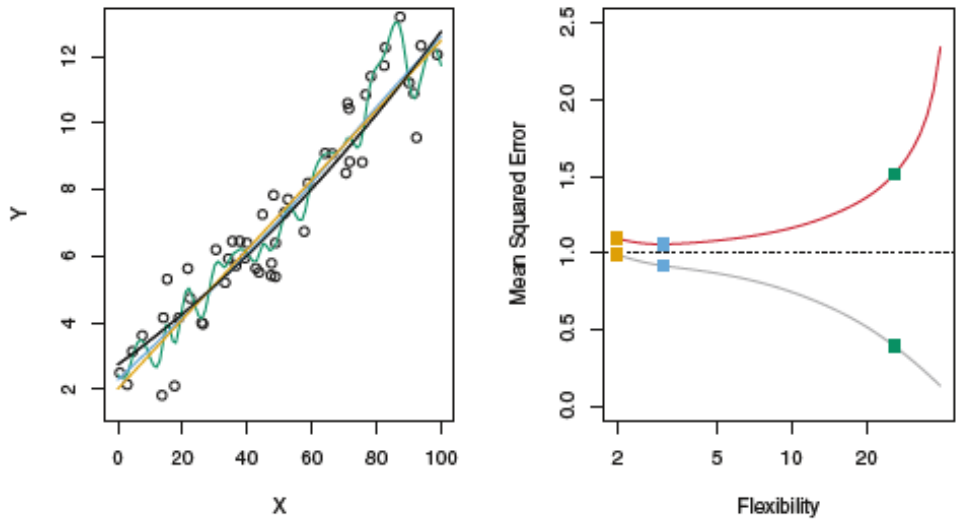
- Sometimes, we may have access to test data, but what happens if we do not?
- There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.
- As the flexibility of the statistical method increases:
  - Training MSE: decreases
  - Test MSE: U-shape



Left: Data simulated from f (black). Estimates: linear regression (orange), two smoothing splines (blue and green). Right: Training MSE(grey), test MSE (red). (James et al, 2013)

Overfitting the Data

- With flexibility increase: training MSE decreases and U-shape test MSE.
  - This is a fundamental property of statistical learning.
- When a method gives small training MSE and large test MSE, we are *overfitting* the data.
  - Overfitting happens when the model works too hard in learning the properties of the data and picks up random noise.
  - One important method to reduce overfitting is *cross-validation*.



The function f in this case is closer to linear. (James et al, 2013)

Variance-Bias Trade-Off

- The U-Shape observed in the test MSE curves results from two competing properties of statistical learning methods.
- Expected test MSE is the sum of three fundamental quantities:
  - Variance of $\hat{f}(x_0)$
  - Square bias of $\hat{f}(x_0)$
  - Variance of the error terms ϵ
- More formally,

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

Bias and Variance

- **Bias** is a model's tendency to learn the same wrong thing consistently.
  - Bias is the error that is introduced by approximating a real-life problem, which is extremely complicated, by a much simpler model.
  - Generally, more flexible models result in less bias.
- **Variance** is the tendency to learn random things irrespective of the real signal.
  - Variance is the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
  - Generally, more flexible models result in higher variance.
  - Ideally, $\hat{f}$ should not vary much between training sets. If it does, small changes in data can result large changes in $\hat{f}$.

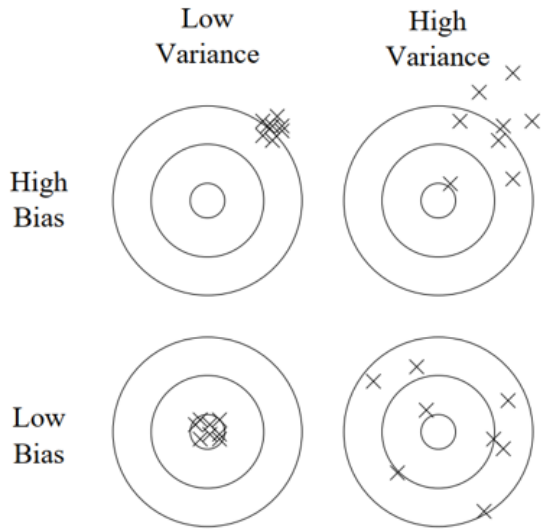


Figure 1: Bias and variance in dart-throwing.

(Domingos, 2012)

References

References

- Agrawal, Ashvin, Rony Chatterjee, Carlo Curino, Avrilia Floratou, Neha Gowdal, Matteo Interlandi, Alekh Jindal,Konstantinos Karanasos, Subru Krishnan, Brian Kroth, Jyoti Leeka, Kwanghyun Park, Hiren Patel, Olga Poppe,Fotis Psallidas, Raghu Ramakrishnan, Abhishek Roy, Karla Saur, Rathijit Sen, Markus Weimer, Travis Wright, Yiwen Zhu. *Cloudy with High Chance of DBMS: A 10-year Prediction for Enterprise-Grade ML*. 2020. 10th Annual Conference on Innovative Data Systems Research (CIDR '20). January 12-15, 2020, Amsterdam, Netherlands.
- Ayers, Ian. *Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart*. US: Bantam, 2008.
- Domingos, Pedro. *A few useful things to know about machine learning*. Communications of the ACM. October 2012. https://doi.org/10.1145/2347736.2347755
- Financial Stability Board (FSB). Artificial intelligence and machine learning in financial services. 2017, available at https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service/
- James, Garreth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* US: Springer, 2013.
- Khun, Max and Julia Silge. *Tidy Modeling with R*. Version 0.0.1.9008 (2021-01-19)
- Mitchell, Tom. *Machine Learning*. US: McGraw-Hill, 1997.
- Wikipedia contributors. (2020, December 30). *Predictive analytics*. In Wikipedia, The Free Encyclopedia. Retrieved 03:27, January 20, 2021, from https://en.wikipedia.org/w/index.php?title=Predictive_analytics&oldid=997309866