Review of Statistical Learning

FA690 Machine Learning in Finance

Dr. Zonghao Yang

2025 Spring

Statistical Learning Problems

- Outcome measurement Y
  - Dependent variable, response, target
- Vector of p predictor measurements $X = (X_1, X_2, \dots, X_p)$
  - Independent variables, inputs, regressors, covariates, features
- In the very general form, the relationship between Y and $X = (X_1, X_2, \dots, X_p)$ can be written as

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

  where $f(\cdot)$ is some fixed but unknown function of $X_1, X_2, \dots, X_p$, and ϵ is a random error term
- Objective of statistical learning: Based on observations $(x_1, y_1), \dots, (x_n, y_n)$, we want to learn the function $f(\cdot)$ so that
  - Accurately predict unseen cases
  - Understand which inputs affect the outcomes, and how
  - Assess the quality of our predictions and inferences

SEMMA: Sample, Explore, Modify, Model, Assess

A framework to analytical modeling

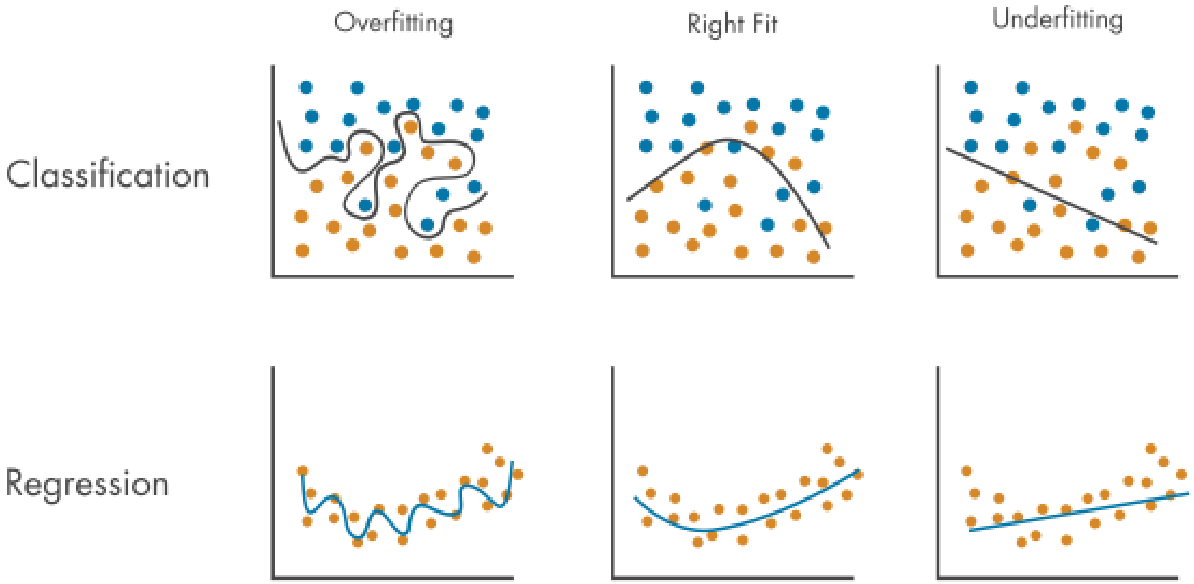
SEMMA: Sample, Explore, Modify, Model, Assess

A framework to analytical modeling

- **S**ample: Take a sample from the dataset; partition into training, validation, and test datasets
- **E**xplore: Examine the dataset statistically and graphically
- **M**odify: Transform the variables and impute missing values
- **M**odel: Fit predictive models (e.g., regression tree, neural network)
- **A**ssess: Compare models using a validation dataset

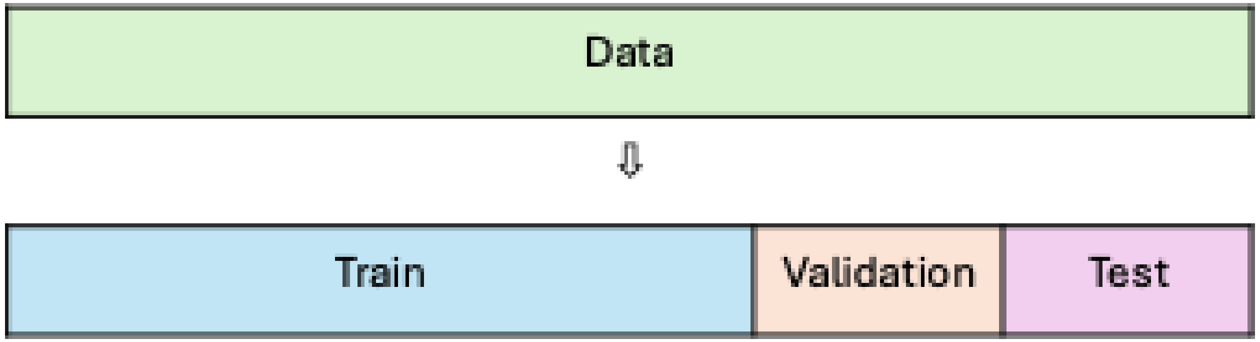
SEMMA: Sample, Explore, Modify, Model, Assess

- Randomly sample data into train-test, or train-validation-test set
- Training and validation data are in-sample; Test data is out-of-sample
- Diagnose and avoid overfitting



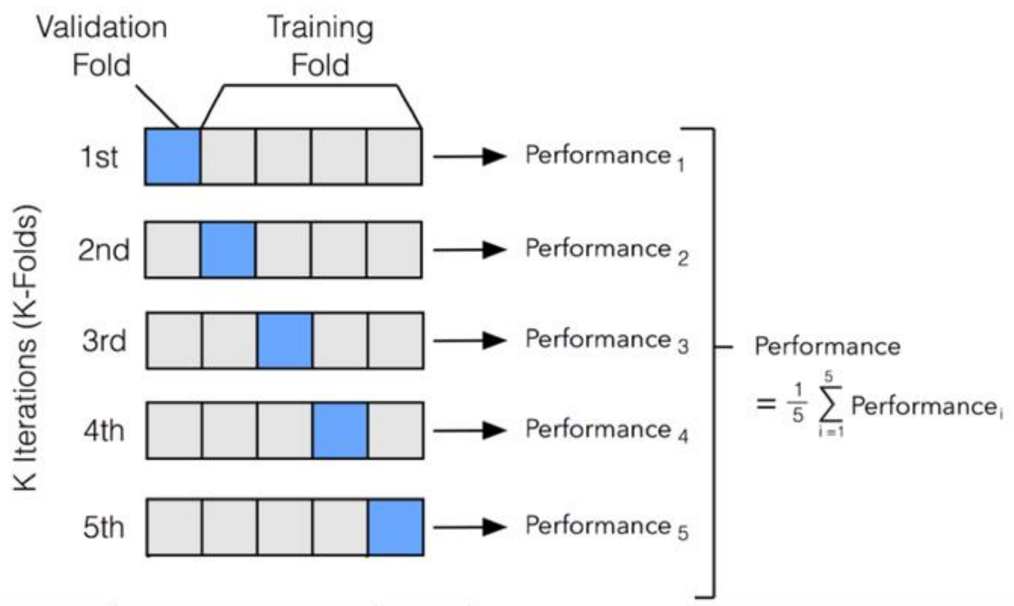
Random Split

- Randomly divide the available set of observations into three parts: A training set, a validation set, and a test set



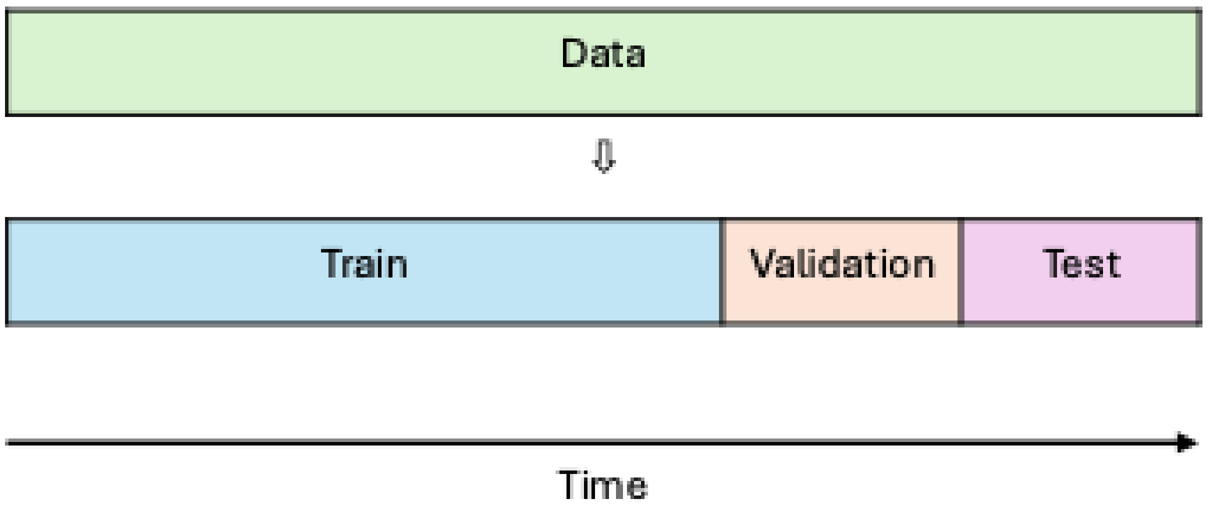
K-Fold Cross-Validation

- Rather than just keeping one part of the training data for validation, we can split the data into *k* folds
- We then train the model *k* times, each time by leaving out one of the folds for validation
- The final performance is the average performance across all *k* validation folds

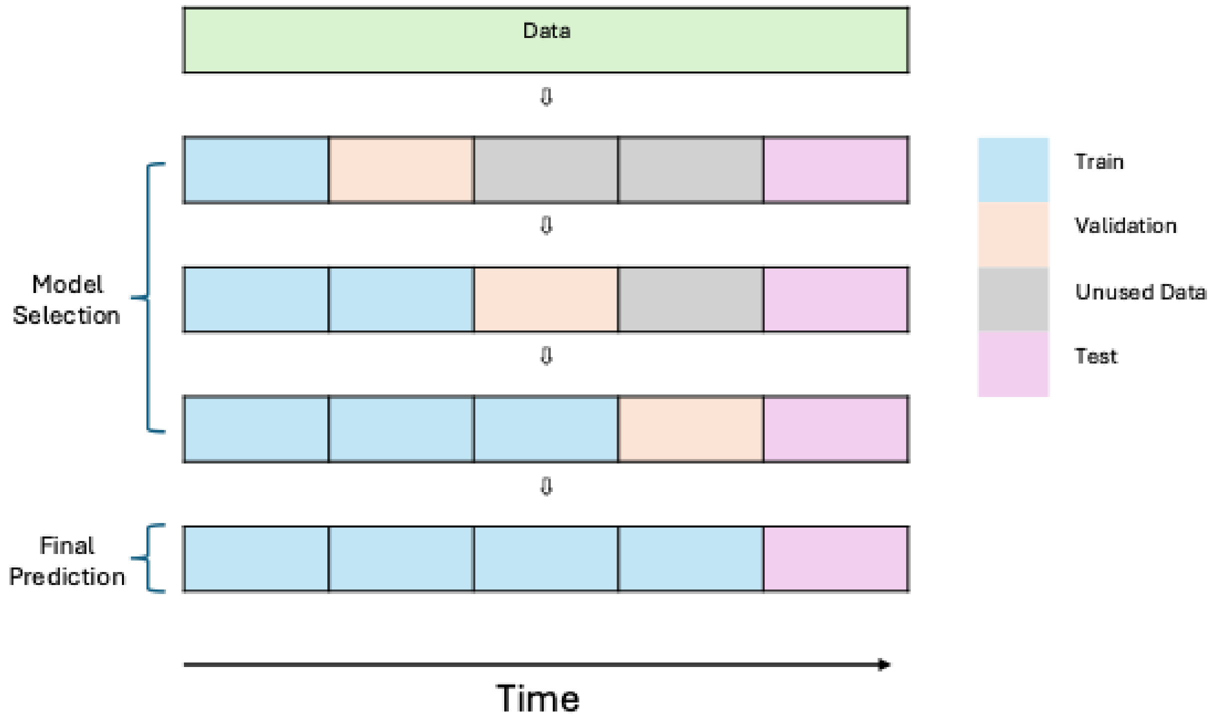


Simple Time Split Validation

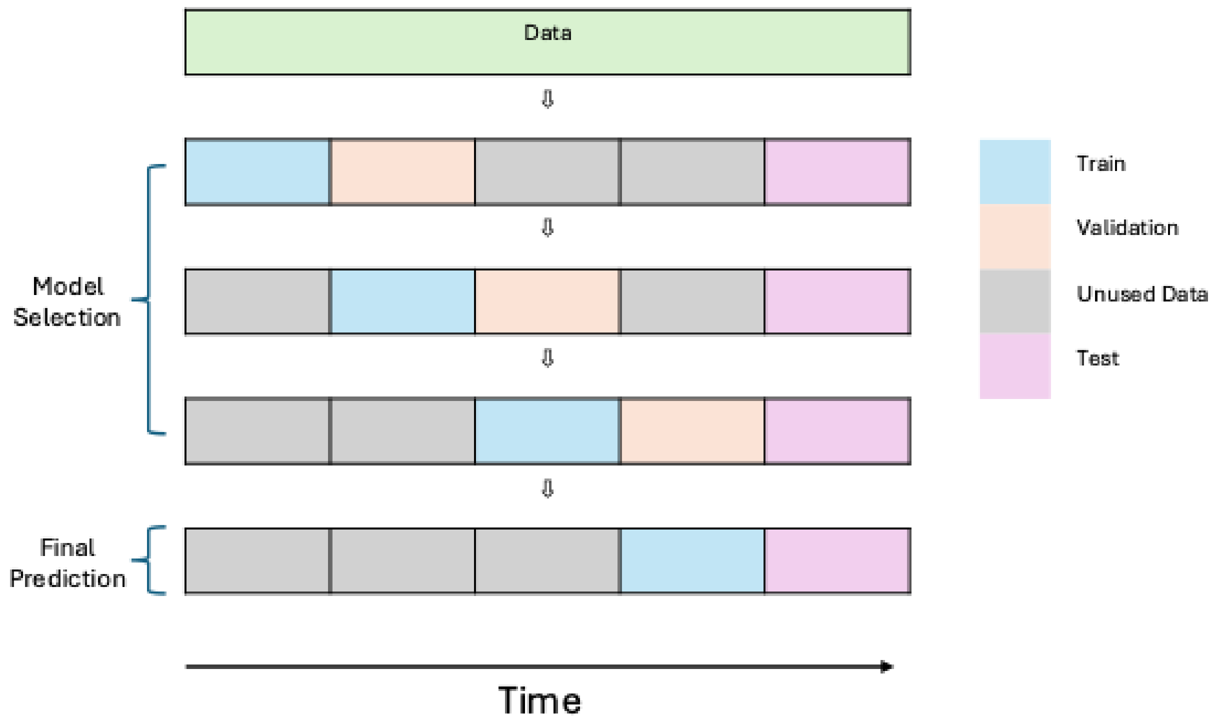
Time series data is a common type of financial data. For example, stock returns and the loan defaults.



Expanding Window



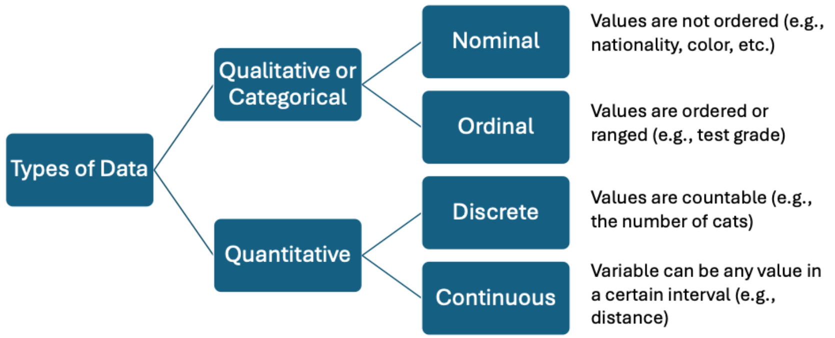
Rolling Window



SEMMA: Sample, Explore, Modify, Model, Assess

There are different ways to segment data

- Qualitative and categorical data



- Structured and unstructured data
  - Structured data: Tabular data
  - Unstructured data: Text, images, audio

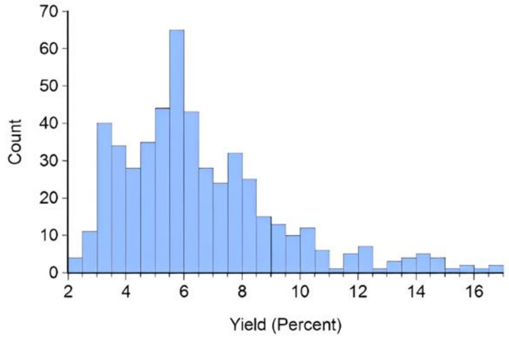
Graphics

- Univariate data
  - Histograms and density estimates can help learn about distributional shape: symmetric, skewed, fat-tailed, etc.
  - Time series plots reveal dynamics such as trend, seasonality, cycle, outliers, . . .
- Multivariate data
  - Scatterplots for relations: Does a relation exist? Is it linear or non-linear? Are there outliers?

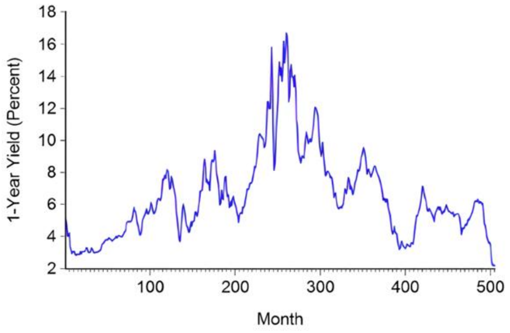
Graphics

Example: One-year government bond yield

- **Histogram** reveals distributional shape



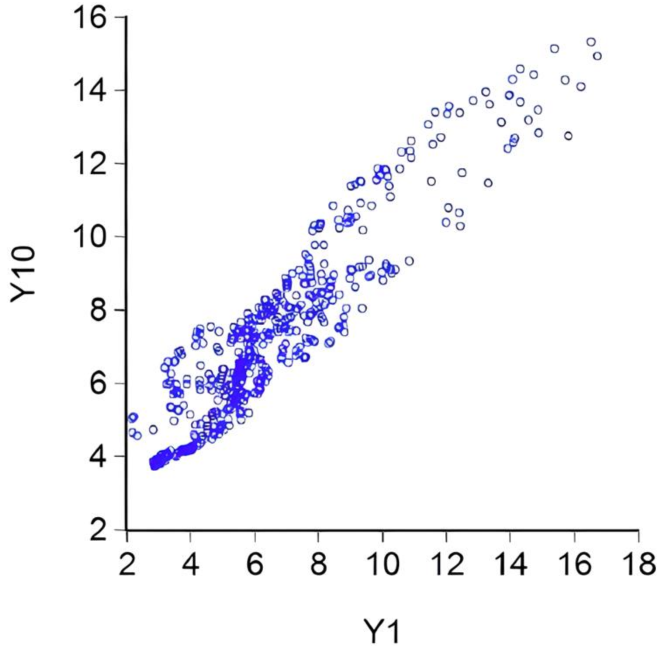
- **Time series plot** reveals dynamics



Graphics

Example: One-year and 10-year government bond yields

- Scatterplot reveals relation between two variables



SEMMA: Sample, Explore, **Modify**, Model, Assess

Missing Values: The absence of data or a lack of recorded information in a dataset

- If the number of datapoints with missing values is small, those datapoints might be omitted
- If the number of missing entries is large, you need to consider removing the feature
- Can replace the missing value with an imputed value, based on the other values for that variable across all data points
  - Mean/median
  - Maximum/minimum
  - 0
- Scenario 1: “The employment length” is missing
- Scenario 2: “The number of months of delinquencies” is missing. According to the data description, these are the borrowers who haven’t delinquent on any loans before.

Feature Scaling

Feature scaling transforms data to the same scale

- Standardization

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

- Normalization

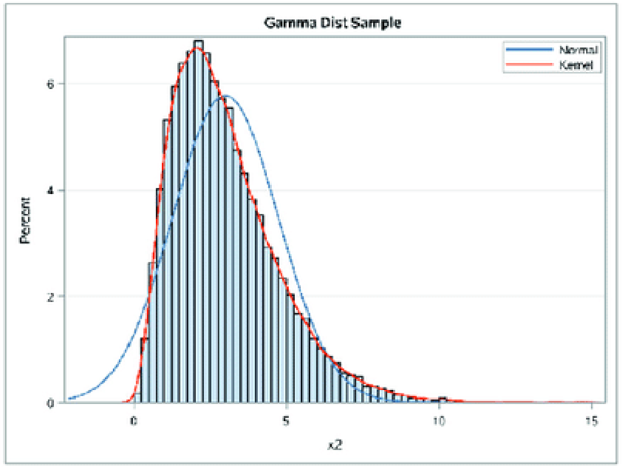
$$\tilde{X} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Almost all algorithms are sensitive to feature scales
  - Tree-based methods are less sensitive to feature scaling

Feature Transformation

- Some data covers various orders of magnitude (e.g., wealth, salary)
- Standardization or normalization can be misleading and will be skewed towards the larger values
- Often, one first applies *log-transform* (and then standardization)

$$\tilde{X} = \log(X)\,, X > 0$$



(a)

(b)

SEMMA: Sample, Explore, Modify, Model, Assess

$$y = f(X) + \epsilon$$

- Supervised learning
  - Labeled outcome variable (Y)
  - Regression: Y is quantitative (e.g., sales)
  - Classification: Y takes values in a finite, unordered set (e.g., spam/not spam email, handwritten digits 0 – 9)
- Unsupervised learning
  - No outcome variable (Y), just a set of predictors (X)
  - Example: k-means clustering algorithm
- Parametric and non-parametric methods
  - Depend on whether a specific functional form for $f(X)$ is specified

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Optimization objective

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p = \underset{\beta_0, \ldots, \beta_p}{\operatorname{argmin}} \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2$$

- Fitted hyperplane

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- Prediction given $X = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}$$

Regularization

- Ridge regression (or L2 regularization)

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p = \underset{\beta_1,\beta_2,\dots,\beta_p}{\operatorname{argmin}} \sum_{i=0}^{n} (y_0 - \hat{y}_i)^2 + \lambda \sum_{j}^{p} \beta_j^2$$

- Lasso regression (or L1 regularization)

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p = \underset{\beta_1,\beta_2,\dots,\beta_p}{\operatorname{argmin}} \sum_{i=0}^{n} (y_0 - \hat{y}_i)^2 + \lambda \sum_{j}^{p} |\beta_j|$$

- Elastic net: A linear combination of the Ridge and Lasso regularization techniques

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p = \underset{\beta_1,\beta_2,\dots,\beta_p}{\operatorname{argmin}} \sum_{i=0}^{n} (y_0 - \hat{y}_i)^2 + \lambda_1 \sum_{j}^{p} \beta_j^2 + \lambda_2 \sum_{j}^{p} |\beta_j|$$

Logistic Regression

- Logistic regression uses sigmoid functions as the functional form

$$P(X) = \Pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

- Binary predictor: Given a decision threshold $\rho \in (0,1]$

$$f(x) = \begin{cases} 0, & P(x) < \rho \\ 1, & P(x) \geq \rho \end{cases}$$

- Parameter estimation: Given a data sample S

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, l(\beta|S) = \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i)\log(1 - p(x_i))$$

Support Vector Machine

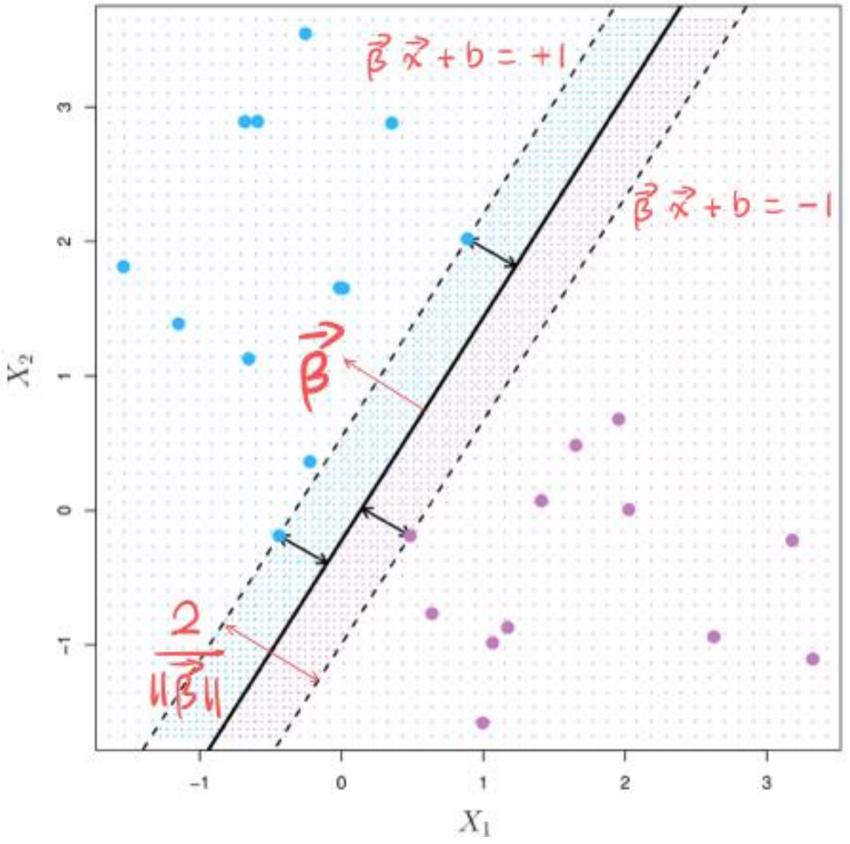
- Assume the data sample S is linearly separable by some margin
- Hyperplane

$$g(x) = \beta x + b = 0$$

- Linear classifier

$$f(x) = \begin{cases} 0, & g(x) < 0 \\ 1, & g(x) \geq 0 \end{cases}$$

- SVM: Find two parallel hyperplanes that correctly classify all the points and maximize the distance between them



SVM Formulation

- SVM as an optimization problem

  maximize the distance: $\frac{2}{|\beta|}$

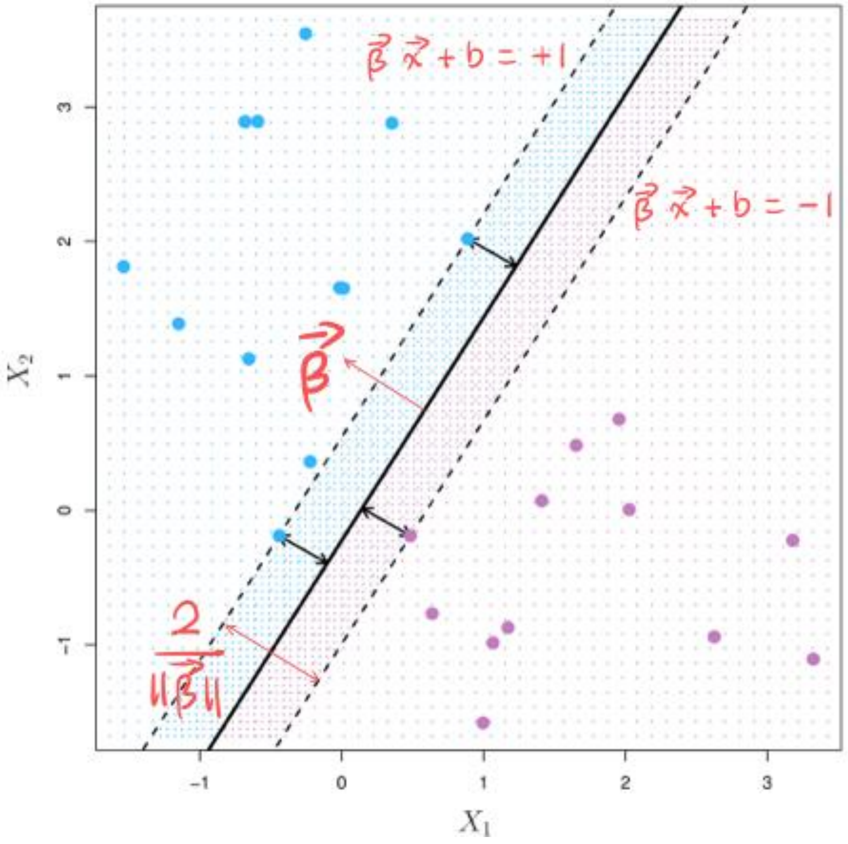
  such that: $y_i(\beta x_i + b) \geq +1$ (for all i)

- Equivalently:

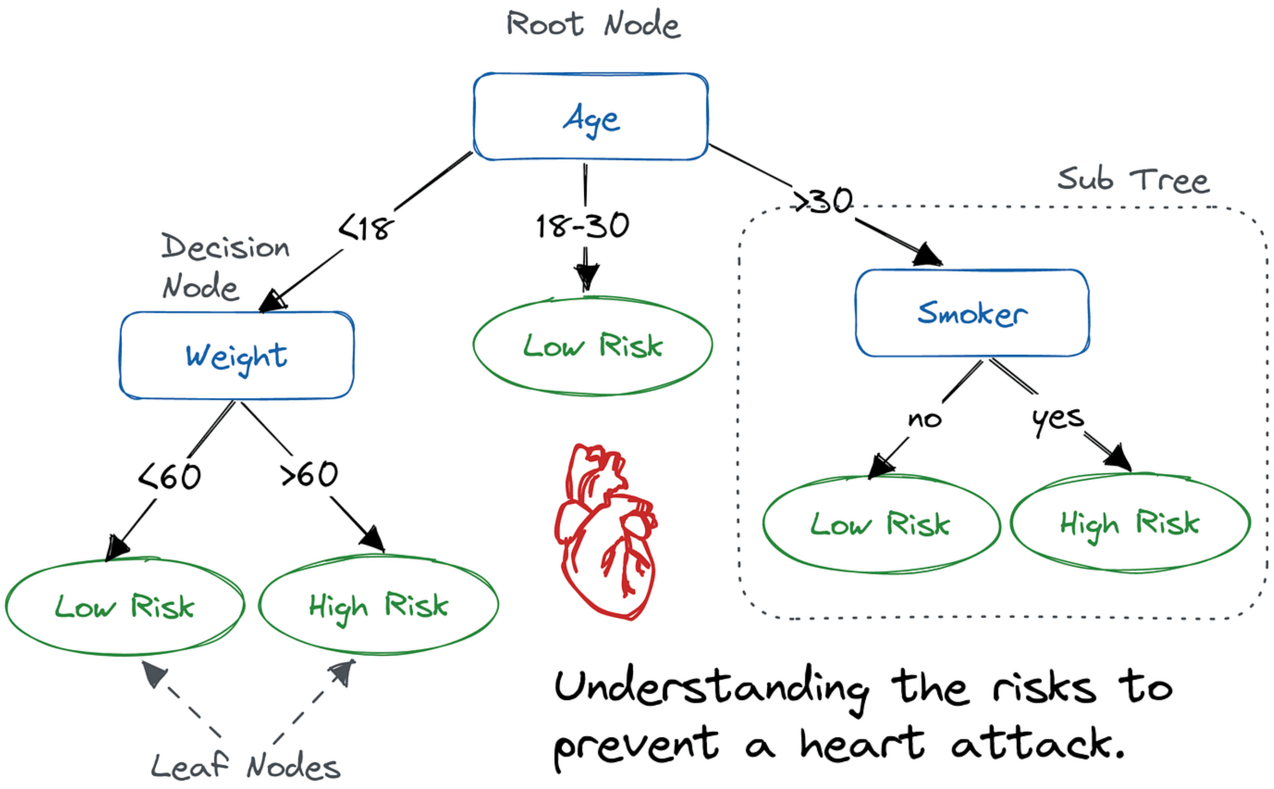
  maximize the distance: $\frac{1}{2}|\beta|^2$

  such that: $y_i(\beta x_i + b) \geq +1$ (for all i)

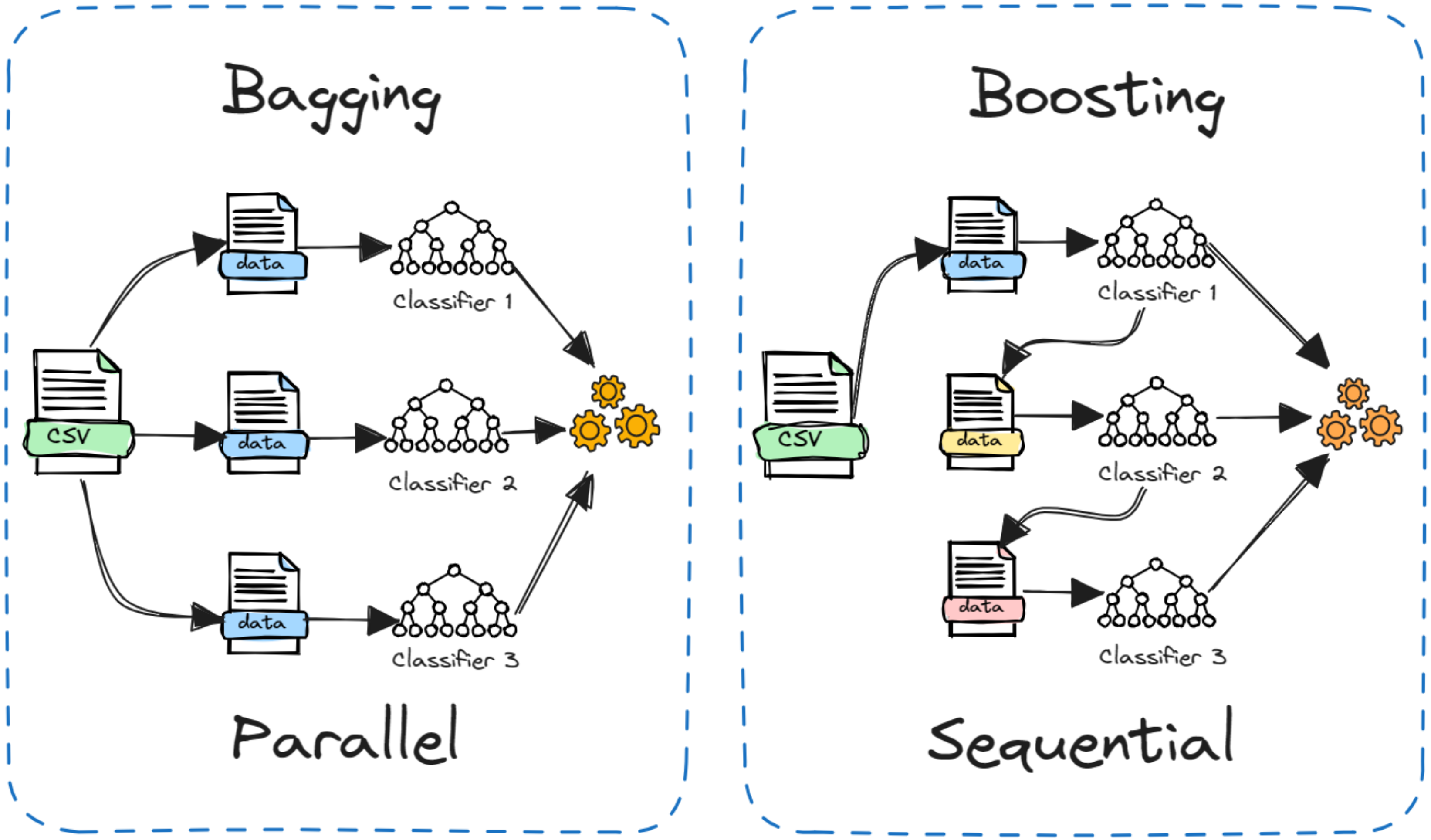
- The constrained optimization problem can be rephrased as a convex quadratic program, and solved efficiently



Decision Trees

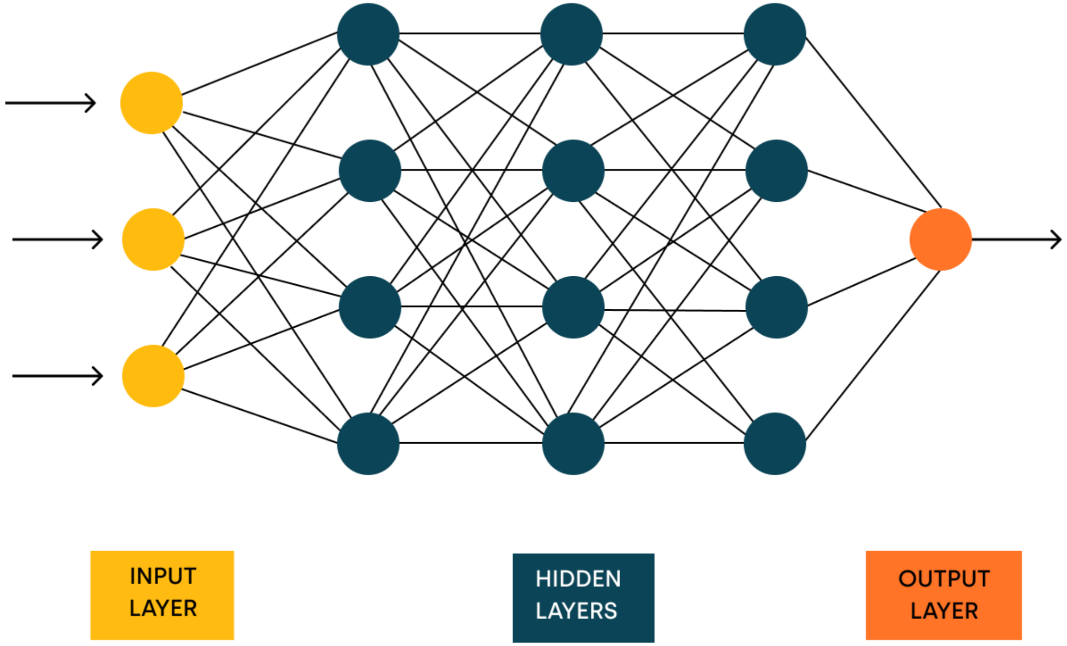


Bagging and Boosting



Neural Networks

- Input layer, hidden layers, output layer
- Neurons: Input neuron, hidden neuron, output neuron
- Activation function:
  - ReLU: $\max(0, x)$
  - Sigmoid: $\frac{1}{1+e^{-x}}$
  - tanh: $\tanh(x)$



K-Means Clustering Algorithm

1. Start with K initial clusters (user chooses K)
2. At every step, each data point is reassigned to the cluster with the “closest” centroid
3. Recompute the centroids of clusters that lost or gained a data point, and repeat Step 2
4. Stop when moving any more data points between clusters increases cluster dispersion

[Visualizing the K-means algorithm][Visualizing DBSCAN Algorithm]

SEMMA: Sample, Explore, Modify, Model, Assess

Evaluation metrics for Regression

- Mean squared error (MSE)

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n$$

- R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^{n} e_i^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Evaluation Metric for Classification

Confusion Matrix

- Confusion matrix is a table that summarizes the performance of classification

		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- Misclassification occurs when the model assigns an incorrect class to the data

Evaluation Metric for Classification

Accuracy

- The most common and intuitive metric is the accuracy score

$$Accuracy = \frac{\sum TP + TN}{\sum TP + FP + FN + TN}$$

- Accuracy score alone is not a good metric when the label imbalance exists
  - For example, if only 10% of the loans are charged off, a naive classifier that predicts all loans will not default has an accuracy of 90%

Evaluation Metric for Classification

Example: Confusion Matrix in Loan Default Prediction

- False positive: Fully paid loans that are predicted to be charged off
- False negative: Defaulted loans that are predicted to be fully paid
- False negative in unsecured loans is costly for banks
- What about cancer detection? Both FN and FP are costly

		Actual Label	
		Default	Fully Paid
Predicted Label	Default	True Positive (TP)	False Positive (FP)
	Fully Paid	False Negative (FN)	True Negative (TN)

Evaluation Metric for Classification

Precision, Recall, and F1 Score

- Precision and recall indicate the occurrence of false positives and false negatives, respectively

$$Precision = \frac{\sum TP}{\sum TP + FP}$$

$$Recall = \frac{\sum TP}{\sum TP + FN}$$

- F1 score: Weighted average (Harmonic mean) of precision and recall

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Evaluation Metric for Classification

Classification Threshold

- Making a classification

$$f(x) = \begin{cases} 0, & P(x) < \rho \\ 1, & P(x) \geq \rho \end{cases}$$

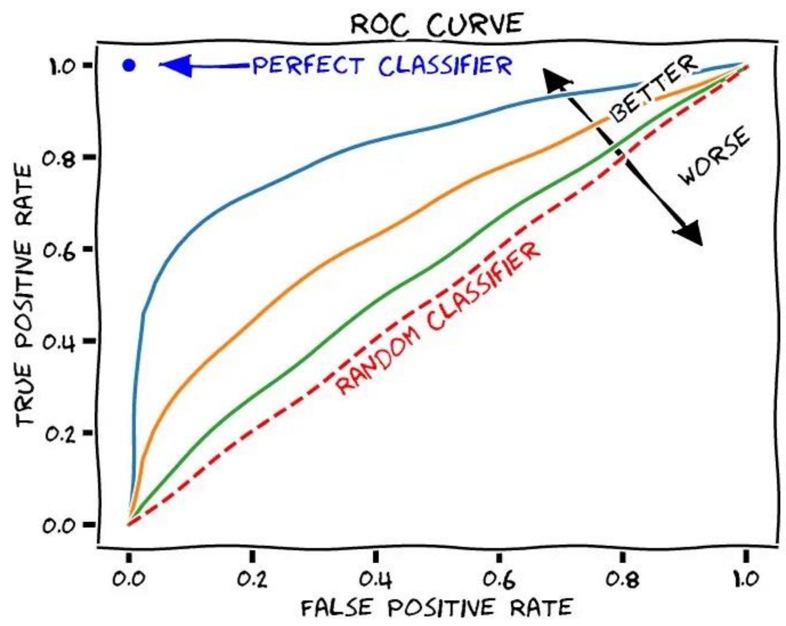
- The threshold (ρ) is a hyperparameter
- As ρ changes between 0 and 1, the evaluation scores also change
  - False-positive rate: $FPR = \frac{\sum FP}{\sum TN+FP}$
  - True-positive rate: $TPR = \frac{\sum TP}{\sum TP+FN}$

Evaluation Metric for Classification

Receiver Operating Characteristics (ROC) Curve

$$\rho \in [0,1], FPR = \frac{\sum FP}{\sum TN + FP}, TPR = \frac{\sum TP}{\sum TP + FN}$$

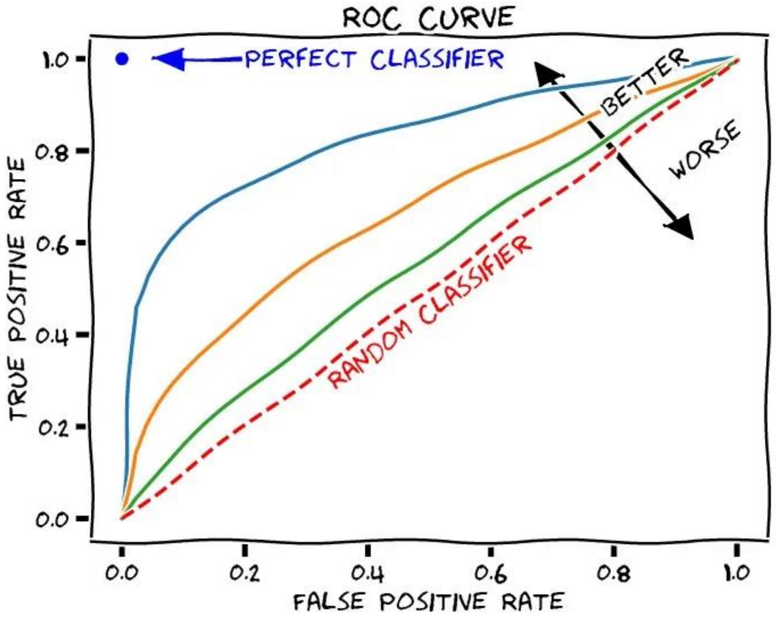
- If $\rho = 0$, then $FPR = 1$ and $TPR = 1$
- If $\rho = 1$, then FPR =0 and TPR =0
- For $\rho \in (0,1)$, there is a trade-off between *FPR* and *TPR*



Evaluation Metric for Classification

AUC Scores

- Area Under the Curve (AUC) Score
  - $AUC = 1$: Perfect classification
  - $AUC = 0.5$: Random guess
  - Random guess is a naïve but often very useful benchmark
- Threshold independent: AUC measures how well the model can distinguish between the classes across all possible classification thresholds



SEMMA: Sample, Explore, Modify, Model, Assess

A framework to analytical modeling

- **S**ample: Take a sample from the dataset; partition into training, validation, and test datasets
- **E**xplore: Examine the dataset statistically and graphically
- **M**odify: Transform the variables and impute missing values
- **M**odel: Fit predictive models (e.g., regression tree, neural network)
- **A**ssess: Compare models using a validation dataset

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK **YOU**

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030