Reading 25 Machine Learning Methods P1

Basics

Model selection in ML is decided by data (not by analyst).

- **Supervised Methods**: Uses labeled data which predefined independent and dependent variable**. *Can be used for prediction.***
- **Unsupervised Methods**: Recognizes patterns in data with no target variable. Not used for prediction but used to learn structure of data.
- **Reinforcement Learning**: Based on trail and error approach in dynamic environment.

Data Preparation and cleaning

Preparation

Scaling is done to make data consistent.

- **Standardization:** Uses mean and SD to scale data.

$$X_{Scaled} = \frac{X_{ij} - \mu}{\sigma_i}$$

- **Normalization:** uses min max transformation to scale data. Outcome is between zero and one.

$$X_{scaled} = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}}$$

Cleaning: : Includes, fixing missing data (by removing observations, replacing it), outliers, duplication, inconsistent values.

Principal Components Analysis

PCA is used for dimension(feature) reduction by keeping only unrelated variables offering maximum information.

K Means Clustering

Unsupervised K means algo used to create cluster of data. K is number of clusters. *Centroids: Center of data cluster.* The objective of k means algo is to minimize distance between data point and centroid. The distance is calculated using

- **Euclidean distance**: diagonal line measure.
- **Manhattan distance**: Uses path of opposite side and adjacent side.

Inertia is measure of total distance between centroid and data points in a cluster. Lower inertia is better fit. *Inertia falls with increasing cluster K hence optimal number is achieved for K where **inertial falls at slower rate (also known as Elbow)** as more K is added* This is done using scree plot.. Alternatively, **Silhoutte coefficient** can be used to choose K. This method compares the distance between the own and next close cluster. Highest coefficient gives optimal K.

Underfitting and Overfitting

Overfitting occurs when model is too complex. Model performs poorly (high error rate) in new data compared to training data. This is major problem for ML model.

Underfitting: when model is too simple it performs poorly in training data. Traditional models are more prone to this issue. ML resolves this issue (compared to traditional models).

Bias variance tradeoff:

Model	Bias error (in training data)	Variance Error (in test data)
Simple	Increases	Decreases
Complex	Decreases	Increases

Model complexity is optimal when total error (combined bias and variance) is minimum.

Data Sub Samples

- Training set: subset of data used to train ML model.
- Validation set: subset of data used to choose model.
- Test Set: subset of data Used to test the model performance.

For large data set the split between subset is not an issue.

K fold cross validation: This method is used when data set is small. In this approach only two subsets 1) training/ validation and 2) Testing are used. For 1st subset, assuming k is 5, data is divided into 5 chunks. Then iteratively, model is trained on k – 1 sets (i.e. leaving one set at a time) and leftover set is used for validation score for each iteration. For final validation performance, scores from each iterations are averaged.

Leave one out cross validation: is method when k = n (total observations in sample set). Rest is same.

Reinforcement Learning (RL)

Is decision making algo with goal of maximize reward via trial-and-error decision-making. Applications in finance include derivatives hedging, trading rules, and large-volume trade management. Requires massive training data; initial performance is weak but improves over time.

Key areas:

- States (S): The environment.
- Actions (A): decisions taken.
- Rewards (R): Feedback maximized when the best decision is made.
- Q-value: Expected value of taking action (A) in state (S). Optimal action maximizes the Q-value.

Exploration vs. Exploitation:

• Exploitation (probability p): Choosing the best action already identified.

• Exploration (probability 1 - p): Trying a new action.

As trials increase and learning improves, p increases (more exploitation).

Evaluation of action:

• Monte Carlo method: Evaluates actions based on total subsequent rewards resulting from an episode. Updates current Q-value by adding alpha X (reward – current Q value)

• Temporal difference learning: Alternative that assumes the current best strategy going forward and only looks one decision ahead. Updates current Q-value by adding alpha X (immediate reward + next state value – current Q value)

Deep Reinforcement Learning:

Occurs when neural networks are used to estimate the complete Q-value table from available observations.

R25 Machine Learning Methods P2

NLP Natural Language Processing

Also known as text mining. Used for processing of human language (written/verbal). Financial/accounting applications include fraud detection, corporate sentiment analysis, text categorization, and intent recognition. Benefits: speed and removing human inconsistency or bias.

Text Preprocessing Steps:

1.**Tokenization**: Removing punctuation, symbols, and spacing, and converting all characters to lowercase.
2.**Stopwords removal**: Eliminating filler words (e.g., "the", "has", "a") that carry no information value.
3.**Stemming**: Replacing words with their base linguistic stems (e.g., "arguing" and "argues" map to "augu"). The stem may not be an actual word.
4.**Lemmatization**: Replacing words with their morphological lemmas (e.g., "worse" maps to "bad"). Similar to stemming, but the lemma is always an actual word.
5.**N-grams**: Grouping adjacent words that carry distinct meaning together rather than individually (e.g., trigram "exceed analyst expectations").

Bag of Words & Sentiment Analysis:

•**Bag of Words**: Final processed text state where word order and grammatical linkages (outside of N-grams) are ignored.
•**Sentiment Analysis:** Scoring text as positive, negative, or neutral by counting pre-classified sentiment word stems. Higher count determines overall sentiment (e.g., 4 positive vs. 6 negative = negative sentiment).
•**Limitation**: Algorithms struggle with positive words used in negative contexts (e.g., "would have resulted in an increase"). Addressed by comparing algorithm outputs against human reviews to refine accuracy.

R26 Machine Learning Predictions P1

Encoding

Transformation of non-numerical (qualitative) data into numbers, required for machine learning or regression models.

Encoding Methods:

- **One-hot encoding**: Used for categorical data with no natural order (nominal data). A separate dummy variable is set up for each category. The observation receives a 1 for its specific category and a 0 for all others.
- **Ordinal encoding**: Used for categorical data that has a natural order. Dummy variables take on sequential numerical values to reflect this hierarchy.

Regularization

Process of simplifying models to reduce overfitting. Shrinks coefficient estimates to substantially reduce model variance on test data. Done by adding a penalty term to the loss function (L).
Techniques:

Ridge Regression (L2):

• Shrinks the magnitude of slope (beta) coefficients closer to zero. Uses analytical approach. Penalty term = Lambda X sum of beta square .

LASSO (L1):

• Sets less important beta exactly to zero, making it a feature selection approach. Penalty term = lambda X sum of |Beta|. Solved using a numerical approach. Higher lambda > simpler model
Hyperparameter (lambda): The tuning parameter used to select the optimal model. It controls the severity of the penalty but is not part of the final model itself.

Logit

Used for binary outcomes in finance like default vs. no default. Cumulative logistic function (sigmoid curve) transforms outputs into a probability range between 0 and 1.
Estimation & Mechanics:

- •Maximum Likelihood Method: Used instead of Ordinary Least Squares (OLS) because the logit model is non-linear.
- Log-Likelihood Function may be used. It is preferred over the standard likelihood function.

Classification & Thresholds (Z):
Once probabilities are estimated, observations are classified based on a probability threshold Z. Z Depends on the asymmetric costs of misclassification. Z = 0.5 implies cost of being wrong or right is equal.
Model Evaluation Metrics: If evaluating continuous financial outputs (like rates of return) rather than binary classifications: Mean Squared Forecast Error (MSFE) or Mean Absolute Forecast Error (MAFE) are used.

Support Vector Machines SVM

Supervised ML models that excel when working with a high number of features. The goal is to find a decision boundary that maximizes the separation between different observation classes (e.g., default vs. no default). Key terms
•**Separation Boundary (Hyperplane**): The center line/plane of the path that splits the data classes.
•**Support Vectors:** points lying directly on the edges of the parallel paths. These points alone define the position of the separation boundary.
•**Tradeoff**: Beyond a simple two-feature model, analysts face a structural tradeoff between maximizing the path width (margin size) and minimizing the number of path-driven misclassifications.

K nearest Neighbour

A supervised ML model used for classification or regression. KNN does not learn explicit underlying dataset relationships during training; it is therefore considered a lazy learner.
Measures the distance between a new observation and all training points in the feature space, identifying the K closest observations to make a prediction.
K value is important because

K Value	Bias	Variance
Too large	High	Low
Too small	Low	High

Recommended: Select K = sqrt of n observations

R26 Machine Learning Predictions P2

Decision Trees

Supervised machine learning technique working with sequential input features for both classification problems and continuous variable estimations (Classification and Regression Trees). Known as "white-box models" because they are highly interpretable, unlike "black-box" neural networks.

Tree Anatomy:

- **Root Node:** The topmost node where the initial split occurs.
- **Decision Nodes**: Internal nodes containing questions/conditions about specific features.
- **Terminal Nodes (Leaf Nodes):** The final outcome nodes where predictions are made.

Splitting Criteria & Information Gain:

The goal at each node is to maximize Information Gain (reducing uncertainty/disorder) to achieve a pure set (a node where all observations belong to a single class).

- **Entropy:** Measures dataset disorder; ranges from 0 (perfectly pure) to 1 (maximum impurity/50-50 split).
- **Gini Impurity:** Measures the probability of misclassifying a randomly chosen element.
- **Continuous Variables:** For continuous features, the algorithm tests different threshold values and selects the one that maximizes information gain.

Overfitting Mitigation (Pruning):

Decision trees carry a high risk of overfitting. This is controlled via two main methods:

- **Pre-pruning (Early Stopping):** Halts tree growth and splitting if a node contains fewer than a pre-specified number of training observations or fails to meet a minimum threshold.
- **Post-pruning:** Allows the tree to grow to its maximum size first, then systematically removes weak or low-value nodes/branches afterward to improve generalization.

Ensemble Method

Combines the outputs of multiple underlying models (learners) into a single, unified model. The core objectives are leveraging the "wisdom of crowds" (averaging predictions) and protecting against overfitting.

Techniques for Building Ensembles:

Bootstrap Aggregation (Bagging): Multiple decision trees are built simultaneously using training data sampled **with replacement. Out-of-Bag (OOB):** Data points left out during bootstrap sampling, which are used to evaluate the model's performance without a separate validation set. **Pasting** is Similar to bagging, but data is sampled **without replacement.**

Random Forests (Ensemble of decision tree): Reduces correlations between individual decision trees to optimize ensemble performance. Achieved by randomly sampling a subset of features or observations without replacement at each split.

Boosting: A sequential ensemble method designed to improve performance by learning from the errors of previously grown trees. **Gradient Boosting:** Successive models are trained on the residual errors (residuals) of the preceding model. **Adaptive Boosting (AdaBoost):** Begins with equal weights on all training observations, then sequentially increases weights on misclassified outputs so subsequent trees focus on tougher cases.

Artificial Neural Networks (ANNs):

Designed to mimic the human brain's computational structure to identify complex, non-linear relationships. Most common type is feedforward network with backpropagation.

Network Structure & Mechanics:

- **Inputs & Hidden Layers**: Input features are passed forward through one or more hidden layers containing nodes.
- **Weights W & Biases**: Applied to inputs at each node. Weights scale the input values, while the constant bias (activation parameter) dictates how easily a node "fires" (generates an output of 1).
- **Activation Functions**: Applied at each layer to introduce non-linearity into the input-output relationships. The logistic (sigmoid) function is a common choice.
- **Backpropagation**: The iterative process by which weights and biases throughout the network are constantly updated by sending error signals backward from the output layer.

Optimization via Gradient Descent: Used to minimize the objective (loss) function by iteratively moving in the direction of the "steepest descent down a valley."

Learning Rate: A hyperparameter controlling the size of the steps taken down the loss valley. **Too large**: Causes drastic movements back and forth between the sides of the valley, potentially missing the minimum. **Too small**: Algorithm takes too long to converge on the global minimum.

Overfitting Mitigation (Early Stopping): Done by tracking the loss functions of both the training and validation datasets simultaneously. The Stopping Rule: As gradient descent progresses, the loss will decrease for both datasets. The algorithm must be stopped at the exact point where the validation set loss begins to increase/diverge, even if the training set loss continues to improve.

Confusion Matrix

Used to evaluate the classification performance of models with binary categorical outputs (0 or 1) by comparing predicted classes against actual classes.

The Four Elements:

1.True Positive (TP) 2.False Positive (FP) 3.True Negative (TN) 4.False Negative (FN)

		Prediction	
		True	False
Outcome	True	TP	FN (Type II Error)
	False	FP (T1 Error)	TN

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

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

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

$$\text{Specificity} = \frac{TN}{TN+FP}$$

ROC Curve and AUC:

- **ROC Curve:** plot illustrating the relationship between the True Positive Rate (Recall) on the y-axis and the False Positive Rate on the x-axis across various threshold settings.
- **Area Under the Curve (AUC):** A single metric measuring the overall ability of the model to discriminate between classes.
  - AUC = 1.0: Perfect model;
  - AUC = 0.5: No predictive power;
  - AUC < 0.5: Performs worse than random guessing (counter-predictive).