Comprehensive Data Science Concepts Summary

This document provides a concise theoretical overview of key concepts in Probability, Statistics, and Machine Learning, including essential formulas.

1 Probability Theory

Basic Concepts

- **Sample Space (Ω)**: Set of all possible outcomes of an experiment.
- **Event (E)**: A subset of the sample space.
- **Probability ($P(E)$)**: A measure of the likelihood of an event occurring, $0 \leq P(E) \leq 1$.
- **Conditional Probability ($P(A|B)$)**: Probability of event A occurring given that event B has occurred.
$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ where } P(B) > 0$$
- **Independent Events**: Two events A and B are independent if $P(A \cap B) = P(A)P(B)$, or equivalently, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.
- **Mutually Exclusive Events**: Two events A and B are mutually exclusive if they cannot occur at the same time, i.e., $P(A \cap B) = 0$.
  - For mutually exclusive events: $P(A \cup B) = P(A) + P(B)$.
- **Complement Rule**: $P(A^c) = 1 - P(A)$.
- **Addition Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Bayes' Theorem

Relates conditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$ (for binary A).

Permutations & Combinations

- **Permutation ($P(n,k)$)**: Number of ways to arrange k items from a set of n distinct items, where order matters.
$$P(n,k) = \frac{n!}{(n-k)!}$$
- **Combination ($C(n,k)$ or $\binom{n}{k}$)**: Number of ways to choose k items from a set of n distinct items, where order does not matter.
$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Probability Distributions

- **Probability Mass Function (PMF)**: For discrete random variables, gives the probability that the variable takes on a specific value.
- **Probability Density Function (PDF)**: For continuous random variables, describes the likelihood of the variable falling within a given range (area under the curve).
- **Expected Value ($E[X]$)**: The long-run average value of a random variable.
- **Variance ($Var[X]$)**: Measures the spread or dispersion of a random variable's values around its expected value. $Var[X] = E[(X - E[X])^2]$.

Basic Distributions

1. **Bernoulli Distribution** ($X \sim \text{Bernoulli}(p)$)
   - Models a single trial with two outcomes (success/failure).
   - PMF: $P(X = k) = p^k(1 - p)^{1-k}$ for $k \in \{0, 1\}$
   - $E[X] = p$
   - $Var[X] = p(1 - p)$
2. **Binomial Distribution** ($X \sim \text{Binomial}(n, p)$)
   - Models the number of successes in n independent Bernoulli trials.
   - PMF: $P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$ for $k \in \{0, 1, \ldots, n\}$
   - $E[X] = np$
   - $Var[X] = np(1 - p)$
3. **Poisson Distribution** ($X \sim \text{Poisson}(\lambda)$)
   - Models the number of events occurring in a fixed interval of time or space, given a constant average rate λ.
   - PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k \in \{0, 1, 2, \ldots\}$
   - $E[X] = \lambda$
   - $Var[X] = \lambda$
4. **Normal (Gaussian) Distribution** ($X \sim \mathcal{N}(\mu, \sigma^2)$)
   - Symmetric, bell-shaped continuous distribution.
   - PDF: $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
   - Properties: Defined by mean (μ) and variance (σ^2). The 68-95-99.7 rule applies.

2 Statistics

Descriptive Statistics

- **Measures of Central Tendency**:
  - **Mean**: Sum of values divided by count. Sensitive to outliers.
  - **Median**: Middle value when data is ordered. Robust to outliers.
  - **Mode**: Most frequent value.
- **Measures of Dispersion**:
  - **Variance (σ^2)**: Average of the squared differences from the mean.
  - **Standard Deviation (σ)**: Square root of variance. In original units.
  - **Interquartile Range (IQR)**: $Q_3 - Q_1$. Robust to outliers.
- **Skewness**:
  - **Right-Skewed (Positive Skew)**: Long tail to the right. Mean > Median > Mode.
  - **Left-Skewed (Negative Skew)**: Long tail to the left. Mean < Median < Mode.

Inferential Statistics

- **Central Limit Theorem (CLT)**: States that the sampling distribution of the sample mean (or sum) will be approximately normally distributed, regardless of the population distribution, as the sample size increases. Crucial for hypothesis testing and confidence intervals.
- **Hypothesis Testing**:
  - **Null Hypothesis (H_0)**: A statement of no effect or no difference.
  - **Alternative Hypothesis (H_1 or H_a)**: A statement that contradicts the null hypothesis.
  - **p-value**: The probability of observing data as extreme as, or more extreme than, the sample data, assuming the null hypothesis is true. A small p-value (typically < 0.05) leads to rejection of H_0.
  - **Type I Error (False Positive)**: Rejecting a true null hypothesis. (α = significance level).
  - **Type II Error (False Negative)**: Failing to reject a false null hypothesis. (β).
- **Confidence Intervals (CI)**: A range of values, derived from sample data, that is likely to contain the true population parameter with a certain level of confidence.
  - **Interpretation**: ”We are X% confident that the true population parameter lies within this interval.” (The confidence is in the method, not a probability about the specific interval).
  - **Factors Affecting Width**:
    - **Confidence Level**: Higher confidence level (e.g., 99% vs 95%) leads to a wider interval.
    - **Sample Size**: Larger sample size leads to a narrower interval.
    - **Standard Deviation**: Larger population standard deviation leads to a wider interval.
- **Sampling Bias**: Systematic errors in data collection that lead to unrepresentative samples.
  - **Selection Bias**: Non-random selection of participants (e.g., surveying only library-goers).
  - **Response Bias**: Participants provide inaccurate answers (e.g., social desirability bias).

3 Machine Learning Theory

ML Paradigms

- **Supervised Learning**: Learns a mapping from input features to known output labels using labeled historical data (e.g., Classification, Regression).
- **Unsupervised Learning**: Discovers hidden patterns or structures in data without labeled outputs (e.g., Clustering, Dimensionality Reduction).
- **Reinforcement Learning**: An agent learns to make decisions by interacting with an environment to maximize a reward signal.

Model Evaluation

Regression Metrics

- **Mean Squared Error (MSE)**: Average of squared differences between predictions and actual values. Penalizes larger errors more. Continuously differentiable.

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

- **Mean Absolute Error (MAE)**: Average of absolute differences. More robust to outliers than MSE. Not differentiable at zero.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE)**: Square root of MSE. In the same units as the target variable.
$$\text{RMSE} = \sqrt{\text{MSE}}$$
- **R-squared (R^2)**: Proportion of variance in the dependent variable predictable from independent variables. Can be negative.
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$
- **Adjusted R-squared**: R^2 adjusted for the number of predictors, penalizing unnecessary variables. Useful for comparing models with different complexities.

Classification Metrics (for Binary Classification, Positive Class)

- **True Positive (TP)**: Actual positive, predicted positive.
- **True Negative (TN)**: Actual negative, predicted negative.
- **False Positive (FP)**: Actual negative, predicted positive (Type I error).
- **False Negative (FN)**: Actual positive, predicted negative (Type II error).
- **Accuracy**: Overall proportion of correct predictions.
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- **Precision**: Proportion of positive predictions that were actually correct. Minimizes false alarms.
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- **Recall (Sensitivity)**: Proportion of actual positives that were correctly identified. Minimizes missed positives.
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
F1-score: Harmonic mean of Precision and Recall. Balances both.
$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model Performance Issues

- **Underfitting (High Bias)**: Model is too simple to capture the underlying patterns. Poor performance on both training and validation data.
  - **Remedies**: Increase model complexity (more features, more powerful algorithm, deeper network), reduce regularization.
- **Overfitting (High Variance)**: Model learns the training data (including noise) too well and fails to generalize to unseen data. Excellent training performance, poor validation performance.
  - **Remedies**: Reduce model complexity (fewer features, simpler algorithm, shallower network), regularization, more data, early stopping.
- **Bias-Variance Trade-off**: A fundamental concept. Simple models have high bias and low variance (underfit). Complex models have low bias and high variance (overfit). The goal is to find an optimal balance.

Data Preprocessing

- **Feature Scaling**:
  - **Min-Max Scaling (Normalization)**: Scales features to a specific range (e.g., [0, 1]).
  $$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$
  - **Standardization (Z-score Normalization)**: Transforms features to have a mean of 0 and standard deviation of 1.
  $$X_{scaled} = \frac{X - \mu}{\sigma}$$
  - **When to Use**: Essential for distance-based models (KNN, SVM, K-Means), gradient-based models (Neural Networks, Logistic Regression) for faster convergence. Not strictly required for tree-based models.
- **Categorical Encoding**:
  - **One-Hot Encoding**: Creates binary columns for each category. Suitable for nominal (unordered) features. Avoids false ordering. Can lead to high dimensionality for high cardinality features.
  - **Ordinal Encoding**: Assigns numerical values based on inherent order. Suitable for ordinal (ordered) features (e.g., 'Low'=1, 'Medium'=2, 'High'=3).
  - **Label Encoding**: Assigns a unique integer to each category. Simple but can imply false order for nominal features.
  - **Binary Encoding**: Converts categories to binary code, then creates columns for each bit. Reduces dimensionality compared to one-hot for high cardinality.
- **Handling Missing Values (Imputation)**: Strategies to fill in missing data.
  - Mean, Median, Mode Imputation
  - K-Nearest Neighbors (KNN) Imputation
  - Regression Imputation
  - Dropping rows/columns (if missingness is extensive or specific).
- **Handling Class Imbalance**: When one class significantly outnumbers others.
  - **Data-level**: Oversampling minority class (e.g., SMOTE, random oversampling), Undersampling majority class (e.g., random undersampling, NearMiss), Data Augmentation (especially for images).
  - **Algorithm-level**: Cost-sensitive learning, using appropriate evaluation metrics (Precision, Recall, F1-score, AUC-PR). **Model-level**: Transfer learning (for image/text with pre-trained models on balanced data).

Regularization

Techniques to prevent overfitting by penalizing model complexity.

- **L1 Regularization (Lasso)**: Adds penalty proportional to the sum of the absolute values of coefficients ($\sum |\beta_i|$). Can drive coefficients exactly to zero, performing feature selection.
- **L2 Regularization (Ridge)**: Adds penalty proportional to the sum of the squared values of coefficients ($\sum \beta_i^2$). Shrinks coefficients towards zero but rarely to exactly zero.
- **Early Stopping**: Halts training of iterative models when performance on a validation set starts to degrade.

Ensemble Methods

Combine multiple models to improve overall performance and robustness.

- **Bagging (Bootstrap Aggregating)**: Builds multiple models independently on bootstrapped (randomly sampled with replacement) subsets of the data. Predictions are averaged (regression) or majority-voted (classification). Primarily reduces **variance**.
  - **Random Forest**: An ensemble of decision trees built using bagging, with an additional random feature subset selection at each split to decorrelate trees.
- **Boosting**: Builds models sequentially, where each new model tries to correct the errors (residuals) of the previous models. Primarily reduces **bias**.
  - **Gradient Boosting Machines (GBMs)**: Iteratively train weak learners (often shallow trees) on the residuals of previous predictions. Examples include XGBoost, LightGBM, CatBoost.
- **General Characteristics**: Often achieve higher accuracy than single models, generally robust to multicollinearity (for tree-based ensembles), but can be less interpretable.

Model Interpretability

- **Shallow Decision Trees**: Highly interpretable, rules are clear.
- **Feature Importance**: Provided by many models (e.g., Random Forest, GBMs) to show which features are globally most influential.
- **SHAP (SHapley Additive exPlanations)**: A powerful post-hoc technique to explain *individual predictions* of any complex machine learning model by attributing the contribution of each feature.

Clustering Methods (Unsupervised Learning)

- **K-Means Clustering**: Partitions data into k clusters based on minimizing the sum of squared distances from each data point to its assigned cluster centroid. Requires specifying k in advance.
- **Hierarchical Clustering**: Builds a hierarchy of clusters. Can be agglomerative (bottom-up, merging clusters) or divisive (top-down, splitting clusters). Results visualized with a dendrogram.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Groups together data points that are closely packed together, marking as outliers points that lie alone in low-density regions. Does not require specifying k in advance.

Neural Networks Basics

- **Activation Functions**: Introduce non-linearity into the network, allowing it to learn complex patterns.
  - **ReLU (Rectified Linear Unit)**: $f(x) = \max(0, x)$. Popular for hidden layers, addresses vanishing gradient for positive inputs.
  - **Leaky ReLU**: $f(x) = \max(0.01x, x)$. Addresses ”dying ReLU” by allowing a small gradient for negative inputs.
  - **Sigmoid**: $f(x) = \frac{1}{1+e^{-x}}$. Outputs values between 0 and 1. Used for binary classification output. Can suffer from vanishing gradients in hidden layers.
  - **Tanh (Hyperbolic Tangent)**: $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Outputs values between -1 and 1. Similar to sigmoid but zero-centered.
  - **Softmax**: Converts a vector of raw scores (logits) into a probability distribution that sums to 1. Essential for multi-class classification output layers.

Model Deployment & Monitoring

- **Data Drift**: Change in the distribution of input features over time.
- **Model Drift (Concept Drift)**: Change in the relationship between input features and the target variable over time.
- **Performance Monitoring**: Continuously tracking key model metrics (accuracy, precision, recall, RMSE) and business outcomes in production to detect degradation.
- **Automated Retraining**: Pipelines to retrain models, often with human oversight, when drift or performance degradation is detected.