DSA 610 Redesign, Lecture 11 Outline

Lecture Outline: Data Ethics and Privacy, with a Focus on the Census and Big Data
Duration: 50 minutes

---

1. Introduction to Data Ethics and Privacy (5 minutes)

- **Objective:** Understand the importance of data ethics and privacy in the context of data analysis and usage.
- **Content:**
  - **Definition:** Data ethics involves the principles and guidelines for responsible data use, including privacy, consent, and fairness.
  - **Importance:** Ensuring data is used ethically protects individuals' rights and maintains trust in data-driven systems.

---

2. Key Principles of Data Ethics (10 minutes)

- **Objective:** Explore the foundational principles of ethical data use.
- **Content:**
  - **1. Consent:**
    - **Explanation:** Data must be collected and used with the explicit consent of individuals.
    - **Example:** Informing users about how their data will be used and obtaining their agreement.
  - **2. Privacy:**
    - **Explanation:** Protecting individuals' personal information from unauthorized access or misuse.
    - **Example:** Anonymizing or pseudonymizing data to prevent identification.
  - **3. Transparency:**
    - **Explanation:** Being open about data collection practices and usage.
    - **Example:** Providing clear privacy policies and disclosures.
  - **4. Fairness:**
    - **Explanation:** Avoiding biases and ensuring equitable treatment in data analysis and decision-making.
    - **Example:** Regularly auditing algorithms for fairness and mitigating discriminatory impacts.
  - **5. Accountability:**
    - **Explanation:** Holding individuals and organizations responsible for ethical data practices.
    - **Example:** Implementing governance structures and regular reviews of data practices.

---

3. Data Privacy Regulations (10 minutes)

- **Objective:** Review major data privacy regulations and their implications.
- **Content:**
  - **1. General Data Protection Regulation (GDPR):**
    - **Overview:** A regulation in the European Union focused on data protection and privacy.
    - **Key Aspects:**
      - **Data Subject Rights:** Right to access, correct, and delete personal data.

        - **Data Protection Impact Assessments (DPIAs):** Assessing risks associated with data processing.
    - **2. California Consumer Privacy Act (CCPA):**
        - **Overview:** A California state law providing data privacy rights to residents.
        - **Key Aspects:**
            - **Right to Know:** Right to know what personal information is collected.
            - **Right to Opt-Out:** Right to opt-out of the sale of personal information.
    - **3. Health Insurance Portability and Accountability Act (HIPAA):**
        - **Overview:** U.S. law protecting sensitive patient health information.
        - **Key Aspects:**
            - **Privacy Rule:** Sets standards for the protection of health information.
            - **Security Rule:** Establishes safeguards for electronic health information.

---

4. Case Study: The U.S. Census (15 minutes)

- **Objective:** Examine the ethical considerations and privacy concerns associated with the U.S. Census.
- **Content:**
    - **1. Overview of the U.S. Census:**
        - **Purpose:** Collects demographic data to inform policy and allocate resources.
        - **Frequency:** Conducted every 10 years.
    - **2. Ethical Considerations:**
        - **Privacy Measures:** Use of anonymization and data suppression techniques to protect individual identities.
        - **Data Sharing:** Balancing the need for detailed data with protecting individual privacy.
    - **3. Big Data and the Census:**
        - **Integration with Big Data:** Use of additional data sources (e.g., social media) to enhance census data.
        - **Challenges:** Ensuring data accuracy and protecting privacy while integrating diverse data sources.
    - **4. Historical Context:**
        - **Past Issues:** Examples of privacy breaches or misuse of census data (e.g., internment of Japanese Americans during WWII).
        - **Current Practices:** Modern safeguards and practices to prevent similar issues.
    - **5. Discussion Points:**
        - **Benefits:** How accurate census data benefits public services and representation.
        - **Risks:** Potential risks of integrating big data with census information and how to mitigate them.

---

5. Best Practices for Ethical Data Use (5 minutes)

- **Objective:** Summarize best practices for ensuring ethical data use.
- **Content:**
    - **1. Data Governance:**
        - **Implementation:** Establishing clear policies and procedures for data handling.
        - **Example:** Regular audits and compliance checks.
    - **2. Ethical Review Boards:**
        - **Implementation:** Creating boards to review data use and research proposals.
        - **Example:** Academic and corporate ethical review committees.

- **3. Training and Awareness:**
    - **Implementation:** Educating data professionals about ethical practices.
    - **Example:** Regular training sessions and workshops.

---

6. Q&A and Discussion (5 minutes)

- **Objective:** Address questions and discuss practical aspects of data ethics and privacy.
- **Content:**
  - **Q&A Session:** Open the floor for student questions.
  - **Discussion:** Explore real-world scenarios and ethical dilemmas in data handling.

Key Takeaways

- **Data Ethics:** Principles of consent, privacy, transparency, fairness, and accountability.
- **Privacy Regulations:** Overview of GDPR, CCPA, and HIPAA.
- **Census Case Study:** Ethical considerations, privacy measures, and challenges related to big data integration.
- **Best Practices:** Data governance, ethical review boards, and training for ethical data use.

Resources:
Federal Data Strategy: https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
The Ethcs of Managing People's Data: https://hbr.org/2023/07/the-ethics-of-managing-peoples-data
Data Ethics: Examples and Principles: https://studyonline.unsw.edu.au/blog/data-ethics-overview
The Importance of Data Ethics: https://www.isba.org.uk/article/importance-data-ethics-why-businesses-must-take-it-seriously
Data Ethics Basics: https://libguides.library.ncat.edu/c.php?g=778712&p=10368600
US Dept of Commerce: Privacy Laws, Policies and Guidance: https://www.commerce.gov/opog/privacy/privacy-laws-policies-and-guidance
US Data Privacy Laws: https://www.forbes.com/sites/conormurray/2023/04/21/us-data-privacy-protection-laws-a-comprehensive-guide/ (you may need to access this through the campus library)
US State Privacy Legislation Tracker: https://iapp.org/resources/article/us-state-privacy-legislation-tracker/
GDPR: https://gdpr.eu/what-is-gdpr/
CCPA: https://oag.ca.gov/privacy/ccpa
HIPAA: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html
FERPA: https://studentprivacy.ed.gov/ferpa
Data Security Guidance for Human Subject Research: https://ovpr.uconn.edu/services/rics/irb/data-security-guidance-for-human-subjects-research/
Differential Privacy: https://www.ncsl.org/technology-and-communication/differential-privacy-for-census-data-explained
Census Data Protection and Privacy: https://www.census.gov/about/policies/privacy.html
Data Security Best Practices: https://coag.gov/app/uploads/2022/01/Data-Security-Best-Practices.pdf

Lecture Outline: Data Rescaling and Distance Metrics
Duration: 50 minutes

---

1. Introduction to Data Rescaling (5 minutes)

- **Objective:** Understand the importance of data rescaling in preprocessing for machine learning algorithms.
- **Content:**

- **Definition:** Data rescaling transforms features to a common scale, improving the performance of various algorithms.
- **Purpose:** Ensures that features contribute equally to the analysis, particularly in distance-based algorithms.

---

2. Rescaling Methods (20 minutes)

a. Standardization (Z-score Normalization)

- **Definition:** Transforms data to have a mean of 0 and a standard deviation of 1.
- **Formula:** $z = \frac{x-\mu}{\sigma}$
  - x is the original value, μ is the mean, and σ is the standard deviation (usually, this is mostly t-score normalization using the mean and standard deviation of the available data)
- **Benefits:**
  - Centers data around zero, making it easier for algorithms that assume data is normally distributed.
  - Useful for algorithms like linear regression and logistic regression that are sensitive to feature scaling.

b. Min-Max Scaling (Rescaling to [0, 1])

- **Definition:** Transforms data to a specific range, typically [0, 1].
- **Formula:** $x' = \frac{x-\text{MIN}}{\text{MAX}-\text{MIN}}$
  - Here x is the original observation and MIN is the minimum of the data, and MAX is the maximum of the data (one column at a time).
- **Benefits:**
  - Ensures all features contribute equally, particularly for algorithms that are sensitive to the scale, such as k-nearest neighbors (KNN) and neural networks.
  - Useful when you need to ensure that features are bounded within a specific range.

c. Min-Max Scaling to [-1, 1]

- **Definition:** Similar to Min-Max scaling but transforms data to the range [-1, 1].
- **Formula:** $x' = \frac{2(x-MIN)}{MAX-MIN} - 1$
  - x is the original value, MIN is the minimum of the variable, MAX is the maximum of the variable
- **Benefits:**
  - Useful when data needs to be centered around zero but still scaled to a bounded range.
  - Helps in scenarios where algorithms perform better with input values in the range [-1, 1].

d. Robust Scaling

- **Definition:** Scales data based on the median and interquartile range (IQR).
- **Formula:** $x' = \frac{x-median}{IQR}$
  - x is the original value, median is the median of the variable, and IQR is the interquartile range of the variable
- **Benefits:**
    - Less sensitive to outliers compared to standardization and Min-Max scaling.
    - Useful in cases where the data contains significant outliers.

e. Log Transformation

- **Definition:** Applies the logarithm function to transform data, often used to handle skewed distributions.
- **Formula:** $x' = \log{(x+1)}$

- **Benefits:**
    - Reduces skewness and handles wide ranges of data.
    - Improves interpretability and stability of statistical models.

---

3. Distance Metrics (15 minutes)

a. Euclidean Distance

- **Definition:** Measures the straight-line distance between two points in Euclidean space.
- **Formula:** $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
    - **Where:**
        - x_i and y_i are values of features for two points
        - n is the number of features
- **Use Case:** Commonly used in KNN, clustering algorithms (e.g., K-Means).

b. Manhattan Distance

- **Definition:** Measures the distance between two points along axes at right angles (grid-like path).
- **Formula:** $d = \sum_{i=1}^{n}|x_i - y_i|$
    - **Where:**
        - x_i and y_i are values of features for two points
        - n is the number of features
    - **Use Case:** Used in situations where grid-like paths are more relevant, such as certain types of clustering algorithms.

c. Minkowski Distance

- **Definition:** Generalization of Euclidean and Manhattan distances.
- **Formula:** $d = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}$
    - **Where:**
        - p is a parameter that defines the distance metric (e.g., p=2 for Euclidean, p=1 for Manhattan)
        - x_i and y_i are values of features for two points
        - n is the number of features
    - **Use Case:** Flexible distance measure for different types of data and models.

d. Cosine Similarity

- **Definition:** Measures the cosine of the angle between two vectors.
- **Formula:** similarity $= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$
    - **Where:**
        - x_i and y_i are values of features for two points
        - n is the number of features

    - **Use Case:** Commonly used in text mining and document similarity.

e. Jaccard Similarity

- **Definition:** Measures similarity between finite sample sets.
- **Formula:** $similarity = \frac{|A \cap B|}{|A \cup B|}$
    - **Use Case:** Useful for binary and categorical data, often in clustering and classification tasks.

---

4. Effects of Rescaling on Algorithms (10 minutes)

- **Objective:** Understand how rescaling affects the performance and interpretation of machine learning algorithms.
- **Content:**
    - **Impact on Distance-Based Algorithms:**
        - **K-Nearest Neighbors (KNN):** Sensitive to feature scaling; rescaling ensures that all features contribute equally to distance calculations.
        - **Clustering (e.g., K-Means):** Rescaling affects cluster formation as algorithms are sensitive to feature scales.
    - **Impact on Gradient-Based Algorithms:**
        - **Neural Networks:** Standardization helps in faster convergence during training by ensuring that weights are updated more consistently.
    - **Impact on Regularization:**
        - **Regularized Regression Models:** Standardization ensures that regularization penalizes features equally, preventing bias towards features with larger scales.
    - **General Considerations:**
        - **Consistency:** Always apply the same scaling to both training and test data to avoid discrepancies.
        - **Model Sensitivity:** Choose the scaling method based on the model's sensitivity to feature scales and the nature of the data.

---

5. Q&A and Discussion (5 minutes)

- **Objective:** Address questions and discuss practical considerations related to data rescaling and distance metrics.
- **Content:**
    - **Q&A Session:** Open the floor for student questions.
    - **Discussion:** Explore scenarios and best practices for rescaling and choosing appropriate distance metrics.

Key Takeaways

- **Data Rescaling:** Methods such as standardization, Min-Max scaling, and log transformation improve model performance and ensure features contribute equally.
- **Distance Metrics:** Different metrics like Euclidean, Manhattan, and Cosine similarity have unique applications and affect algorithm performance.
- **Impact on Algorithms:** Rescaling affects algorithms differently; understanding this helps in selecting appropriate preprocessing steps for optimal model performance.

Resources:
Z-score Normalization: https://www.statology.org/z-score-normalization/
Min-Max Normalization: https://www.codecademy.com/article/normalization
-1 to 1: https://www.statology.org/normalize-data-between-1-and-1/
Robust Scaling: https://proclusacademy.com/blog/robust-scaler-outliers/
Distance Metrics: https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/
Euclidean Distance: https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/
Manhattan Distance: https://www.datacamp.com/tutorial/manhattan-distance
Minkowski Distance: https://www.datacamp.com/tutorial/minkowski-distance

Cosine Similarity: https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded
Jaccard Similarity: https://medium.com/@mayurdhvajsinhjadeja/jaccard-similarity-34e2c15fb524
Hamming Distance: https://www.tutorialspoint.com/what-is-hamming-distance
Levenshtein Distance: https://www.geeksforgeeks.org/introduction-to-levenshtein-distance/
Mahalanobis Distance: https://www.statisticshowto.com/mahalanobis-distance/
Haversine Distance: https://www.geeksforgeeks.org/haversine-formula-to-find-distance-between-two-points-on-a-sphere/

Lecture Outline: Validation Methods and Metrics in Python for Various Models
Duration: 50 minutes

---

1. Introduction to Model Validation (5 minutes)

- **Objective:** Understand the purpose of model validation and its importance in assessing the performance of machine learning models.
- **Content:**
    - **Definition:** Model validation involves assessing how well a model performs on unseen data to ensure it generalizes well.
    - **Purpose:** To avoid overfitting, select the best model, and ensure robustness of predictions.

---

2. Validation Methods (15 minutes)
a. Hold-Out Validation

- **Definition:** Splitting the dataset into training and testing sets.
- **Process:**
    - **Training Set:** Used to train the model.
    - **Testing Set:** Used to evaluate the model's performance.
- **Pros and Cons:**
    - **Pros:** Simple to implement.
    - **Cons:** Results may vary based on the particular split.

b. K-Fold Cross-Validation

- **Definition:** Divides the dataset into k subsets (folds). The model is trained on k−1 folds and validated on the remaining fold, repeating this process k times.
- **Process:**
    - **Steps:** Split data into k folds, train and validate k times.
    - **Pros and Cons:**
        - **Pros:** Reduces variance in performance estimates, utilizes all data for both training and validation.
        - **Cons:** Computationally expensive.

c. Leave-One-Out Cross-Validation (LOOCV)

- **Definition:** Special case of K-Fold Cross-Validation where k equals the number of data points. Each data point is used as a test set exactly once.
- **Process:**
    - **Steps:** Train on n−1 samples, validate on the remaining single sample.
    - **Pros and Cons:**
        - **Pros:** Uses almost all data for training.
        - **Cons:** Computationally intensive for large datasets.

d. Stratified Cross-Validation

- **Definition:** Variation of K-Fold Cross-Validation that ensures each fold is representative of the whole dataset's class distribution.
- **Process:**
    - **Steps:** Similar to K-Fold but preserves class distribution in each fold.
    - **Pros and Cons:**
        - **Pros:** Useful for imbalanced datasets.
        - **Cons:** Slightly more complex than regular K-Fold.

---

3. Validation Metrics (15 minutes)

a. Classification Metrics

- **Accuracy:** Proportion of correctly classified instances.
    - **Formula:** $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$
- **Precision:** Proportion of true positive predictions among all positive predictions.
    - **Formula:** $Precision = \frac{TP}{TP+FP}$
- **Recall (Sensitivity):** Proportion of true positive predictions among all actual positives.
    - **Formula:** $\frac{TP}{TP+FN}$
- **F1 Score:** Harmonic mean of precision and recall.
    - **Formula:** $F1\ Score = \frac{2(Precision \times Recall)}{Precision+Recall}$
- **ROC Curve and AUC (Area Under the Curve):** Measures performance across different thresholds. AUC indicates the model's ability to discriminate between classes.

b. Regression Metrics

- **Mean Absolute Error (MAE):** Average of absolute errors between predicted and actual values.
    - **Formula:** $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Mean Squared Error (MSE):** Average of squared errors between predicted and actual values.
    - **Formula:** $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Root Mean Squared Error (RMSE):** Square root of MSE, providing error in the same units as the target variable.
    - **Formula:** $RMSE = \sqrt{MSE}$
- **R-Squared (Coefficient of Determination):** Proportion of variance in the dependent variable that is predictable from the independent variables.
    - **Formula:** $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

---

4. Implementing Validation and Metrics in Python (10 minutes)

- **Objective:** Provide an overview of how to implement validation methods and calculate metrics in Python.
- **Content:**
    - **Using Scikit-Learn:**
        - **K-Fold Cross-Validation:**

```
from sklearn.model_selection import cross_val_score
scores = cross_val_score(model, X, y, cv=5)
```

Classification Metrics:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score
accuracy = accuracy_score(y_true, y_pred)
```

```
precision = precision_score(y_true, y_pred)
recall = recall_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred)
roc_auc = roc_auc_score(y_true, y_pred_prob)
```

Regression Metrics:

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
mae = mean_absolute_error(y_true, y_pred)
mse = mean_squared_error(y_true, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_true, y_pred)
```

5. Q&A and Discussion (5 minutes)

- **Objective:** Address questions and discuss practical considerations of validation methods and metrics.
- **Content:**
  - **Q&A Session:** Open the floor for student questions.
  - **Discussion:** Explore real-world scenarios where different validation methods and metrics are applied.

Key Takeaways

- **Model Validation:** Techniques such as Hold-Out, K-Fold Cross-Validation, and LOOCV help in evaluating model performance and avoiding overfitting.
- **Validation Metrics:** Metrics for classification and regression help assess model accuracy, precision, recall, and other performance aspects.
- **Python Implementation:** Scikit-Learn provides tools for implementing validation and calculating metrics, making it easier to assess model performance.

Resources:

Validation Methods in Machine Learning: https://medium.com/@mehmetalitor/top-5-model-validation-methods-in-machine-learning-77eed8d08937
Cross Validation: https://www.w3schools.com/python/python_ml_cross_validation.asp
Time Series Split: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html
Bootstrapping: https://www.digitalocean.com/community/tutorials/bootstrap-sampling-in-python
Test-Train Split: https://www.geeksforgeeks.org/how-to-do-train-test-split-using-sklearn-in-python/
Leave-one-out cross validation: https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/
Holdout Validation: https://github.com/learn-co-students/dsc-1-12-10-holdout-validation-online-ds-pt-112618
Model Evaluation Methodologies: https://medium.com/analytics-vidhya/different-model-evaluation-methodologies-part-2-679fcb064c55
Stratified Cross Sampling: https://towardsdatascience.com/what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e/
Validation Metrics (Classification): https://www.geeksforgeeks.org/metrics-for-machine-learning-model/
Regression Metrics: https://machinelearningmastery.com/regression-metrics-for-machine-learning/