Training Data

SAMPLING, LABELING AND SOLUTIONS FOR IMBALANCES

STUDENT: PAULO EDUARDO DA SILVA JUNIOR

Summary

1. Importance of Training Data;
2. Data Sampling;
3. Data Labeling;
4. Imbalance Classification and Solutions;

Importance of Training Data

1. Data quality is essential for model performance.
2. Data collection is challenging and can compromise ML operations.

Data Sampling

1. Non-Probability Sampling: Convenience, Snowball.
2. Simple Random Sampling: Equal Chance of Selection.
3. Stratified Sampling: Inclusion of All Classes.
4. Weighted and Reservoir Sampling: Prioritization and Continuous Flow.

Data Labeling

1. Manual: Necessary for specific data, but time-consuming.
2. Natural: Based on system signals, such as clicks.
3. Weak Supervision: Fast labeling via heuristics.
4. Implicit and Explicit Labeling: Feedback and lack of response.

Table 4-3. The advantages of programmatic labeling over hand labeling

Hand labeling	Programmatic labeling
Expensive: Especially when subject matter expertise required	**Cost saving**: Expertise can be versioned, shared, and reused across an organization
Lack of privacy: Need to ship data to human annotators	**Privacy**: Create LFs using a cleared data subsample and then apply LFs to other data without looking at individual samples
Slow: Time required scales linearly with number of labels needed	**Fast**: Easily scale from 1K to 1M samples
Nonadaptive: Every change requires relabeling the data	**Adaptive**: When changes happen, just reapply LFs!

Imbalance Classification and Solutions

1. Undersampling and Oversampling: Balancing Classes.
2. Cost-Sensitive Learning: Increased Weight for Minority Class Errors.

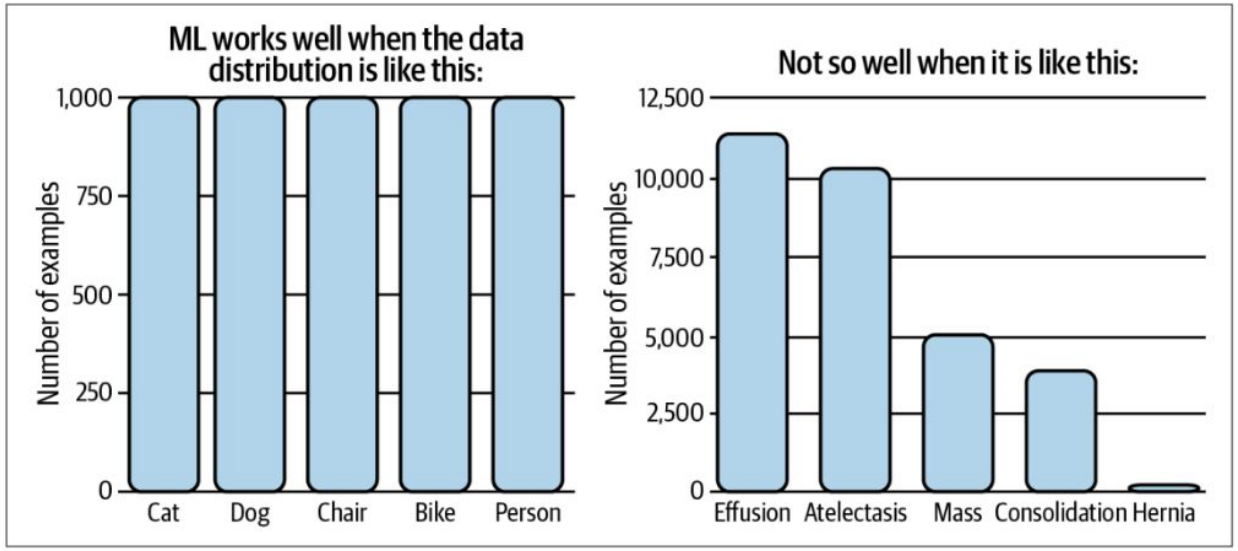

Figure 4-8. ML works well in situations where the classes are balanced. Source: Adapted from an image by Andrew Ng[26]

Conclusion

Training data requires an iterative and structured process.

Effective sampling and labeling ensures robust models.

Class balance is essential to avoid bias in the model.