Data Labeling & Annotation

CS 203: Software Tools and Techniques for AI

Prof. Nipun Batra, IIT Gandhinagar

The Labeling Bottleneck

The Reality:

- Data is abundant (unlabeled).
- Labels are scarce (expensive).
- 80% of AI project time is Data Prep.

Why Labeling Matters:

- **Supervised Learning**: Needs ground truth (y).
- **Evaluation**: Even unsupervised methods need a test set to verify.
- **Ambiguity**: Labeling forces you to define your problem clearly.

Types of Annotation

Computer Vision:

1. **Classification**: "Cat" vs "Dog".
2. **Bounding Box**: Object Detection (x, y, w, h).
3. **Polygon**: Segmentation (pixel-precise).
4. **Keypoints**: Pose estimation (skeleton).

NLP:

1. **Text Classification**: Sentiment, Intent.
2. **NER (Named Entity Recognition)**: Highlighting spans (Person, Org).
3. **Seq2Seq**: Translation, Summarization.

Annotation Interfaces: Label Studio

Label Studio: Open-source, flexible, web-based tool.

Configuration (XML):

```xml
<View>
  <Image name="img" value="$image"/>
  <RectangleLabels name="tag" toName="img">
    <Label value="Car" background="red"/>
    <Label value="Person" background="blue"/>
  </RectangleLabels>
</View>
```

Why usage XML?

- Allows custom UI layouts.
- Can mix modalities (Image + Text + Audio).

Quality Control: The Gold Standard

How do we know if labels are "correct"?

1. **Gold Standard (Ground Truth)**:
    - Experts label a small subset (e.g., 100 items).
    - Annotators are tested against this set.
2. **Consensus (Majority Vote)**:
    - 3 annotators label the same item.
    - If 2 say "Cat" and 1 says "Dog", label is "Cat".

Inter-Annotator Agreement (IAA)

Measure of reliability. Do annotators agree with each other?

Percent Agreement:

$$ext Agreement = \frac{\text{Agreed Items}}{\text{Total Items}}$$

Problem: Doesn't account for chance agreement.

Cohen's Kappa (κ):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- p_o: Observed agreement.
- p_e: Expected agreement by chance.

Cohen's Kappa Example

Scenario: 2 Annotators, 100 items (Yes/No).

- Both say Yes: 45
- Both say No: 45
- Disagree: 10
- $p_o = 0.90$

Chance Calculation:

- A says Yes 50% of time.
- B says Yes 50% of time.
- Chance they agree on Yes = $0.5 \times 0.5 = 0.25$.
- Chance they agree on No = 0.25.

Managing Labeling Teams

Workflow:

1. **Guidelines**: Write detailed instructions (e.g., "Does a reflection count as a car?").
2. **Pilot**: Label 50 items, calculate Kappa. Refine guidelines.
3. **Production**: Label large batch.
4. **Review**: Spot check 10% of labels.

Human-in-the-Loop:

- Use Model to pre-label (Predictions).
- Humans verify/correct (faster than starting from scratch).

Active Learning: Smart Sampling

Problem: Labeling all data is expensive.

Solution: Label only the most informative examples.

Uncertainty Sampling:

```python
def uncertainty_sampling(model, unlabeled_pool, n=100):
    """Select examples where model is least confident."""
    predictions = model.predict_proba(unlabeled_pool)

    # Entropy-based uncertainty
    entropy = -np.sum(predictions * np.log(predictions + 1e-10), axis=1)

    # Select top-n most uncertain
    top_indices = np.argsort(entropy)[-n:]

    return unlabeled_pool[top_indices]
```

Benefit: Can achieve 90% accuracy with 20-30% of labeled data

Active Learning Strategies

1. **Uncertainty Sampling**:
    - **Least Confidence**: $1 - P(\hat{y}|x)$
    - **Margin**: $P(y_1|x) - P(y_2|x)$ (difference between top 2)
    - **Entropy**: $-\sum P(y|x) \log P(y|x)$
2. **Query-by-Committee**:
    - Train ensemble of models
    - Select examples with highest disagreement
3. **Expected Model Change**:
    - Select examples that would change model most if labeled
4. **Diversity Sampling**:

Active Learning Implementation

```
from sklearn.ensemble import RandomForestClassifier
from scipy.stats import entropy

class ActiveLearner:
    def __init__(self, model, X_pool, y_pool=None):
        self.model = model
        self.X_pool = X_pool
        self.y_pool = y_pool
        self.X_labeled = []
        self.y_labeled = []

    def query(self, n_samples=10, strategy='entropy'):
        """Query most informative samples."""
        if strategy == 'entropy':
            probs = self.model.predict_proba(self.X_pool)
            scores = entropy(probs.T)
        elif strategy == 'margin':
            probs = self.model.predict_proba(self.X_pool)
            probs_sorted = np.sort(probs, axis=1)
            scores = probs_sorted[:, -1] - probs_sorted[:, -2]

        # Select top-n uncertain samples
        query_idx = np.argsort(scores)[-n_samples:]
        return query_idx

    def teach(self, idx, labels):
        """Add labeled samples to training set."""
        self.X_labeled.extend(self.X_pool[idx])
        self.y_labeled.extend(labels)

        # Remove from pool
        self.X_pool = np.delete(self.X_pool, idx, axis=0)

        # Retrain model
        self.model.fit(self.X_labeled, self.y_labeled)
```

Weak Supervision: Programmatic Labeling

Idea: Write labeling functions instead of manually labeling.

Example (Spam detection):

```python
# Labeling Function 1: Check for money mentions
def lf_contains_money(text):
    if re.search(r'\$\d+|\bmoney\b', text, re.I):
        return 1  # SPAM
    return -1  # NOT_SPAM

# Labeling Function 2: All caps
def lf_all_caps(text):
    if text.isupper() and len(text) > 10:
        return 1  # SPAM
    return -1  # NOT_SPAM

# Labeling Function 3: Urgency words
def lf_urgency(text):
    urgent_words = ['urgent', 'act now', 'limited time']
```

Snorkel Framework

Snorkel: Framework for weak supervision.

Workflow:

1. Write labeling functions (LFs)
2. Apply LFs to unlabeled data
3. Learn generative model to denoise labels
4. Train final classifier on probabilistic labels

```
from snorkel.labeling import labeling_function, PandasLFApplier
from snorkel.labeling.model import LabelModel

@labeling_function()
def lf_keyword_spam(x):
    keywords = ['free', 'win', 'click here']
    return 1 if any(k in x.text.lower() for k in keywords) else -1
```

Weak Supervision: Benefits and Challenges

Benefits:

- **Scalability**: Label thousands of examples in minutes
- **Flexibility**: Encode domain knowledge
- **Iteration**: Easy to update rules vs re-labeling
- **Cost**: Much cheaper than manual annotation

Challenges:

- **Quality**: LFs may be noisy or conflicting
- **Coverage**: LFs may abstain on many examples
- **Bias**: Rules encode human biases
- **Complexity**: Need to balance precision vs coverage

Semi-Supervised Learning

Setting: Small labeled set + Large unlabeled set.

Self-Training:

1. Train model on labeled data
2. Predict on unlabeled data
3. Add high-confidence predictions to labeled set
4. Repeat

```
def self_training(model, X_labeled, y_labeled, X_unlabeled, threshold=0.9):
    """Iterative self-training."""
    for iteration in range(10):
        # Train on current labeled set
        model.fit(X_labeled, y_labeled)

        # Predict on unlabeled
        probs = model.predict_proba(X_unlabeled)
        max_probs = np.max(probs, axis=1)
        preds = np.argmax(probs, axis=1)
```

Co-Training: Multi-View Learning

Idea: Use different views of same data.

Example (Web page classification):

- View 1: Page text
- View 2: Anchor text from links to page

Algorithm:

1. Train classifier on each view independently
2. Each classifier labels unlabeled data
3. Add high-confidence predictions from one view to other view's training set

```python
def co_training(X1, X2, y, X1_unlabeled, X2_unlabeled, n_iter=10):
    """Co-training with two views."""
    model1 = RandomForestClassifier()
    model2 = RandomForestClassifier()
```

Crowdsourcing Platforms

Major Platforms:

Platform	Use Case	Cost	Quality
Amazon MTurk	Generic tasks	Low	Variable
Scale AI	High-quality CV/NLP	High	High
Labelbox	Enterprise labeling	Medium	Medium-High
Hive	Image/video annotation	Medium	High
Appen	Multilingual, global	Medium	Medium

Key Considerations:

- **Task complexity**: Simple (MTurk) vs Expert (Scale AI)
- **Quality control**: Gold standard questions, redundancy

Crowdsourcing: Quality Control

Challenge: Workers may be inattentive or malicious.

Solutions:

1. **Gold Standard Questions** (Test questions):

```
# Mix 10% gold questions into tasks
gold_questions = [
    {"text": "The sky is blue", "label": "POSITIVE"},  # Known
]

# Check worker accuracy on gold questions
if worker_accuracy < 0.8:
    reject_worker()
```

2. **Redundancy** (Multiple workers per item):

- Assign same task to 3-5 workers

Advanced IAA: Fleiss' Kappa

Fleiss' Kappa: Extends Cohen's Kappa to >2 annotators.

Formula:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Where:

- $\bar{P}$: Overall agreement
- $\bar{P}_e$: Expected agreement by chance

Implementation:

```
from statsmodels.stats.inter_rater import fleiss_kappa

# Data format: (n_items, n_categories)
```

Krippendorff's Alpha

Most general IAA metric:

- Works with any number of annotators
- Handles missing data
- Works with different data types (nominal, ordinal, interval, ratio)

Formula:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where:

- D_o: Observed disagreement
- D_e: Expected disagreement by chance

Labeling Bias: Types and Impact

Types of Bias:

1. **Selection Bias**:
   - Annotators label non-representative subset
   - Example: Only labeling easy cases
2. **Confirmation Bias**:
   - Annotators favor pre-existing beliefs
   - Example: Political bias in sentiment labeling
3. **Anchoring Bias**:
   - First label influences subsequent labels
   - Example: Seeing "spam" first makes next emails look more like spam

Mitigating Labeling Bias

Strategies:

1. **Blind Annotation**:

```python
# Don't show annotators previous labels or predictions
# Randomize order to prevent patterns
def randomize_annotation_order(tasks):
    return random.sample(tasks, len(tasks))
```

2. **Calibration Sessions**:

   - Train annotators together
   - Discuss edge cases and establish consensus

3. **Regular Audits**:

```python
def audit_annotator_quality(annotations, gold_standard):
```

Label Noise: Detection and Handling

Sources of Noise:

- Genuine ambiguity in data
- Annotator mistakes
- Poorly defined guidelines

Confident Learning (cleanlab):

```
from cleanlab.classification import CleanLearning

# Wrap any sklearn classifier
cl = CleanLearning(RandomForestClassifier())

# Automatically identify and remove noisy labels
cl.fit(X_train, y_train)

# Get indices of likely label errors
```

Consensus Mechanisms: Dawid-Skene Model

Problem: Annotators have different error rates.

Dawid-Skene Model:

- Probabilistic model for aggregating labels
- Learns confusion matrix per annotator
- Estimates true labels given noisy annotations

```
from crowdkit.aggregation import DawidSkene

# Data format: DataFrame with columns ['task', 'worker', 'label']
annotations = pd.DataFrame([
    {'task': 1, 'worker': 'A', 'label': 'spam'},
    {'task': 1, 'worker': 'B', 'label': 'spam'},
    {'task': 1, 'worker': 'C', 'label': 'not_spam'},
    {'task': 2, 'worker': 'A', 'label': 'spam'},
    {'task': 2, 'worker': 'B', 'label': 'not_spam'},
```

Cost-Quality Tradeoffs

Annotation Budget: B dollars

Decision Variables:

- n: Number of samples to label
- k: Annotators per sample
- c: Cost per annotation

Constraint: $n \times k \times c \leq B$

Quality vs Quantity:

- **High n, low k**: More data, less reliable
- **Low n, high k**: Less data, more reliable

Optimal strategy depends on task:

Annotation Guidelines: Best Practices

Key Elements:

1. Clear Definitions:

```
# Spam Classification Guidelines

## Definition
Spam: Unsolicited commercial email sent in bulk.

## Examples
✅
 Spam: "Buy cheap meds now! Click here!"
❌
 Not Spam: Newsletter you subscribed to
⚠️
 Edge case: Promotional email from company you bought from
```

2. Decision Trees:

Measuring Annotation Productivity

Metrics:

1. **Throughput**:

```
throughput = annotations_completed / time_spent_hours
# Target: 50-100 simple labels/hour
```

2. **Learning Curve**:

```
def plot_learning_curve(annotations_df):
    """Plot annotator speed over time."""
    annotations_df['time_per_label'] = (
        annotations_df.groupby('annotator')['timestamp']
        .diff()
        .dt.total_seconds()
    )

    # Group by batch
```

Data Programming: Advanced Patterns

Labeling Function Patterns:

1. **Pattern Matching**:

```
@labeling_function()
def lf_regex_email(x):
    """Detect emails in text."""
    pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b'
    return 1 if re.search(pattern, x.text) else -1
```

2. **External Knowledge**:

```
from nltk.corpus import wordnet

@labeling_function()
def lf_wordnet_positive(x):
    """Use WordNet to detect positive sentiment."""
    positive_words = set(['good', 'great', 'excellent', 'amazing'])
```

Few-Shot Learning for Labeling

Goal: Train model with very few examples (5-10 per class).

Meta-Learning Approach:

```python
from sentence_transformers import SentenceTransformer

# Few-shot classifier using embeddings
class FewShotClassifier:
    def __init__(self, model_name='all-MiniLM-L6-v2'):
        self.model = SentenceTransformer(model_name)
        self.support_embeddings = None
        self.support_labels = None

    def fit(self, support_texts, support_labels):
        """Train on few examples."""
        self.support_embeddings = self.model.encode(support_texts)
        self.support_labels = np.array(support_labels)

    def predict(self, query_texts, k=3):
        """Predict using k-nearest neighbors."""
        query_embeddings = self.model.encode(query_texts)

        predictions = []
        for query_emb in query_embeddings:
            # Compute cosine similarity
            similarities = np.dot(self.support_embeddings, query_emb) / (
                np.linalg.norm(self.support_embeddings, axis=1) *
                np.linalg.norm(query_emb)
            )

            # Get top-k neighbors
            top_k_idx = np.argsort(similarities)[-k:]
```

Prompt-Based Labeling with LLMs

Modern Approach: Use GPT-4 or Claude for labeling.

```python
import anthropic

def llm_label(text, task_description, examples):
    """Use LLM for zero/few-shot labeling."""
    client = anthropic.Anthropic()

    prompt = f"""Task: {task_description}

Examples:
{examples}

Now classify this text:
Text: {text}

Classification:"""

    message = client.messages.create(
        model="claude-3-5-sonnet-20241022",
        max_tokens=10,
        messages=[{"role": "user", "content": prompt}]
    )

    return message.content[0].text.strip()

# Example usage
task = "Classify sentiment as POSITIVE or NEGATIVE"
examples = """
Text: "I love this product!" → POSITIVE
Text: "Terrible experience." → NEGATIVE
"""
```

Synthetic Data Generation

Goal: Generate labeled data programmatically.

Text Augmentation:

```
import nlpaug.augmenter.word as naw

# Synonym replacement
aug_synonym = naw.SynonymAug(aug_src='wordnet')
text = "The movie was great"
augmented = aug_synonym.augment(text)
# Output: "The film was excellent"

# Back-translation
from googletrans import Translator
translator = Translator()

def back_translate(text, intermediate_lang='fr'):
    """Translate to intermediate language and back."""
    # English -> French
    translated = translator.translate(text, dest=intermediate_lang).text
    # French -> English
    back = translator.translate(translated, dest='en').text
```

Transfer Learning to Reduce Labeling

Pre-trained Models: Already learned general features.

Fine-tuning Strategy:

```python
from transformers import AutoModelForSequenceClassification, Trainer

# Load pre-trained model
model = AutoModelForSequenceClassification.from_pretrained(
    'distilbert-base-uncased',
    num_labels=2
)

# Fine-tune on small labeled set (100-1000 examples)
trainer = Trainer(
    model=model,
    train_dataset=small_labeled_dataset,
    eval_dataset=validation_dataset,
)
```

Annotation Tools Comparison

Tool	Type	Strengths	Weaknesses	Cost
Label Studio	Open-source	Flexible, customizable	Setup required	Free
CVAT	Open-source	Video annotation, tracking	CV-focused only	Free
Labelbox	Commercial	Enterprise features, QC	Expensive	$$$
V7	Commercial	Auto-annotation, versioning	Learning curve	$$$
Prodigy	Commercial	Active learning built-in	License per user	$$
Scale AI	Service	Full-service labeling	Least control	$$$$

Recommendation:

- **Prototyping**: Label Studio
- **Production CV**: CVAT or Labelbox

Multi-Task Annotation

Scenario: Annotate multiple attributes simultaneously.

Example (Product reviews):

```json
{
  "text": "Great phone but battery life is poor",
  "annotations": {
    "overall_sentiment": "MIXED",
    "aspects": [
      {"aspect": "phone", "sentiment": "POSITIVE"},
      {"aspect": "battery", "sentiment": "NEGATIVE"}
    ],
    "rating": 3,
    "would_recommend": false
  }
}
```

Benefits:

Handling Edge Cases and Ambiguity

Define Ambiguity Threshold:

```python
# In annotation interface, allow "UNCERTAIN" label
labels = ["POSITIVE", "NEGATIVE", "NEUTRAL", "UNCERTAIN"]

# Later, handle uncertain labels
def resolve_uncertain_labels(annotations):
    """Send uncertain cases to expert annotators."""
    uncertain = annotations[annotations['label'] == 'UNCERTAIN']

    # Send to expert pool
    expert_annotations = expert_annotate(uncertain)

    # Merge back
    annotations.loc[uncertain.index, 'label'] = expert_annotations

    return annotations
```

Escalation Protocol

Annotation Audit Trails

Track Everything:

```
annotation_log = {
    "id": "ann_12345",
    "task_id": "task_789",
    "annotator_id": "user_42",
    "timestamp": "2024-01-15T10:30:00Z",
    "label": "SPAM",
    "confidence": 0.9,  # Annotator confidence
    "time_spent_seconds": 12,
    "annotations_history": [  # If label was changed
        {"timestamp": "2024-01-15T10:29:50Z", "label": "NOT_SPAM"},
        {"timestamp": "2024-01-15T10:30:00Z", "label": "SPAM"}
    ],
    "notes": "Unclear sender but obvious commercial intent"
}
```

Benefits:

Label Taxonomy Design

Hierarchical Labels:

```
Product Review
├── Electronics
│   ├── Smartphones
│   │   ├── Android
│   │   └── iOS
│   └── Laptops
│       ├── Gaming
│       └── Business
└── Clothing
    ├── Men's
    └── Women's
```

Considerations:

- **Depth**: Too many levels = confusion

Domain Adaptation for Labeling

Problem: Labels from one domain don't transfer well.

Example: Sentiment in product reviews vs movie reviews.

Solution: Active learning + Transfer learning.

```
# Train on source domain (movie reviews)
model.fit(X_source, y_source)

# Active learning on target domain (product reviews)
for iteration in range(10):
    # Select uncertain examples from target
    uncertain_idx = uncertainty_sampling(model, X_target_unlabeled)

    # Label selected examples
    y_selected = manual_label(X_target_unlabeled[uncertain_idx])

    # Add to target training set
    X_target_labeled = np.vstack([X_target_labeled,
```

Summary: Annotation Best Practices

Before Labeling:

1. Define clear taxonomy and guidelines
2. Set up quality control (gold standard, IAA)
3. Choose appropriate tools and platform
4. Budget allocation (n samples, k annotators)

During Labeling:

1. Monitor IAA and quality metrics
2. Regular calibration sessions
3. Handle edge cases and ambiguity
4. Track productivity and fatigue

Advanced Techniques Summary

Technique	Use Case	Effort Reduction
Active Learning	Limited budget	50-80%
Weak Supervision	Domain expertise available	70-90%
Semi-Supervised	Large unlabeled pool	40-60%
Few-Shot Learning	Very few examples	90-95%
Transfer Learning	Pre-trained models exist	80-90%
LLM Labeling	High budget, quality needed	60-80%
Synthetic Data	Augmentation tasks	30-50%

Combine techniques for maximum efficiency!

Lab Preview

Today's Goals:

1. **Install Label Studio**.
2. **Config**: Set up a Sentiment Analysis project.
3. **Label**: Annotate 10 examples.
4. **Export**: Get JSON/CSV data.
5. **Analysis**: Write a Python script to calculate Cohen's Kappa between two "simulated" annotators.

Let's start labeling!