Choosing Between Anomaly Detection and Supervised Learning

Deciding between anomaly detection and supervised learning depends on the nature of your dataset, specifically the number of positive examples (y=1) and negative examples (y=0). Here are some considerations to help guide your choice:

1. Anomaly Detection

- **Small Number of Positive Examples:**
  - Suitable when you have a very limited number of positive examples (0-20 is common).
  - Positive examples are mainly used in the cross-validation and test sets for parameter tuning and evaluation.
- **Diverse Anomalies:**
  - Appropriate when there are many different types of anomalies or positive examples.
  - If new types of anomalies may emerge, making it challenging to cover all possibilities with a small set of positive examples.
- **Modeling Normal Examples:**
  - Focuses on modeling normal (negative) examples (y=0) and flags anything deviating significantly from normal as an anomaly.

2. Supervised Learning

- **Larger Number of Positive Examples:**
  - More applicable when you have a relatively larger number of positive examples.
  - Supervised learning assumes that future positive examples are likely to be similar to those in the training set.
- **Similarity to Training Set:**
  - Effective when positive examples in the training set are representative of possible future positive instances.
  - Assumes that the distribution of positive examples in the future will be similar to the distribution observed in the training set.

Choosing Based on Application

- **Examples:**
  - **Financial Fraud Detection (Anomaly Detection):**
    - Many different ways individuals attempt fraud, and new forms emerge frequently.
  - **Email Spam Classification (Supervised Learning):**
    - Types of spam emails may vary, but they often share common characteristics over time.
- **Manufacturing Defect Detection:**
  - **Anomaly Detection:**
    - Identifying new, previously unseen defects in products.
  - **Supervised Learning:**
    - Detecting known defects based on a set of labeled examples.
- **Security Monitoring:**
  - **Anomaly Detection:**
    - Detecting unusual behavior in machines that may indicate a hack.
  - **Supervised Learning:**
    - Used less frequently due to the constantly evolving nature of hacking techniques.
- **Weather Prediction (Supervised Learning):**
  - Limited types of weather conditions, making it suitable for supervised learning.
- **Medical Diagnosis (Supervised Learning):**
  - Predicting diseases based on known symptoms observed in labeled patient data.

Summary

- **Anomaly Detection:**
  - Ideal for scenarios with few positive examples and diverse anomalies.
  - Effective when anticipating new and unpredictable forms of anomalies.
- **Supervised Learning:**
  - Suitable when a larger number of positive examples is available.
  - Assumes future positive instances will be similar to those in the training set.

Understanding the characteristics of your data and the potential diversity of anomalies will help you make an informed decision between anomaly detection and supervised learning for your specific application.