1. 異常検知の概要

1.1 異常・異常検知タスクとはなにか

異常とは、正常データの分布以外から発生したデータのことです。また、興味のあるデータ点が正常か異常かを判断するタスクを異常検知タスクといいます。実社会だと、

- クレジットカードの不正検知
- SNSにおけるスパム検知
- 産業システムにおける部品の故障・劣化・損傷検知
- 医用画像解析・脳波記録におけるレアパターンの検知

などが異常検知タスクの代表例です。

ところで、異常検知タスクは「正常」と「異常」という2種類のカテゴリがあることから、2値分類タスクの一種だと思われるかもしれません。しかし、

- 2値分類の場合、厳密には多岐にわたる全ての種類の「異常」を訓練データに揃える必要があります。しかし、例えばクレジットカードの不正検知やSNSにおけるスパム検知では、不正やスパムの振る舞いに一貫したパターンがないと考えられています。
- また、異常は稀にしか発生せず、データ点として少数しか準備できないことがあります。産業システムにおける部品の故障や医用画像解析・脳波記録におけるレアパターンは、大抵の場合に少数しか異常を準備できません。このような問題は、2値分類だと不均衡データの問題として知られ、難しいタスクであることが知られています。

以上の観点から、異常検知では2値分類以外のアプローチがさまざま検討されてきました。最も古典的なアプローチは

- 異常検知モデルの入力 = （ほとんどが）正常からなるデータ
- 異常検知モデル = 正常データの分布 $p^+(x)$ の推定量
- 異常検知モデルの出力 = データ点 x の正常データにおける起こりやすさ（例えば $-\log p^+(x)$ ）の推定値

です。このアプローチは、**教師なし異常検知**とよばれています。また、このような異常検知モデルの出力を**異常度**といいます。あるデータ点が異常かどうかを判定するときは、あらかじめ閾値 t を決めておき、t より大きければ異常、小さければ正常と判定します。閾値 t を決めるときには、1.2節で説明する異常検知モデルの「誤り」を考慮して決めます。

1.2 異常検知モデルによる判定の誤りと評価

異常検知モデルによる異常の判定には、2種類の誤りが考えられます。

正解\判断	異常	正常
異常		**第2種の誤り**
正常	**第1種の誤り**	

第1種の誤りと第2種の誤りのどちらが重要かは、異常検知タスクの目的に応じて異なります。クレジットカード不正の検知を例に考えてみましょう。

1. 不正の疑いのあるアカウントを停止したい場合：目標は、不正を行ったと予測されるアカウントを改めて人が調べ、実際に不正を行ったと断定できた場合に、そのアカウントを停止することです。人手で調べられるアカウントの数には限りがあるので、不正を行ったと予測されたアカウントは、実際に不正を行なっていることが強く望まれます。つまり、第1種の誤りを避けたいと言い換えることができます。
2. 詐欺に用いられている疑いのあるアカウントの顧客に自動で通知を行いたい場合：目標は、不正を見落とすことで、顧客からの不信を買うような事態を防ぐことです。このとき望まれることは、実際に不正を行なっているアカウントに対して、なるべく正確に不正と予測することでしょう。つまり、第2種の誤りを避けたいと言い換えることができます。

異常検知モデルを評価するときには、正常なデータ点と異常なデータ点を含むテストデータを準備し、

- **適合度**（precision）＝異常と判断したもののうち実際に異常だったものの割合
- **再現率**（recall）＝実際に異常だったもののうち異常と判断したものの割合

を評価します。適合度は第1種の誤り、再現率は第2種の誤りの評価に対応します。第1種の誤り・第2種の誤りは異常検知モデルの異常度の閾値 t の設定によってコントロールすることができます。t を大きくすると適合度は大きくなりますが、再現率が下がります。逆に t を小さくすると再現率は大きくなりますが、適合度が小さくなります。