

# “精度評価” Is All You Need

## はじめに

MLの文脈での「評価」は

- オフラインテスト
- オンラインテスト

に分けられると思います。

今回の対象は前者の「オフラインテスト」です。

## 評価指標

機械学習モデルやそれを扱うシステムの開発時に最適化の対象になるものとして以下の三つが挙げられる

- 目的関数(損失関数)
- 評価指標
- KPIなどのビジネス的な数字

## 目的関数

機械学習モデルの「学習」とは「目的関数を最小化する関数を見つけること」とも言い換えられる。

現段階では最適化アルゴリズムに勾配を利用したものが一般的であるため、多くのニューラルネットの学習では微分可能である必要がある。

## 評価指標

追すべき数字を定式化しモデルの評価を可能にしたもの。

一般的な指標を使用してもいいし、複雑性が高いのであれば自分で定式化することも考える必要が出てくる。

微分可能である必要はない。課題の構造が単純で滑らかな関数で表せるのであれば、目的関数と共に構わない。

## KPIなどの数字

向上させることがゴールになるもの

売上でなくても良いと思います。ユーザーにとっての価値（ECのレコメンドで言えばどれくらい実際に買ってくれるかとか、AIリストならユーザーがその中から実際どれだけ契約まで辿り着いたのか）とか。

ここが改善しないのであればどれだけ最先端のモデル・システムでもビジネス的には価値を生まない。もちろん技術的な挑戦・蓄積という面も見るべきではあると思う。

## なぜ評価指標を置くのか

- そもそも評価指標を置いておらず、本当に価値を提供できているのか不透明
- 指標が適切でないため、開発者側とユーザーとの間で体感に差が出る
- 評価指標が改善する追加開発を行なっても肝心のユーザーエクスペリエンスやKPIが変化しない

## いい評価指標とは

ビジネス空間に存在する課題・タスクから技術的な性能を示す空間への写像が的確に対応していること。

## なぜモデル開発ではなくAIサービスでも重要なのか

ブラックボックスを適切に評価する上で必要不可欠であるから。

機械学習モデルの中でもLLMは特に説明可能性が低く、評価指標がほぼ唯一の客観的なビジネス要件への適合度や品質保証の判断基準になる。ロジックを組む場合であれば、Aという入力であればBになります、という判断ができるがMLはそうではない。

ベンチマークでの精度向上を謳う新モデルはLLMユーザーの開発したサービス・アプリケーションなどの個別タスクの精度向上を保証するものではない。

汎用的なタスクをこなせるモデルの登場によってより複雑なタスクを任せられるようになつた一方で、有用性・倫理観・自然かどうか、のような曖昧な概念を扱う必要が生まれ、構造を定式化したり評価指標を置くことの難易度が上がる。

場合によってはHuman In The LoopやLLM-as-a-Judgeのような形をとることも考えなければならないが、コストがかかる上に評価にブレが生じやすい。定式化した評価指標・HITLやLLM-as-a-Judgeによる評価・それぞれにかかるコストなど、評価プロセスの設計そのものにまで話が膨らむ。

**サービスとしてMLモデルを使用する場合、損失としての目的関数を意識することは少なくなる（多くの場合意識する必要すらなくなる）が、LLMの登場により扱うタスクが複雑になっており、適切な評価指標を置く難易度が上がる**

---

## 適切な性能評価のためのデータ分割

用意したデータを全てモデル・システムの開発に使用してはいけない。

- モデルの開発が必要な場合
  - 学習用データ
    - Training
    - Validation
  - テスト用データ（評価用データ）
- モデルの開発が不要な場合
  - ロジック・プロンプトの開発に使用するデータ
  - テスト用データ（評価用データ）

いずれにせよ、テスト用のデータを開発が始まる時点で用意しておく必要がある。

全体のデータを分割した後、テスト用データは基本的に中身を確認してはいけない。

もう一方のデータでモデル・システム開発を行い、マスクしたテスト用データで評価指標がどのように変化するか確認する。

学習用・開発用データと同じ精度が出る→汎用的なモデル・システム開発ができる

学習用・開発用データと比べて精度が低い→過学習したモデル・開発データに特化しそうなプロンプトになっている可能性がある

## **ML/AI系のプロジェクトが始まったら、まずデータを二つに分ける。片方は見ない。**

適当に分けていいわけではない。

例えば企業関連のデータを二つに分ける場合、片方に大企業が多めに入ってしまった場合、LLMは知識として大企業の情報を保持しているため、Fine-TuningしたモデルやRAGを使用したシステムの適切な評価ができない。

## **評価指標が完璧でも用意したデータ次第で信頼性が損なわれてしまう。**

例えば上であげた企業関連データの分割の場合、上場企業とそうでない企業のフラグなどを使用して、それぞれの分布の比率を維持して分割する方法（stratified k-fold）を使用するなどの対策方法があります。

Trainig/Validationデータの分割方法を意味しますが考え方は同じなので、「交差検証」などで調べるとデータの分割方法について学ぶことができると思います。

さらにテストデータでの評価をあまり繰り返してしまうと、意図せずテストデータへの過適合を起こすこともよく問題になる。データが潤沢にあるかなどを考慮しながら、厳格かつ柔軟に判断する必要がある。

**評価指標は機械学習・データ分析をするから  
必要なわけではない。  
ブラックボックスが含まれる処理を客観的に  
評価するために必要なのであって、MLモデ  
ルはブラックボックスの一例にすぎない。**

**ブラックボックスを評価する力が大切！**

例え話としてのソフトウェアベンチ