評価関数の重要性

- 評価関数とは
  - 評価関数
    > モデルの性能（予測精度）を定量的に表すための関数
  - 評価関数の値が改善するようにモデル開発を行う
- 評価関数の重要性
  - 予測精度を定量評価する意義
    - モデルの予測結果がどれほど信頼できるのか把握することで、実際の運用が可能か判断する
    - 予測精度とKPIが関連付けられれば、ビジネスインパクトやROI(費用対効果)を算出することもできる
- タスクによって異なる評価関数
- 目的やテーマに合わせて、適切な評価関数を設定する必要がある

1. 回帰：売上・需要予測、リスク予測等。実測値と予測値のずれの評価が必要
   > 代表的な評価関数；MAE、RMSE、MAPE、R２(決定係数)
2. 分類：画像分類、不正・故障検知等。正しいラベル付けができているかの評価が必要
   > 代表的な評価関数；Accuracy、Precision、Recall、F値、AUC
3. 推薦/検索：検索エンジン、推薦エンジン。ユーザーの嗜好と適合(マッチ)しているかの評価が必要

> 代表的な評価関数；mAP@N、nDCG

4. 物体検出：顔検出、通行人検出等。物体の位置や領域の整合性の評価が必要。

> 代表的な評価関数：IoU、mAP

- 回帰問題における評価関数
  - MAE (Mean Absolute Error)
    - 実測値と予測値の「誤差の絶対値」を元に算出する手法
    - 平均絶対誤差とも呼ばれ、値が小さい(0に近い)ほど予測精度が高い
    - 後述のRMSEとともに、回帰問題で一般的に使用される
  - RMSE (Root Mean Squared Error)
    - 実測値と予測値の「誤差の2乗」を元に算出する手法
    - 平均平方二乗誤差とも呼ばれ、値が小さい(0に近い)ほど予測精度が高い
  - MAEとRMSEの計算結果の比較
    - RMSEは大きな外れ値があると悪い評価となる
  - MAEとRMSEで評価する課題例
    - MAE：誤差を平均的に評価

        - RMSE：局所的な誤差も評価
    - MAPE (Mean Absolute Percentage Error)
        - 実測値に対する「誤差の割合(誤差率)」を元に算出する指標
        - 平均絶対パーセント誤差とも呼ばれ、値が小さい(0に近い)ほど予測精度が高い(通常はパーセント[%]で精度を表現する)
        - スケールが異なるデータの誤差を比較しやすい
        - ただし、実測値に"0"が含まれる場合は使用できない
    - R2 (決定係数：coefficient of determination)
        - 「データそのもののばらつき(分散)」と「予測値のずれ」を元に算出する指標
        - モデルの予測値のあてはまり度合い（適合性）を評価
        - 値が1に近いほど予測精度が高い
        - MAEやRMSEと違い、あくまで相対的な指標であるため、決定係数から直接誤差の大きさをはかることはできない

- 分類問題における評価関数

    - 混同行列 (Confusion Matrix)
        - 分類結果(正しく分類した数/誤って分類した数)をまとめた表のこと
        - 特に2値分類(0か1かの分類)のモデルの性能を測る指標として使用される(2x2の表形式で作成する)
        - 次の4つの区分が、評価関数を算出するための基礎となる
            1. TP(True Positive)：Positiceなものを正しくPositiveと予測した数
            2. FN(False Negatove)：Positiveなものを誤ってNegativeと予測した数
            3. FP(False Positive)：Negativeなものを誤ってPositiveと予測した数
            4. TN(True Negative)：Negativeなものを正しくNegativeと予測した数
    - Accuracy (正解率)
        - 予測結果が実際にあたっていた割合
        - (TP+TN) / (TP+FP+FN+TN)
        - 単純に正解率だけではモデルの性能を評価できない
    - Precision (適合率)
        - 陽性であると予測したもののうち、実際に陽性だった割合
        - TP / (TP+FP)
        - 誤判定を避けたい場合は、Precisionを重視
    - Recall (慎陽性率、再現率)
        - 実際に陽性であるもののうち、正しく要請であると予測できた割合
        - TP / (TP+FN)
        - 見落としを避けたい場合は、Recallを重視
    - PrecisionとRecallはトレードオフの関係
        - 2値分類のモデルでは、陽性(Positive)である確率が、ある基準(閾値)より大きいか、小さいかで分類を行う
        - 閾値が高い⇒Precisionは高いが、Recallは低い
        - 閾値が低い⇒Recallは高いが、Precisionは低い
    - F値 (F1-score、F-measure)
        - PrecisionとRecallの両方を加味した評価関数
        - 0から1の値を取り、値が大きいほど予測精度が高い
        - F値が高いモデルは、PrecisionとRecallのバランスがとれた良いモデル
    - ROC曲線 (Receiver Operating Characteristic)

        - 2値分類問題で、閾値を変化させたときに、モデルの性能がどのように変わるかを可視化する手法
        - 閾値を変化させたときに、真陽性率(Recall)と偽陽性率の組み合わせがどのように変化するかを曲線で表現する
        - 偽陽性率(FPR)=FP / (FP+TN)
    - AUC (Area Under the Curve)
        - ROC曲線の下の領域面積
        - ROC曲線全体を一つの値で示せる
        - 2値分類の評価指標
        - 0～1の値を取り、値が大きいほど良い
        - AUC=0.5：ランダムに判断したのと同じ(予測が役に立っていない)
        - AUC=1：理想だが、現実的にはほぼあり得ない
        - AUCが1.0にどれだけ近いかで、モデルの性能の良し悪しを判断する

- 推薦・検索問題における評価関数
    - 適合アイテム推薦

        > 推薦したアイテムがユーザに適合するかどうか

    - 評価値予測

        > 予測したアイテムをユーザがどう評価するか

    - Precision@k
        - 推薦上位k件に占める、適合アイテムの割合
    - Recall@k
        - すべての適合アイテムに占める、推薦度上位k件に含まれる適合アイテムの割合
    - AP (Average Precision)
        - 適合アイテムが得られた時点での適合率の平均をとった値
    - MAP (Mean Average Precision)
        - ユーザごとに計算したAPの平均をとった値
    - DCG (Discounted Cumulative Gain)
        - 推薦したアイテムの関連度に重みをつけて合計する
        - 関連度はユーザがアイテムにどの程度適合しているかを表す数値
        - 下位に順位付けしたアイテム程重みが減衰する
    - nDCG (normalized Discounted Cumulative Gain)
        - DCGを理想的な順位付けを行ったときのDCGperfectで割った値
        - DCGを0～1の値をとるように正規化したもの
    - 評価値予測の評価関数
        - RMSEやMAEなどが用いられる

- 物体検出問題における評価関数
    - 物体検出アルゴリズムの出力
        - クラスの位置情報

            > どこに写っているか

        - クラスとラベルの信頼度

> 何が写っているか その信頼度も一緒に出力

- IoU (Intersection over Union)
  - 検出領域都心の領域の一致度を測る指標
- mAP
  - マルチクラスの物体検出問題において、各クラスについて算出したAverage Precisionの平均値をmAPとする
  1. 検出の成否を判定(IoUが閾値を超えているか)
  2. Precision@kを計算
  3. Average Precisionを計算