Visual Mining Studio Tips 集

ROC 曲線を描画し AUC を求める

ここで使う！	予測における判断の直接の根拠となる量、の良し悪しを評価する 1 つの方法として ROC 曲線というものが知られています。VMS では、Tree & Random Forest や Neural Network のアイコンで分類モデル[※1]を作成するとカテゴリ毎の予測確率値を出力することができます。このとき「予測における判断の直接の根拠となる量」はその予測確率値になりますが、「予測」アイコンでは、この予測確率値と予測結果をもとに、ROC 曲線を作成するための情報を出力することができます。これを用いて、より詳細な分類モデルの評価を行うことができます。 [※1] 分類モデル、すなわち目的変数に文字列を指定している場合に有効です。また、Tree & Random Forest, Neural Network といった、予測確率値が出力されるタイプの「モデリング」系アイコンを利用する必要があります。

ROC 曲線とは？

予測確率値は 0 と 1 の間の値をとります。(異常, 正常) のような 2 値の予測を行う場合、通常は確率値が 0.5 より大きければ異常、小さければ正常と判断するかと思います。ROC 曲線[※2]はこの確率値そのものがどの程度信頼がおけるものであるかを図示します。

例えば、0.5 ではなく、かなり厳しく判定し「0.95 以上のものだけを異常とみなす」と決めたとすると、確率値そのものが実情を反映したものであれば、本当に正常であるデータが異常である予測されることは非常に少なく、その逆に、異常と判定されるデータは本当に異常であるデータの中からほぼ選ばれるはずです。

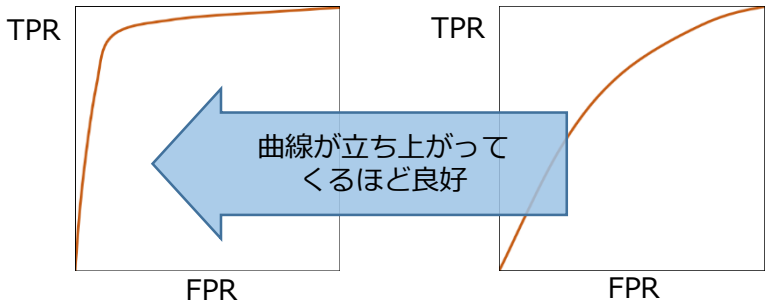
ここで、今の 0.95 のようにある判定の確率値を固定したとき、元来の異常データの中で、異常と予測される割合を「TPR」[※3]、元来の正常データの中で、異常と予測される割合を「FPR」[※3]と呼びます。

[※2] Receiver Operating Characteristic の略で、日本語では「受信者操作特性」などと表記されます

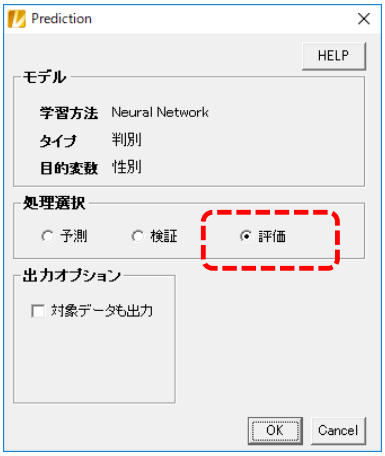
[※3] それぞれ True Positive Rate, False Positive Rate の略

ここで、確率値判定の条件を 0.95 から徐々に緩めていきます。「異常確率が高い範囲に、本当に異常データが集中している」ならば、確率値を緩めていってもやはり異常データは捕捉できるはずなので TPR の値は上昇していき、また正常データが異常と予測されることはあまりないはずなので FPR の値はあまり上昇しないはずです。逆に、「異常確率が高い範囲に、正常データが混じってしまっている」ような思わしくない状態であれば、値を緩めたときに TPR の値はあまり上昇せず、FPR の方が上昇してしまうことになります。

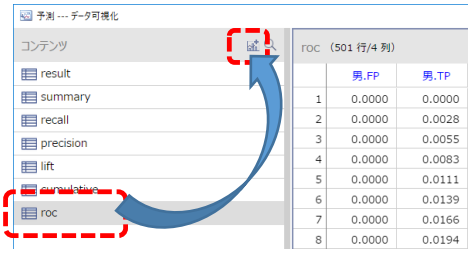
なので、確率値判定の基準を緩めた際に、TPR が早く上昇すればそれは良いモデル、そうでなく FPR の方が早く上昇すれば好ましくないモデル、ということになります。この状況は、TPR を縦軸、FPR を横軸にとり、判定値を緩めていった際の TPR, FPR の値を順次プロットしていくことで視覚的に確認することができます。これが ROC 曲線です。最も判定基準が厳しい場合は全データが「正常」なので、TPR も FPR もゼロです。最も判定基準が緩い場合は全データが「異常」なので TPR も FPR も 1 です。なので、ROC 曲線は（FPR, TPR）=最も厳しい場合（0,0）と最も緩い場合（1,1）を結ぶ曲線になり、この形で分類モデルの評価を行うことができます。



どうする？



ROC 曲線を描画するための情報は、「予測」アイコンでオプション「評価」を選択することで出力されるようになります。

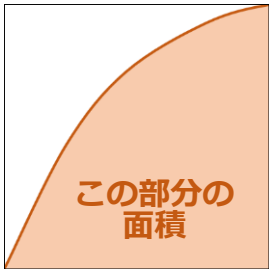


結果のテーブルから「ROC」を選択し、新規グラフ作成ボタンを押します。

グラフ設定で 折れ線グラフ を選択し、X 軸に「予測対象の値.FP」、Y 軸に「予測対象の値.TP」を指定して「OK」を押すことで ROC 曲線が描画できます。

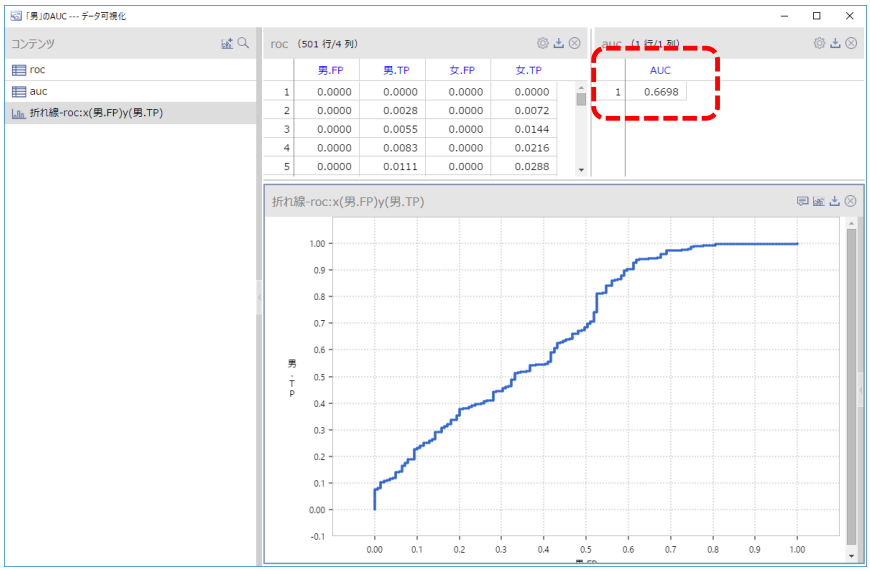
ROC 曲線は「予測対象のあるカテゴリ値」1 つについて考えるものです。X 軸・Y 軸に、同じ名前の FP, TP の列をそれぞれ指定することにご注意ください。

AUC 値とは？



ROC 曲線から得られる特性を簡単に比較する値として、AUC 値[※4]という指標があります。これは、ROC 曲線から下の部分の面積を示す量で、この値が大きいほど（1 に近いほど）予測モデルの評価は良好であるということがいえます。

[※4] Area Under Curve の略

どうする？	サンプルプロジェクトに、VAP スクリプトを用いて AUC 値を求めるアイコン「AUC 値を求める」を追加いたしました。男・女の 2 値を持つ属性「性別」について、「男性」を予測する場合の AUC 値を求めています。  スクリプト内の ``` // AUC（ROC曲線の下の面積）を求める列名を指定して下さい。 targetColName = "男"; ``` 上記の部分を修正すれば、一般の「予測」アイコンの出力「roc」に対応可能ですので、是非参考にしてください。

本文書について

本文書は、（株）NTT データ数理システム（以下「弊社」）が開発・販売するデータマイニングツール **Visual Mining Studio** のユーザーに対する情報提供として弊社が作成を行ったものです。弊社による事前の許可なしに、本文書の再配布や引用の範囲を超える複製・改変といった行為を禁じます。

本文書は、下記の URL よりダウンロード・閲覧が可能です。

http://www.msi.co.jp/vmstudio/tips05_roc_and_auc.pdf

お問い合わせ

株式会社 NTT データ数理システム Visual Mining Studio 担当
〒160-0016 東京都新宿区信濃町 35 番地 信濃町煉瓦館 1 階
TEL 03-3358-6681 FAX 03-3358-1727
<e-mail> vmstudio-info@msi.co.jp
<URL> http://www.msi.co.jp/vmstudio/