**XÁC SUẤT - THỐNG KÊ**

1. THỐNG KÊ
   1. Mean
   2. Median
   3. Quantile
   4. Mode
   5. Normal distribution (Phân phối chuẩn)

Phân phối chuẩn (Tiếng Anh: normal distribution) còn gọi là phân phối Gauss hay (Hình chuông Gauss), là một phân phối xác suất cực kì quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số vị trí (giá trị trung bình μ) và tỉ lệ (phương sai σ2).

Phân phối chuẩn tắc (standard normal distribution) là phân phối chuẩn với giá trị trung bình (μ) bằng 0 và độ lệch chuẩn (σ) bằng 1. Phân phối chuẩn còn được gọi là đường cong chuông (bell curve) vì đồ thị của mật độ xác suất có dạng chuông.

**Đặc tính của phân phối chuẩn**

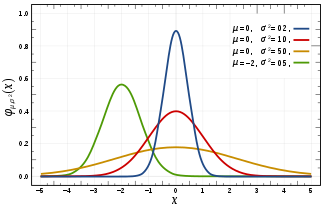
Có nhiều cách để thể hiện các đặc tính của một phân phối xác suất. Cách dễ thấy nhất là thông qua hàm mật độ xác suất (vẽ ở hình đầu tiên), nó cho biết khả năng xảy ra của mỗi giá trị của biến ngẫu nhiên. Hàm phân phối tích lũy cũng cho cùng thông tin, nhưng hình ảnh của nó thì thông tin chứa đựng không được dễ nhận thấy cho lắm (hình đi sau). Cách tương đương khi chỉ định một phân phối chuẩn là thông qua: mômen, ước lượng, hàm đặc trưng, hàm khởi tạo mômen, và hàm khởi tạo ước lượng và định lý Maxwell. Một số rất hữu ích về mặt lý thuyết, nhưng không trực quan. Xem phân phối xác suất.

Mọi ước lượng của phân phối chuẩn đều bằng 0, ngoại trừ hai cái đầu tiên.

*https://vi.wikipedia.org/wiki/Ph%C3%A2n\_ph%E1%BB%91i\_chu%E1%BA%A9n#:~:text=Ph%C3%A2n%20ph%E1%BB%91i%20chu%E1%BA%A9n%20(Ti%E1%BA%BFng%20Anh,(ph%C6%B0%C6%A1ng%20sai%20%CF%832).*

**Phân phối chuẩn:**

**a. Hàm mật độ xác suất**



* 1. Biểu đồ tần suất (frequency histogram)
  2. Kiểm tra Shapiro-Wilk
  3. Biểu đồ Q-Q (quantile-quantile plot)

1. **XÁC SUẤT**

BÀI TẬP TOÁN

1. **KIỂM ĐỊNH MÔ HÌNH**
   1. **ROC**

ROC (Receiver Operating Characteristic) là một biểu đồ được sử dụng để đánh giá hiệu suất của một mô hình phân loại (Có nghĩa là mô hình phải là nhãn 1,0 thì mới có thể sử dụng), đặc biệt là trong các bài toán với các tập dữ liệu không cân bằng. Biểu đồ ROC được sử dụng rộng rãi trong thống kê và học máy để đánh giá khả năng phân biệt của một mô hình.

Thông thường các mô hình Scorecard với nhãn gắn cho khách hàng là 0,1 tức rủi ro hoặc không thường được sử dụng ROC để đánh giá hai mô hình.

**Các thành phần chính của biểu đồ ROC:**

1. **True Positive Rate (TPR)**: Còn gọi là Sensitivity hoặc Recall, TPR là tỷ lệ giữa số lượng True Positives (TP) trên tổng số lượng của các Positive samples. Công thức:

**A black background with white text

Description automatically generated**​

Trong đó FN là False Negatives.

1. **False Positive Rate (FPR)**: FPR là tỷ lệ giữa số lượng False Positives (FP) trên tổng số lượng của các Negative samples. Công thức:

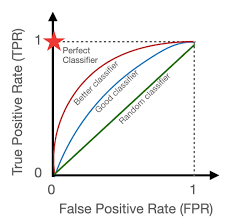
**A math equation with white text

Description automatically generated**​

Trong đó TN là True Negatives.

**Biểu đồ ROC:**

* Trục X: False Positive Rate (FPR).
* Trục Y: True Positive Rate (TPR).



**Đường cong ROC:**

* Đường cong ROC vẽ biểu đồ TPR (Sensitivity) trên trục y so với FPR (1-Specificity) trên trục x tại các ngưỡng phân loại khác nhau.
* Một mô hình phân loại tốt sẽ có đường cong ROC gần với góc trên bên trái của biểu đồ.
* Một mô hình có đường cong ROC là đường chéo từ góc dưới bên trái đến góc trên bên phải biểu thị một mô hình phân loại ngẫu nhiên.

**Diện tích dưới đường cong ROC (AUC - Area Under the Curve):**

* AUC là một giá trị tóm tắt của biểu đồ ROC, biểu thị khả năng phân biệt của mô hình. AUC dao động từ 0 đến 1.
* AUC = 1: Mô hình hoàn hảo.
* AUC = 0.5: Mô hình ngẫu nhiên, không có khả năng phân biệt.
* AUC < 0.5: Mô hình kém hơn cả ngẫu nhiên, thường do sự cố hoặc mô hình bị lỗi.

**Ví dụ minh họa:**

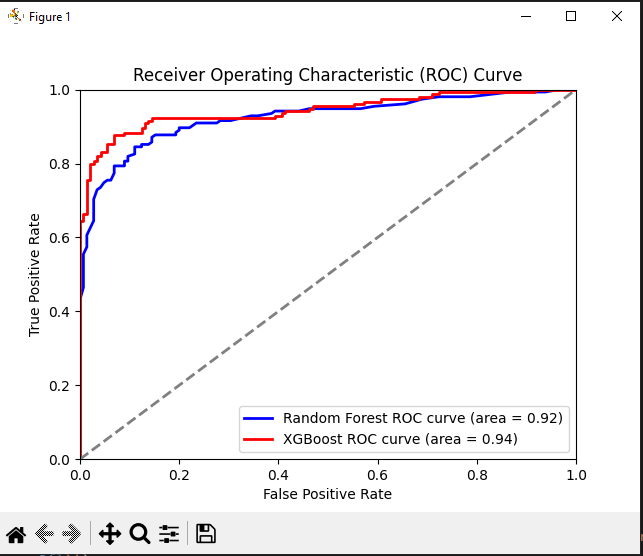
Cách sử dụng ROC và AUC trong mô hình Scorecard Profit Base

1. **Huấn luyện mô hình**: Sử dụng các thuật toán như XGBoost và Random Forest để huấn luyện mô hình dự đoán trên dữ liệu huấn luyện.
2. **Dự đoán xác suất**: Đối với mỗi khách hàng, mô hình sẽ dự đoán xác suất của việc khách hàng đó sẽ trả nợ đúng hạn.
3. **Tính toán TPR và FPR**: Thay đổi các ngưỡng dự đoán (ví dụ, từ 0 đến 1) và tính toán các giá trị TPR và FPR tương ứng.
4. **Vẽ đường cong ROC**: Sử dụng các cặp giá trị TPR và FPR để vẽ đường cong ROC. Đường cong này sẽ cho biết khả năng phân biệt của mô hình tại các ngưỡng khác nhau.
5. **Tính toán AUC**: Tính diện tích dưới đường cong ROC. Giá trị AUC sẽ giúp đánh giá hiệu suất tổng thể của mô hình. Mô hình nào có AUC cao hơn sẽ được ưu tiên vì nó có khả năng phân biệt tốt hơn giữa các khách hàng sẽ trả nợ và không trả nợ.
6. **Lựa chọn mô hình**: So sánh AUC của các mô hình khác nhau (ví dụ, XGBoost và Random Forest) để chọn mô hình tốt nhất. Mô hình có AUC cao hơn thường được coi là mô hình tốt hơn.

**Ví dụ cụ thể bằng Python**

Dưới đây là một ví dụ đơn giản về cách tính và vẽ ROC, AUC sử dụng thư viện sklearn với mô hình XGBoost và Random Forest:

|  |
| --- |
| import numpy as np  import matplotlib.pyplot as plt  from sklearn.metrics import roc\_curve, auc  from sklearn.ensemble import RandomForestClassifier  from xgboost import XGBClassifier  from sklearn.model\_selection import train\_test\_split  from sklearn.datasets import make\_classification  # Tạo dữ liệu giả lập  X, y = make\_classification(n\_samples=1000, n\_features=20, random\_state=42)  # Chia dữ liệu thành tập huấn luyện và tập kiểm tra  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)  # Huấn luyện mô hình Random Forest  rf = RandomForestClassifier()  rf.fit(X\_train, y\_train)  rf\_probs = rf.predict\_proba(X\_test)[:, 1]  # Huấn luyện mô hình XGBoost  xgb = XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss')  xgb.fit(X\_train, y\_train)  xgb\_probs = xgb.predict\_proba(X\_test)[:, 1]  # Tính ROC và AUC cho Random Forest  rf\_fpr, rf\_tpr, \_ = roc\_curve(y\_test, rf\_probs)  rf\_auc = auc(rf\_fpr, rf\_tpr)  # Tính ROC và AUC cho XGBoost  xgb\_fpr, xgb\_tpr, \_ = roc\_curve(y\_test, xgb\_probs)  xgb\_auc = auc(xgb\_fpr, xgb\_tpr)  # Vẽ ROC Curve  plt.figure()  plt.plot(rf\_fpr, rf\_tpr, color='blue', lw=2, label=f'Random Forest ROC curve (area = {rf\_auc:.2f})')  plt.plot(xgb\_fpr, xgb\_tpr, color='red', lw=2, label=f'XGBoost ROC curve (area = {xgb\_auc:.2f})')  plt.plot([0, 1], [0, 1], color='grey', lw=2, linestyle='--')  plt.xlim([0.0, 1.0])  plt.ylim([0.0, 1.0])  plt.xlabel('False Positive Rate')  plt.ylabel('True Positive Rate')  plt.title('Receiver Operating Characteristic (ROC) Curve')  plt.legend(loc="lower right")  plt.show() |



Trong ví dụ này:

* Tạo một tập dữ liệu giả lập.
* Huấn luyện hai mô hình (Random Forest và XGBoost).
* Dự đoán xác suất trên tập kiểm tra.
* Tính toán các giá trị TPR và FPR để vẽ đường cong ROC.
* Tính toán AUC để đánh giá hiệu suất mô hình.

Mô hình XGBoost có AUC cao hơn được coi là mô hình tốt hơn để sử dụng trong dự đoán phê duyệt tín dụng khách hàng.

* 1. **MEAN ABSOLUTE ERROR (MAE)**

MAE đo lường độ lệch trung bình của các dự đoán so với các giá trị thực tế:

A black background with white text

Description automatically generated

* 1. **MEAN SQUARE ERROR (MSE)**

**MSE đo lường độ lệch trung bình của các dự đoán so với các giá trị thực tế, nhưng trọng số lớn hơn được đặt cho các sai lệch lớn hơn:**

****

* 1. **R-SQUARE (R2)**

R² đo lường tỷ lệ phần trăm của tổng biến động trong giá trị thực tế có thể được giải thích bằng mô hình:

A black background with white text

Description automatically generated