Naive Bayes Classifier

Explainable Machine Learning – Fabian Keck, Eva Fiserova, Natalie Martin

Wissenschaftliche Kriterien von Explainability – *Was macht ein Modell transparent?*

Simulierbarkeit

Das Bayes'sche Modell ist für die Menschen leicht zu verstehen und zu kopieren.
Bei kleineren Daten können Menschen bedingte Wahrscheinlichkeiten von Klassen und dessen Ausprägungen errechnen und Vorhersagen treffen.

Zerlegbarkeit

Die Aufschlüsselung des Modells in verständliche Teile, wie Eingaben und Berechnungen, hilft bei der Erklärung seiner Funktionsweise.
Der Datensatz kann in seine einzelnen Merkmale zerlegt werden. Wahrscheinlichkeiten können für einzelne und mehrere Merkmale berechnet werden.

Algorithmische Transparenz

Das Modell zeigt uns genau, wie es Entscheidungen trifft, so dass der Prozess nachvollziehbar ist.
Die fit()-Funktion von Scitkitlearn verhindert Einblicke in den Trainingsprozess, allerdings können Wahrscheinlichkeiten abgefragt werden.

Vorgehen & Accuracy

Es wurden Methoden angewendet, um die Accuracy zu verbessern.

Nur Gaussian Naive	79,84%
Reduziert auf die wichtigsten Merkmale	79,97%
Mit Hyperparametertuning	82,72%
Mit Principal Component Analysis	80,69%

Der Datensatz wurde auf die wichtigsten Merkmale reduziert.

Bayes Theorem & Bayes Classifier

Das Bayes Theorem basiert auf dem Konzept der bedingten Wahrscheinlichkeiten.

Posterior **Likelihood** **Prior Class**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$ ← **Prior Feature**

Der Naive Bayes Classifier verwendet dieses Theorem, um im Trainingsprozess die Wahrscheinlichkeit für die Zugehörigkeit zu einer Klasse A für eines bestimmtes Merkmales B zu bestimmen. Eine Prämisse hierfür ist die Unabhängigkeit der Merkmale.

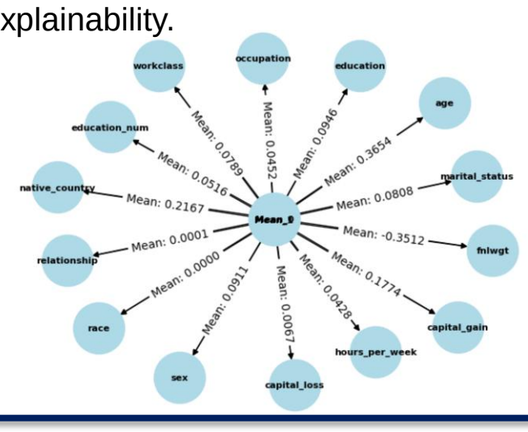
Das Theorem im Trainingsprozess:
1. Wahrscheinlichkeit der Klasse P(A)* – **Prior Class**
2. Inverse bedingte Wahrscheinlichkeit P(B|A)* – **Likelihood**
3. Wahrscheinlichkeit des Merkmales P(B)* – **Prior Feature**
4. Berechnung der gesuchte Wahrscheinlichkeit P(A|B) – **Posterior**

*anfängliche Wahrscheinlichkeit (Historische Daten, Annahmen)

Dataset

Das verwendete *adult income dataset* umfasst demografische und berufliche Daten wie Alter, Bildungsniveau und Beschäftigungsstatus zur Einkommensvorhersage.

Der *Directed Acyclic Graph (DAG)* veranschaulicht durchschnittliche Merkmalswerte zum Einkommen, erklärt Einflussfaktoren und unterstützt die Explainability.



Bayes vs. EBM

Explainable Boosting Machine (EBM) ist ein baumbasiertes, zyklisches Gradient Boosting Generalized Additive Model. Es ist ein Glassbox-Modell mit hoher Genauigkeit, das mit den modernsten Methoden des maschinellen Lernens wie Random Forest vergleichbar ist, und gleichzeitig hochgradig intelligent und erklärbar ist.
Als erklärbares Modell wird dieses zum Vergleich herangezogen.

$P(A\|B) = \frac{P(B\|A) \cdot P(A)}{P(B)}$	$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$
Accuracy bei Dataset: 80%	Accuracy bei Dataset: 87%
Verteilung der Daten kann mit sns.countplot analysiert werden, für jede einzelne Variable	Funktion Class Histogram: Analyse der Verteilung der Daten
Marginal Plot mit seaborn Funktion und Berechnung des Pearson-Korrelationskoeffizienten, für jede einzelne Variable	Funktion Marginal: Verteilung der Datenmerkmale, Berechnung des Pearson-Korrelationskoeffizienten
Feature Importance kann mit plt gezeigt werden, Beitrag, den jedes Merkmal zu den Vorhersagen leistet, nicht	Explain.global(): zeigt Feature Importance und Beitrag, den jedes Merkmal zu den Vorhersagen leistet
Keine ähnliche Funktion	Explain.local(): zeigt zeigt die Aufschlüsselung, wie viel jeder Begriff zur Vorhersage für eine einzelne Stichprobe beiträgt

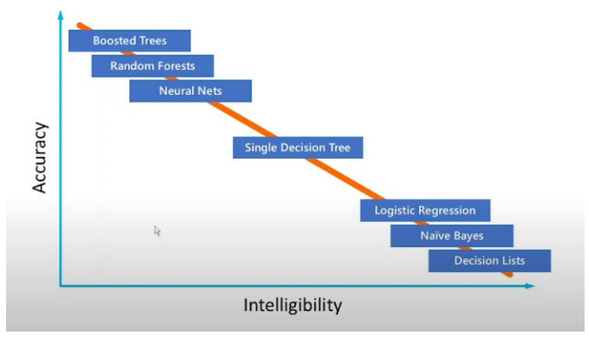
Explainability

Da ein Naive Bayes Classifier auf dem mathematischen Theorem basiert, ist bereits ein hoher Grad an Explainability gegeben. Durch eine genaue Analyse können Wahrscheinlichkeiten errechnet werden, womit Vorhersagen getroffen werden. Darüber hinaus bietet ein DAG erklärende Erkenntnisse, inwiefern unterschiedliche Merkmale in ein Klassifizierungsproblem hineinspielen und wie diese sich auf eine Zielvariable auswirken.

Da EBM speziell auf die Interpretierbarkeit ausgelegt ist, liefert es viele Funktionen, mit welchen man sowohl das Dataset als auch die Ergebnisse des trainierten Modells betrachten kann. Auch beim Gaussian Naive Bayes kann man viele Funktionen anwenden, die die Explainability unterstützen, jedoch werden diese Funktionen von anderen Bibliotheken bereitgestellt und man muss sie oft für jede Variable einzeln ausführen. Folglich ist es aufwendig, dennoch möglich einige Explainability-Aspekte gemäß des EBM im Bayes nachzustellen, was den Grad an Explainability weiter erhöht.

Fazit

Der Naive Bayes Classifier ist für seine Einfachheit und leichte Implementierung bekannt, bietet jedoch nicht dasselbe Maß an Interpretierbarkeit wie einige andere Modelle, insbesondere solche, die ausdrücklich auf Interpretierbarkeit ausgelegt sind.



Die Nachstellbarkeit gewisser Funktionen führt zu einem hohen Grad an Explainability. Dennoch lassen sich ausgewählte Aspekte der Vorhersagen nicht vollständig erklären.