Naive Bayes Classifier

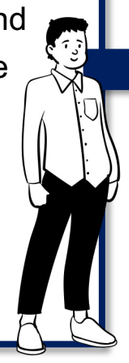
Explainable Machine Learning – Fabian Keck, Eva Fišerová , Natalie Martin

Inwiefern sind die Vorhersagen eines Gaussian Naive Bayes Classifiers erklärbar? Wie transparent ist das Modell?

Story & Ziel

Wir betrachten:

- John (28), Amerikaner, Einzelkind, ledig, High School Abschluss und arbeitet in der Farming-Industrie verdient mit einer Wahrscheinlichkeit von 58,2% mindestens 50.000€ im Jahr
- Warum ist das so? Warum sagt der Classifier das voraus?



Adult Income Dataset

- 14 demografische und berufliche Merkmale, u.a. Alter, Bildung, Beschäftigungsstatus
- Ziel: Bayes-Klassifikator für Einkommensklassifizierung
- Entscheidung: über oder unter 50.000 € Jahresgehalt
- Methode: Verwendung bedingter Wahrscheinlichkeiten

Bayes Theorem

Der Bayes-Klassifikator nutzt für die Klassifizierung bedingte Wahrscheinlichkeiten auf der Grundlage des Bayes-Theorems.

Das Bayes-Theorem: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Für das bessere Verständnis der Klassifizierung des Einkommens wird das Merkmal Alter isoliert betrachtet – angelehnt an John (28). Gesucht ist also die Wahrscheinlichkeit, dass John mit 28 Jahren bereits über 50.000 € im Jahr verdient, somit:

$$P(A|B) = P(\text{Einkommen} > 50.000\ € \mid \text{Alter} = 28\ \text{J.})$$?

Zudem sind folgende Wahrscheinlichkeiten bereits berechnet:

- P(B) = P(Alter = 28 J.) = 30 %
- P(A) = P(Einkommen > 50.000 €) = 40 %
- P(B|A) = P(Alter = 28 J. | Einkommen > 50.000 €) = 25 %

Damit kann nun die gesuchte P(A|B) berechnet werden:

$$P(\text{Einkommen} > 50.000\ € \mid \text{Alter} = 28\ \text{J.}) = \frac{P(\text{Alter} = 28\ \text{J.} \mid \text{Einkommen} > 50.000\ €) \cdot P(\text{Einkommen} > 50.000\ €)}{P(\text{Alter} = 28\ \text{J.})}$$

Setzt man die gegebenen Wahrscheinlichkeiten ein, erhält man:

$P(\text{Einkommen} > 50.000\ € \mid \text{Alter} = 28\ \text{J.}) = \frac{0{,}25 \cdot 0{,}4}{0{,}3} = 33{,}33\ \%$ ✓

Die Wahrscheinlichkeit, dass John mit 28 Jahren bereits jährlich über 50.000 € verdient, liegt bei 33,33 % - in diesem vereinfachten Beispiel. Der Bayes-Klassifikator bezieht in diese Entscheidung alle Merkmale ein und kommt deshalb zu abweichenden Ergebnissen. Die Berechnung erfolgt jedoch grundlegend nach dem dargestellten Prinzip der bedingten Wahrscheinlichkeiten auf der Grundlage des Bayes-Theorems.

Bayes vs. EBM

Ziel: Da EBM speziell auf die Interpretierbarkeit ausgelegt ist, wird es zum Vergleich herangezogen
Explainable Boosting Machine (EBM) ist ein baumbasiertes, zyklisches Gradient Boosting Generalized Additive Model.

Bayes	EBM
Accuracy bei Dataset: 80%	Accuracy bei Dataset: 87%
Marginal Plot: mit Funktion jointplot() kann die Beziehung einer Variablen aus den Trainingsdaten in Beziehung auf das Income gezeigt werden	Marginal Plot: mit Funktion Marginal() kann die Beziehung einer Variablen aus den Trainingsdaten in Beziehung auf das Income gezeigt werden
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 die Aufschlüsselung, wie viel jeder Begriff zur Vorhersage für eine einzelne Stichprobe beiträgt

Fazit Vergleich:

- EBM stellt viele Funktionen zur Unterstützung der Explainability (global und local explanation) bereit
- Bayes: durch Bibliotheken kann man einige Funktionen nachbauen → aufwendig und für jedes einzelne Merkmal, da kein Drop-Down Menü

Explainability

Es gibt drei Kriterien der Transparenz:

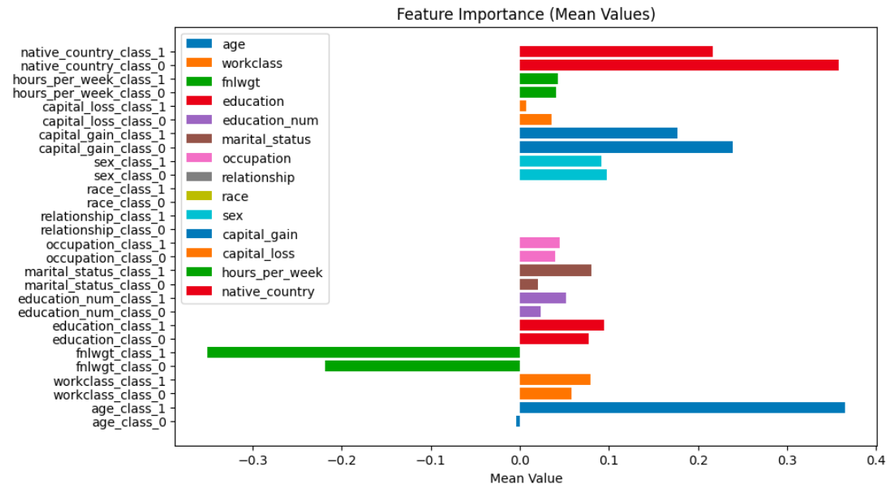
- Simulierbarkeit durch Menschen ☑
- Zerlegbarkeit des Modells in verständliche Attribute ☑
- Algorithmische Transparenz ☑

} Bayes ist ein transparentes Modell

Verwendeter Ansatz:
→ Post-hoc Explainability bei einem transparenten Modell

Explainability gegeben durch:

- Basiert auf mathematisches Theorem → hoher Grad an Explainability
- Mithilfe von Feature Importance erklärende Erkenntnisse: stellt dar, welchen Beitrag unterschiedliche Merkmale zu den Vorhersagen leisten



- Man kann mit Funktionen aus unterschiedlichen Bibliotheken die Explainability unterstützen

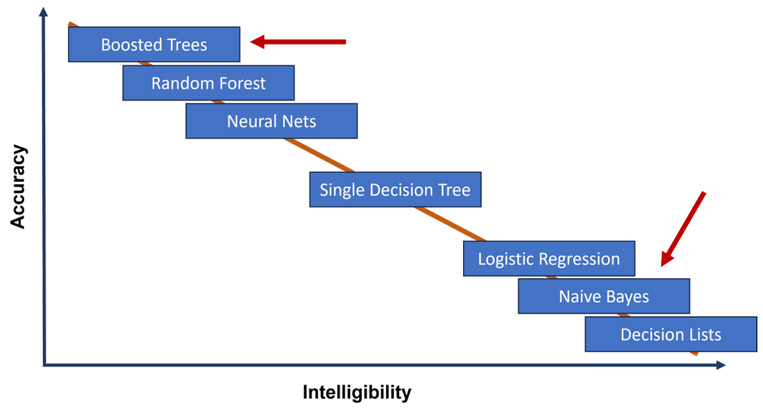
Defizite Explainability Bayes:

- Aufwendig, mit Funktionen die Explainability-Aspekte gemäß des EBM nachzustellen
- Wahrscheinlichkeiten können nicht einzeln errechnet werden → mangelnde Transparenz
- Bayes funktioniert für einige Anwendungsfälle gut, wird jedoch in der Praxis nicht häufig angewendet

Fazit

Naive Bayes Classifier:

- für seine Einfachheit und leichte Implementierung bekannt
- bietet jedoch nicht dasselbe Maß an Interpretierbarkeit wie einige andere Modelle, die auf Interpretierbarkeit ausgelegt sind



- Nachstellbarkeit gewisser Funktionen vom EBM führt zu einem gewissen Grad an Explainability
- dennoch lassen sich ausgewählte Aspekte der Vorhersagen nicht vollständig erklären, da…
  - der Trainingsprozess nicht vollständig einsehbar ist
  - das Theorem in einem solchem Rahmen angewendet wird, dass es für Menschen nur schwer nachvollziehbar erscheint

Inwiefern ist Johns Vorhersage nun erklärbar?
→ In vereinfachter Form (nur zwei Merkmale) ist das Theorem selbst anwendbar
→ Einflussnahmen gewisser Merkmale lassen sich herausfinden
→ Es bleibt Rest-Intransparenz übrig