Blatt 1: Entscheidungsbäume

Machine Learning – 2024

Amos Dinh

Erklärung

Auf diesem Arbeitsblatt wird eine Einfache Anwendung der Shannon-Entropie im diskreten Fall (Equation 1) vorgestellt.

$$H(X) = \mathbb{E}_{x\sim P_X}[I(x)] = \mathbb{E}_{x\sim P_X}[-\log(x)] = -\sum_{x\in X} \log p(x) \cdot p(x) \tag{1}$$

Problem 1 999 points

Gegeben ist folgendener Datensatz bestehend aus Beobachtungen/ Realisierungen der kategorischen Variablen *Outlook*, *Temperature*, *Humidity*, *Wind* und *Go outside(No studying)*. Wir wollen ??? vorhersagen:

Outlook	**Temperature**	**Humidity**	**Wind**	**Go outside (No studying)**
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no
rainy	mild	normal	strong	???

Strategie: Finden eines Modells für unsere Daten, das die Shannon-Entropie bezüglich der Variable *Go outside* verringert. Das bedeutet, wir wollen die Daten basierend auf den anderen Attributen so aufteilen, dass wir eine gute Vorhersage treffen können ob *yes* oder *no* zutrifft. Hierfür verwenden wir den Decision Tree Algorithmus.

Info: Dieser Algorithmus basiert auf der Shannon-Entropie, ist aber vorallem empirisch begründet. D.h. Abwandlungen und Erweiterungen dieser Idee sind äußerst effektiv auf echten (kleinen) Datensätzen ("SOTA/State of the Art").

Decision Tree Algorithmus

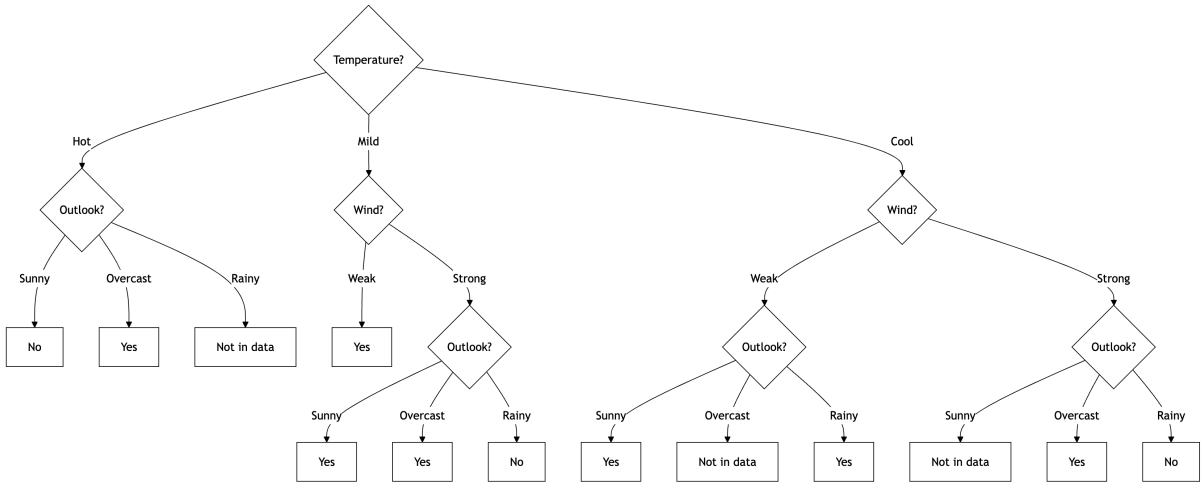

Figure 1: Schematisches Beispiel

Ziel des Algorithmus ist es, solch einen Entscheidungsbaum zu erstellen. Im Fall der Observablen *(Outlook:rainy, Temperature:mild, Humidity:normal, Wind:strong)* würden wir uns in Figure 1 für *no* entscheiden.

Wir suchen nach einer Aufteilungsstrategie, welche die Gesamtentropy "der Daten" bezüglich *Go outside* verringert.
Beispiel:
Wir teilen im ersten Schritt nach *Temperature* auf und vergleichen die Entropie nach dem Split mit der Ausgangsentropie. Hier ist X die Variable *Go outside*.

$$
\begin{aligned}
H(X) &= -\log\left(\tfrac{5}{14}\right) \cdot \tfrac{5}{14} - \log\left(\tfrac{9}{14}\right) \cdot \tfrac{9}{14} && = 0.65 \\
\hat{H}_{\text{hot}} &= -\log\left(\tfrac{2}{4}\right) \cdot \tfrac{2}{4} - \log\left(\tfrac{2}{4}\right) \cdot \tfrac{2}{4} && \approx 0.69 \\
\hat{H}_{\text{mild}} &= -\log\left(\tfrac{2}{6}\right) \cdot \tfrac{2}{6} - \log\left(\tfrac{4}{6}\right) \cdot \tfrac{4}{6} && \approx 0.64 \\
\hat{H}_{\text{cool}} &= -\log\left(\tfrac{1}{4}\right) \cdot \tfrac{1}{4} - \log\left(\tfrac{3}{4}\right) \cdot \tfrac{3}{4} && \approx 0.56 \\
\hat{H} &= \tfrac{4}{14} \cdot \hat{H}_{\text{hot}} + \tfrac{6}{14} \cdot \hat{H}_{\text{mild}} + \tfrac{4}{14} \cdot \hat{H}_{\text{cool}} && \approx 0.63 \\
\text{Information Gain} &= 0.65 - 0.63 = 0.02
\end{aligned}
$$

Als rudimentäre Strategie wählen wir in jedem Schritt das Attribut aus, welches den höchsten Information Gain besitzt.

Aufgabe: Erstelle den rudimentären Entscheidungsbaum für die gegebenen Daten und treffe eine Entscheidung für ???.

Bemerkung: Per Ast (Branch) können potentiell verschiedene Splits angewand werden. Die Lösung ist auf der nächsten Seite.

Bemerkung: Der Algorithmus kann auch auf numerische Feature Variablen erweitert werden (Feature Variable: jene die wir zur Vorhersage benutzen, Target Variable: Jene die vorhergesagt wird). Bspw. für Temperatur benutzen wir die gemessenen numerischen Werte und finden beliebig viele numerische Splits nach dem Schema "Temperatur $\geq x°$".

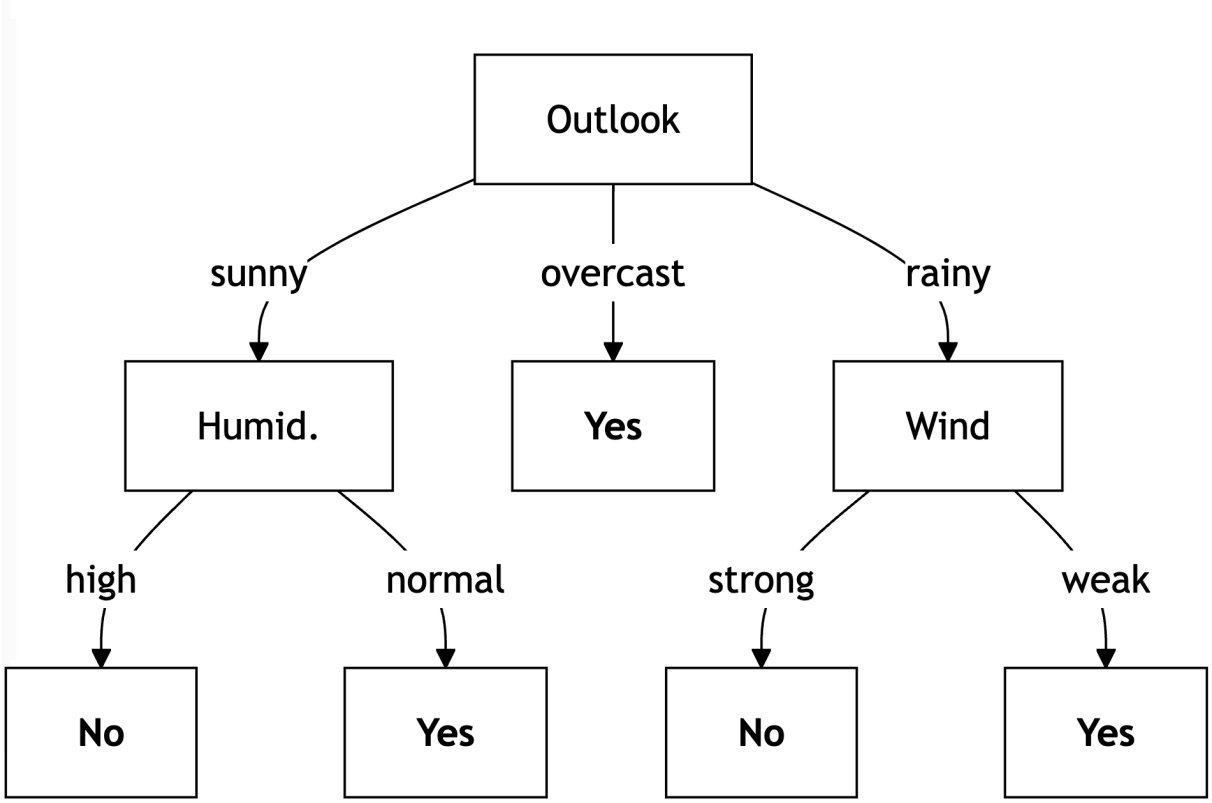

Figure 2: Lösung