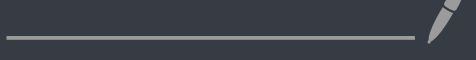


Teil 1: Rechenaufgabe

In der folgenden Tabelle finden Sie einen Datensatz mit zwei metrisch skalierten Merkmalen *feature_one*, *feature_two* und einer metrisch skalierten Zielgröße *target*.

Tabelle 1 - Beispieldatensatz

ID	feature_one	feature_two	target
0	1	16	14
1	2	10	12
2	5	5	3
3	6	7	6
4	9	14	15
5	8	17	19
6	4	16	13
7	5	13	12
8	6	6	5
9	2	11	9
10	5	4	4

a) Gehen Sie davon aus, dass *feature_one* als erstes Split-Merkmal gewählt wurde, um einen Entscheidungsbaum zur Schätzung von *target* zu berechnen. Bestimmen Sie die optimale Split Location anhand der (gewichteten) empirischen Varianz der Zielgröße.

Hinweis: Verwenden Sie als Kandidaten für Split Locations die Mittelwerte zwischen den sortierten, eindeutigen Werten des Merkmals.

a) feature_one_sortiert_eindeutig = {1, 2, 4, 5, 6, 8, 9}

```
feat_one_unique = df['feature_one'].unique()
feat_one_unique.sort()
feat_one_unique = feat_one_unique.tolist()

split_values = []
for i in range(len(feat_one_unique) - 1):
    split_values.append((feat_one_unique[i] + feat_one_unique[i+1]) /2)
print(split_values)

for i in split_values:
    items_below = df[df['feature_one']<= i].shape[0]
    items_above = df[df['feature_one']> i].shape[0]
    var_below = df[df['feature_one']<= i]['target'].var(ddof=0)
    var_above = df[df['feature_one']> i]['target'].var(ddof=0)

    weighted_var = (items_below * var_below + items_above * var_above) / (items_below + items_above)
    print(f'Split at {i}: Weighted Variance = {weighted_var}')
```

0.0s

```
[1.5, 3.0, 4.5, 5.5, 7.0, 8.5]
Split at 1.5: Weighted Variance = 22.69090909090909
Split at 3.0: Weighted Variance = 23.3219696969697
Split at 4.5: Weighted Variance = 22.25974025974026
Split at 5.5: Weighted Variance = 23.496753246753244
Split at 7.0: Weighted Variance = 13.818181818181818
Split at 8.5: Weighted Variance = 21.827272727272728
```

Die geringste Varianz liegt bei einem Split von 7 vor (13,818)

b) Bestimmen Sie analog die optimale Split Location für den Fall, dass feature_two als erstes Split-Merkmal ausgewählt wurde. Welches der beiden Merkmale ist anhand Ihrer Berechnungen nach der Logik des CART Algorithmus als erstes Split-Merkmal zu bevorzugen? Begründen Sie kurz.

c) Gehen Sie im Folgenden davon aus, dass ein Entscheidungsbaum verwendet wird, der in der ersten Stufe nach CART-Logik bei feature_one an der Stelle 4 verzweigt. In der zweiten Stufe wird in beiden Teilmengen nach feature_two beim Wert 12 verzweigt. Geben Sie die Prognosewerte der Endknoten an.

```
feat_two_unique = df['feature_two'].unique()
feat_two_unique.sort()
feat_two_unique = feat_two_unique.tolist()

split_values = []
for i in range(len(feat_two_unique) - 1):
    split_values.append((feat_two_unique[i] + feat_two_unique[i+1]) /2)
print(split_values)

for i in split_values:
    items_below = df[df['feature_two']<= i].shape[0]
    items_above = df[df['feature_two']> i].shape[0]
    var_below = df[df['feature_two']<= i]['target'].var(ddof=0)
    var_above = df[df['feature_two']> i]['target'].var(ddof=0)

    weighted_var = (items_below * var_below + items_above * var_above) / (items_below + items_above)
    print(f'Split at {i}: Weighted Variance = {weighted_var}')
```

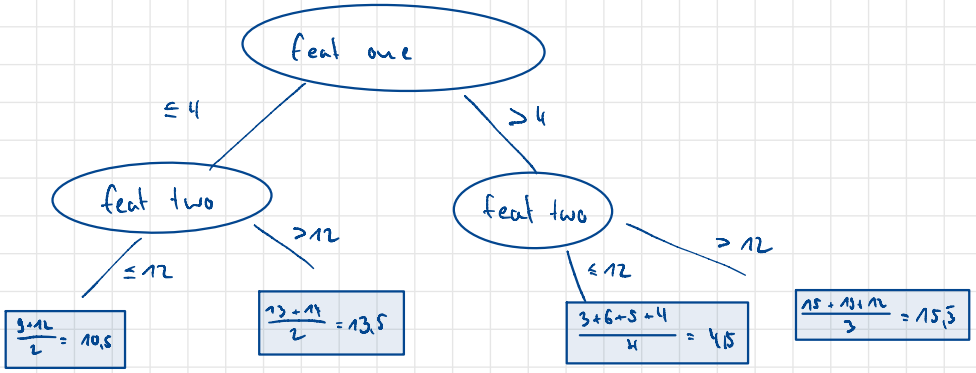
0.0s

```
[4.5, 5.5, 6.5, 8.5, 10.5, 12.0, 13.5, 15.0, 16.5]
Split at 4.5: Weighted Variance = 20.327272727272735
Split at 5.5: Weighted Variance = 14.227272727272727
Split at 6.5: Weighted Variance = 9.818181818181818
Split at 8.5: Weighted Variance = 5.7012987012987
Split at 10.5: Weighted Variance = 9.575757575757576
Split at 12.0: Weighted Variance = 7.881818181818182
Split at 13.5: Weighted Variance = 9.470779220779223
Split at 15.0: Weighted Variance = 14.196969696969695
Split at 16.5: Weighted Variance = 16.372727272727275
```

Split bei 8,5, da geringste Varianz (5,70)

→ Wahl von feature_two als erstes Split Merkmal (Varianz hier kleiner als bei feature_one → geringere Verunreinigung)

c)


feat one
≤ 4
> 4
feat two
feat two
≤ 12
> 12
≤ 12
> 12
$\frac{9+12}{2} = 10,5$
$\frac{13+14}{2} = 13,5$
$\frac{3+6+3+4}{4} = 4,5$
$\frac{15+13+12}{3} = 15,\overline{3}$