# Online-Test 3: Entscheidungsbäume und Evaluation

## Test und Assessment – Druckansicht

### Online-Test 3: Entscheidungsbäume und Evaluation

Datum: Tue Nov 2 07:55:40 2021  Maximale Punktezahl: 10

Frage 1 - Entscheidungsbäume - Trainingskomplexität (1 Punkt) [ID: 931232]

Gegeben sei ein Datensatz mit $m$ Datenobjekten und $n$ Attributen. Welche Zeitkomplexität hat das Trainieren des Baumes, angenommen, er ist balanciert?

- $O(m * n)$
  *(0 Punkte)*
- $O(m * n^2)$
  *(0 Punkte)*
- $O(m * \log m * n)$
  *(1 Punkt)*
- $O(m * \log m * n * \log n)$
  *(0 Punkte)*
- $O(m^2 * n^2)$
  *(0 Punkte)*

Ein balancierter Baum für $m$ Datenobjekte hat die Tiefe $\log m$. Für jeden Knoten muss zum Bestimmen des Splits über die $m$ Datenobjekte (bzw. weiter unten im Baum werden es dann weniger) und die $n$ Attribute iteriert werden. Die Attributwerte sollten jeweils sortiert sein, aber das lässt sich als einmaliger Voraufwand in $O(m * \log m * n)$ bewerkstelligen.

Frage 2 - Entscheidungsbäume - Vorhersagekomplexität (1 Punkt) [ID: 931234]

Gegeben sei ein Entscheidungsbaum, der auf einem Datensatz mit $m$ Datenobjekten und $n$ Attributen trainiert wurde und balanciert ist. Welche Zeitkomplexität hat die Vorhersage für ein neues Datenobjekt?

- $O(\log m)$
  *(1 Punkt)*
- $O(m)$
  *(0 Punkte)*
- $O(m * n)$
  *(0 Punkte)*
- $O(1)$
  *(0 Punkte)*
- $O(m * \log m * n)$
  *(0 Punkte)*

Bei einem balancierten Baum beträgt die Tiefe $\log m$. Die Anzahl der Attribute ist beim Vorhersagen nicht mehr relevant, da in jedem Split nur ein Attribut abgefragt wird.

Frage 3 - Entscheidungsbaum - Vorhersagequalität (1 Punkt) [ID: 946066]

Sie trainieren einen Entscheidungsbaum ohne jegliches Pre- oder Post-Pruning (d.h., der Baum darf beliebig oft verzweigen) auf einem beliebigen Datensatz mit zwei Klassen und einem beliebigen Split in Trainings- und Testdaten.

Im schlechtesten Fall kann die Gesamterfolgsquote 50% *(0.5 Punkte)* auf den Trainingsdaten bzw. 0% *(0.5 Punkte)*

auf den Testdaten betragen.

Lücke 1: Die Erfolgsquote auf den Trainingsdaten kann unter 100% fallen, selbst wenn man ohne Beschränkung trainiert. Dies ist der Fall, wenn sich Datenobjekte, die zu verschiedenen Klassen gehören, nicht über ihre gegebenen Attributwerte unterscheiden lassen. Man redet hier vom nicht reduzierbaren Fehler bzw. Bayes Error. Bei einer binären Klassifikation und unter Annahme balancierter Klassen (50:50-Verteilung) heißt das, dass 50% der Datenobjekte falsch klassifiziert werden und somit nur 50% richtig. Wenn die Klassen dagegen zu unterschiedlichen Proportionen auftreten, kann die Erfolgsquote auch bei nichts diskriminierenden Attributen über 50% gesteigert werden, indem die häufigere Klasse geraten wird.

Lücke 2: Auch auf den Testdaten kann der Entscheidungsbaum nur raten, wenn die Attribute auf den Trainingsdaten nicht diskriminierend waren. Davon abgesehen kann es sein, dass durch die Aufteilung zwischen Trainings- und Testdaten die Datenobjekte aus den Trainingsdaten anhand anderer Regeln klassifiziert werden können als die Objekte aus den Testdaten (d.h. die Verteilung der Attributwerte bzw. der Zusammenhang zwischen Attributen und Klassenlabels kann in Trainings-Split und Test-Split unterschiedlich sein). Dadurch ist die Erfolgsquote auf den Testdaten im schlimmsten Fall 0%, selbst wenn sie auf den Trainingsdaten sehr hoch war. Durch einen zufälligen Split, der idealerweise bzgl. der Klassenlabels stratifiziert ist, d.h., auch die Proportionen der Klassen aus den Trainingsdaten beibehält, soll dies vermieden werden.

## Frage 4 - Conditional Risk (1 Punkt) [ID: 931284]

Sie sind mit dem Lernen auf die AGD-Prüfung fast fertig. Es verbleiben lediglich 10 Stunden reine Lernzeit. Sollten Sie dennoch durchfallen, müssen Sie die Prüfung später noch einmal absolvieren, mit einem Vorbereitungsaufwand von zusätzlich 40 Stunden. Sie sind sich zu 80% sicher, dass sie die Prüfung bestehen. Wie hoch ist das Conditional Risk, ausgedrückt als Arbeitsaufwand in Stunden?

Der Wert muss zwischen 18 und 18 liegen

80 % * 10 h + (100 % - 80 %) * (10 h + 40 h) = 18 h bzw.

100 % * 10 h + (100 % - 80 %) * 40 h = 18 h

Der Aufwand für die Vorbereitung auf die erste Prüfung fällt auf jeden Fall an. Die zweite Prüfung muss nur vorbereitet werden, wenn die erste nicht bestanden wurde.

## Frage 5 - Training, Validierung und Test (1 Punkt) [ID: 931236]

Ordnen Sie die Datensplits ihrer Hauptfunktion zu!

| | | | |
|---|---|---|---|
| Trainingsdaten | passt zu | Erstellen des Modells | (0.34 Punkte) |
| Validierungsdaten | passt zu | Vergleich von Modellen und Parametereinstellungen | (0.33 Punkte) |
| Testdaten | passt zu | Bewertung des final ausgewählten Modells | (0.33 Punkte) |

Siehe Vorlesung. Nach der Modell- und Parameterauswahl können die Validierungsdaten in das Training einbezogen werden. Nach dem Bestimmen der Qualität des final ausgewählten Modells können die Testdaten in das Training einbezogen werden.

## Frage 6 - Bias und Varianz (1 Punkt) [ID: 931247]

Angenommen, ein Vorhersagemodell hat

a) eine Erfolgsquote auf den Trainingsdaten von deutlich unter 100%, z.B. nur 60%.

b) eine große Differenz zwischen Erfolgsquote auf Trainings- und Testdaten.

Was ist das jeweils zugrundeliegende Phänomen?

- ○ a) Bias, b) Bias
  *(0 Punkte)*

- [x] a) Bias, b) Varianz
  *(1 Punkt)*
- [ ] a) Varianz, b) Bias
  *(0 Punkte)*
- [ ] a) Varianz, b) Varianz
  *(0 Punkte)*

Wenn das Modell schon auf den Trainingsdaten eine niedrige Erfolgsquote hat, deutet dies an, dass das Modell womöglich nicht komplex genug ist, um die Daten zu lernen (hohes Bias). Natürlich kann auch beim Trainingsprozess selbst etwas schiefgelaufen sein oder der Datensatz enthält einfach keine Attribute, die Rückschlüsse auf die Zielvariable erlauben.

Besteht ein großer Unterschied bzgl. der Vorhersagequalität zwischen Training und Validierung/Test, ist das Modell für gewöhnlich zu stark an die Trainingsdaten angepasst (Overfitting), was zu nicht generalisierbaren Vorhersagen führt (hohe Varianz).

Frage 7 - Erfolg durch Raten (1 Punkt) [ID: 946083]

Bei welchem der folgenden Evaluationsmaße erreicht man einen Wert von 1, indem man einfach immer "positiv" vorhersagt (angenommen, beide Klassen treten in den Daten auf)?

- [ ] Gesamt-Erfolgsquote (Accuracy)
  *(0 Punkte)*
- [ ] Kappa-Koeffizient
  *(0 Punkte)*
- [ ] Precision
  *(0 Punkte)*
- [x] Recall
  *(1 Punkt)*

Recall ist definiert als $\frac{TP}{TP+FN}$. Wenn man immer "positiv" vorhersagt, gibt es natürlich keine false negatives. Man erhält zwar false positives, aber diese gehen beim Recall, im Gegensatz zu den anderen genannten Evaluationsmaßen, nicht in die Berechnung mit ein.

Frage 8 - Kappa-Koeffizient (1 Punkt) [ID: 931279]

Gegeben sei folgendes Ergebnis einer Klassifikation:

Anzahl TP = 60, FP = 10, FN = 10, TN = 20

Berechnen Sie den Kappa-Koeffizienten.

Runden Sie Ihre Lösung auf zwei Nachkommastellen. Verwenden Sie einen Punkt als Dezimaltrennzeichen.

Der Wert muss zwischen 0.52 und 0.52 liegen

Der Klassifikator sagt zu 70% die Klasse "True" voraus (TP + FP) und zu 30 % die Klasse "False" (TN+ FN). Real gehören 70 Objekte zur Klasse "True" (TP + FN) und 30 zur Klasse "False" (TN + FP). Wenn man gemäß obiger Wahrscheinlichkeiten die Klassenzugehörigkeit rät, klassifiziert man 70% * 70 + 30% * 30 = 58 Objekte richtig. Unser Klassifikator hat wiederum 80 Objekte richtig klassifiziert (TP + TN). Daraus ergibt sich ein Kappa-Koeffizient von $\frac{80-58}{100-58} = 0.52$. (Man kann auch direkt in Anteilen rechnen, also $\frac{0.8-0.58}{1-0.58} = 0.52$)

Frage 9 - ROC und Lift (1 Punkt) [ID: 931245]

Ordnen Sie die Metriken auf den Achsen der jeweiligen Grafik zu! (Hinweis: jede Grafik hat zwei Achsen)

ROC-Grafikpasst zuTrue-Positive-Rate (Recall) (0.25 Punkte)
ROC-Grafikpasst zuFalse-Positive Rate (0.25 Punkte)
Lift-Grafik passt zuAnzahl von True Positives (0.25 Punkte)
Lift-Grafik passt zuAnzahl von Datenobjekten(0.25 Punkte)

Siehe Vorlesung. Eine weitere typische Auswertungsgrafik, die nicht Bestandteil der Vorlesung war, ist die Precision-Recall-Kurve.

Frage 10 - MDL (1 Punkt) [ID: 931238]

Sei D der Datenbestand und M ein Vorhersagemodell. Welche Wahrscheinlichkeiten spielen nach dem MDL-Prinzip eine Rolle, um das Vorhersagemodell auszuwählen?

- [x] Pr(M)
  *(Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)*
- [ ] Pr(D)
  *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)*
- [x] Pr(M|D)
  *(Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)*
- [x] Pr(D|M)
  *(Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)*

Wir wollen das wahrscheinlichste Vorhersagemodell auswählen, also P(M|D) maximieren. Entsprechend der Bayes'schen Regel kann man diese Wahrscheinlichkeit auch über P(D|M), P(D) und P(M) ausdrücken. P(D) können wir durch die Modellauswahl allerdings nicht beeinflussen, der Datensatz ist gegeben.