Übung 1: Anwendungen von Predictive Analytics

Aufgabe 1

Es sei V_D^K das Volumen einer D-dimensionalen Kugel mit Durchmesser d und V_D^W das Volumen eines D-dimensionalen Würfels mit Kantenlänge d. Weisen Sie nach, dass

$$\lim_{D\to\infty} \frac{V_D^K}{V_D^W} = 0$$

gilt und interpretieren Sie das Ergebnis.

Hinweis: Verwenden Sie

$$V_D^K(r) = \frac{\pi^{D/2}}{\Gamma(\frac{D}{2}+1)} r^D \qquad \text{mit} \quad \Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \quad \text{für } x > 0$$

sowie die Stirling-Formel

$$\Gamma(x+1) \approx \sqrt{2\pi} e^{-x} x^{x+1/2}$$

für $x \gg 1$.

Aufgabe 2

In einer Studie wurde für verschiedene europäische Regionen untersucht, wie viele Störche dort zu Hause sind und wie hoch die Geburtenrate ist. Es zeigte sich, dass eine statistisch signifikante positive Korrelation zwischen der Anzahl der Störche und der Anzahl der Babys in einer Region besteht. Das heißt, je mehr Störche in einer Region leben, umso mehr Babys gibt es dort. Erörtern Sie, ob aus dieser positiven Korrelation auf einen kausalen Zusammenhang zwischen Zahl der Geburten und Anzahl an Storchen in einer Region geschlos- sen werden kann (i.a.W. die Kinder werden vom Storch gebracht).

Aufgabe 3

Beurteilen Sie, ob Machine Learning zur Untersuchung folgender Problemstellungen geeignet ist:

a) Mittels Machine Learning soll aus den soziographischen Merkmalen eines potenziellen Kreditnehmers (z.B. Geschlecht, Alter, Einkommen, Familienstand, Haushaltsgröße, Schulbildung, Beruf) die Wahrscheinlichkeit geschätzt werden, dass dieser den Kredit vollständig zurückzahlt.

b) Mittels Machine Learning soll aus Gesundheitsdaten die Lebenserwar- tung geschätzt werden.

c) Mit Hilfe von Machine Learning soll die Wahrscheinlichkeit für einen noch nie dagewesenen Kernkraftwerksunfall geschätzt werden.

d) Mittels Machine Learning sollen die Lottozahlen vorausgesagt werden.

e) Mit Hilfe von Machine Learning sollen die guten und schlechten Teile einer Produktion vorhergesagt werden.

Aufgabe 4

Beurteilen Sie die Aussagefähigkeit der Ergebnisse der beiden folgenden Vorgehensweisen:

a) Es wird eine Bank betrachtet, die mit Hilfe von verschiedenen Risikomerkmalen beurteilt, ob ein Antragsteller kreditwürdig ist oder nicht. Hierzu hat die Bank in der Vergangenheit die Antragsteller, denen ein Kredit gewährt wurde, über die Zeit beobachtet und festgehalten, ob es nachfolgend zu einem Kreditausfall kam oder nicht. Mit Hilfe dieser Daten wird ein existierendes Modell zur Prognose der Kreditwürdigkeit neuer Antragsteller getestet. Die Ergebnisse dieses Modells für die bereits vorhandenen Kreditnehmer sind exzellent.

b) Um die Performance der „Buy & Hold-Strategie“ zu testen, wird für alle aktuell an der Börse gelisteten Unternehmen analysiert, welcher Profit erzielt worden wäre, wenn man die Aktien dieser Unternehmen vor 50 Jahren gekauft und bis heute gehalten hätte. Als Ergebnis erhält man eine exzellente durchschnittliche Rendite.

Aufgabe 5

Geben Sie für die folgenden Situationen jeweils an, ob man generell erwarten würde, dass ein komplexes/flexibles oder ein einfaches/unflexibles Modell besser abschneidet. Begründen Sie Ihre Antwort.

a) Der Umfang N des Trainingsdatensatzes ist extrem groß, und die Anzahl der Inputvariablen p ist gering.

b) Die Anzahl der Inputvariablen p ist sehr groß, und der Umfang N des Trainingsdatensatzes ist klein.

c) Der funktionale Zusammenhang zwischen der Outputvariablen und den Inputvariablen ist hochgradig nichtlinear.

d) Die Varianz des zufälligen Fehlers ε, d.h. der irreduzible Fehler, ist extrem hoch.

Aufgabe 6

Was sind die Vor- und Nachteile eines komplexen/flexiblen Regressions- oder Klassifikationsmodells gegenüber eines weniger flexiblen Ansatzes? Unter welchen Umständen könnte ein flexibleres Modell einem weniger flexiblen Modell vorgezogen werden und unter welchen Umständen könnte ein weniger flexibleres Modell vorteilhaft sein?