DHBW Mannheim
Studiengang Wirtschaftsinformatik/Data Science

27. November 2022

Introduction to Data Science
Übungsblatt 2
Lösungsskizze

1. Kreuzvalidierung [5 Punkte]

Was versteht man unter Kreuzvalidierung?

2. Modellgüte [5 Punkte]

Erklären Sie qualitativ, wie allgemein eine Güte von Modellen des Supervised Learning und des Unsupervised Learning jeweils definiert werden kann.

3. Gütemaß numerischer Modelle [8 Punkte]

Geben Sie ein Gütemaß für ein numerisches Modell M an und beschreiben Sie, wie damit auf einem Validierungs- oder Testdatensatz die Güte ermittelt wird.

4. Wahrheitsmatrix binärer Klassifikatoren [18 Punkte]

Sei **K** ein trainierter binärer Klassifikator, und sei **TE** die Menge der annotierten Testdatensätze für **K**, die aus Tupeln (O, w_O) von Datenobjekten O mit tatsächlichen annotierten Klassifikationen w_O besteht. Dabei kann w_O die Werte `positive` oder `negative` annehmen.

Beweisen Sie durch stichhaltige logische oder mathematische Argumentation, dass für die Mengen TP (true positive), TN (true negative), FP (false positive), FN (false negative) in TE folgende Sachverhalte gelten:

1. TP, TN, FP, FN sind zueinander disjunkte Untermengen von TE
2. $TP \cup TN \cup FP \cup FN = TE$
3. $TP \cup FP = \{O \in TE \mid K(O) = positive\}$
4. $TN \cup FN = \{O \in TE \mid K(O) = negative\}$
5. $TP \cup FN = \{O \in TE \mid w_O = positive\}$
6. $TN \cup FP = \{O \in TE \mid w_O = negative\}$
7. $TP \cup TN = \{O \in TE \mid K(O) = w_O\}$
8. $FP \cup FN = \{O \in TE \mid K(O) \neq w_O\}$
9. $tp + tn + fp + fn = |TE|$

Dabei ist $tp = |TP|$, $tn = |TN|$, $fp = |FP|$, $fn = |FN|$.