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?

Lösung:

Kreuzvalidierung ist eine Technik, mit der n verschiedene Modelle erlernt und verglichen werden können. Dabei werden die Gesamttrainingsdaten in n etwa gleichgroße Blöcke unterteilt. Jeder Block dient als Validierungsdatensatz zur Zwischen-Gütebestimmung für jeweils eines der Modelle, die restlichen Blöcke dienen vorab zum Erlernen des jeweiligen Modells. Das Modell mit der besten Validierung/Zwischen-Güte wird dann mit dem Testdatensatz endgültig bewertet.

2. Modellgüte [5 Punkte]

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

Lösung:

1. Güte wird stets mit Testdatensätzen (oder Validierungsdatensätzen) ermittelt, die disjunkt zu den Trainingsdaten sind, mit denen das Modell mit Hilfe von Lernalgorithmen erlernt wurde.
2. Im Supervised Learning wird die Güte durch geeigneten aggregierten Vergleich bzw. Diskrepanz der Sollwerte der annotierten Datenobjekte und deren Istwerten ermittelt.
3. Im Unsupervised Learning wird die Güte dadurch ermittelt, dass die Struktur des Modells der Lernphase mit der Struktur des Modells der Testphase in geeigneter Weise verglichen wird.

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.

Lösung:

Für den annotierten Test- oder Validierungsdatensatz $(X_k, Y_k)_{k=1,\dots,n}$ wird die Güte des Modells M beispielsweise durch die mittlere quadratische Abweichung ermittelt:

$$\mathbf{MQA}_M((X_k, Y_k)_{k=1,\dots,n}) = \frac{1}{n}\sum_{k=1}^{n}(M(\mathbf{X}_k) - \mathbf{Y}_k)^2$$

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

Lösung:

Siehe Vorlesung.