0.1 Pros

0.2 Cons

0.3 Hypothesen

Die Performance von Support Vector Machines wird anhand verschiedener Datenszenarien untersucht. Dabei werden Logistic Regression und k-nearest Neighbour als Vergleichsalgorithmen hinzugezogen. Genauer gesagt, wird in neun verschiedene Szenarien unterschieden, welche sich durch zwei Unterteilungen ergeben: die Form der Entscheidungsgrenze sowie das Verhältnis zwischen der Anzahl an Dimensionen (p) und der Anzahl an Beobachtungen (n).Dabei werden ausschließlich binäre Klassifikationen untersucht. Es ergibt sich folgende Aufteilung:

	linear	polynomial	radial
p « n	S1	S2	S3
p = n	S4	S5	S6
p » n	S7	S8	S9

Im Folgenden werden Studien hinzugezogen, um eine Einschätzung der Performance in den verschiedenen Szenarien vorzunehmen und Hypothesen abzuleiten.
Es konnte gezeigt werden, dass in einem Szenario, indem erheblich mehr Beobachtungen als Dimensionen und eine lineare Entscheidungsgrenze vorliegen (S1), deutliche Unterschiede zwischen SVM, k-NN und LogR bei der Diskriminationsfähigkeit auftreten (Entezari-Maleki et al., 2009). k-NN und lineare SVM zeigen AUC-Werte nahe 1 auf, was für eine nahezu perfekte Differenzierung der Klassen spricht. LogR hingegen hat einen Wert knapp über 0.5, was nur etwas besser als eine Zufallsauswahl ist. Darüber hinaus ist festzustellen, dass die Unterschiede deutlicher werden, je höher die Anzahl an Beobachtungen ist.
Für den Fall einer radialen Entscheidungsgrenze (S3) sind die Ergebnisse ähnlich. So erreicht in diesem Beispiel eine SVM mit radialem Kernel im Vergleich zu einer LogR eine um 34% höhere Genauigkeit (Fávero et al., 2022).

Liegt ein Szenario vor, indem die Anzahl der Dimensionen erheblich größer ist, als die Anzahl der Beobachtungen, mit einer linearen Entscheidungsgrenze (S7), sind die Ergebnisse differenzierter zu betrachten. So schneidet die SVM mit polynomialem Kern am besten unter den genannten Algorithmen ab, jedoch die lineare SVM am schlechtesten (als Kriterium wurde die mittlere Performance über 100 Datensätze evaluiert) (Scholz & Wimmer, 2021). Während k-NN auch in diesem Szenario eine gute Performance hat, schneiden LogR und SVM mit radialem Kern mittelmäßig ab.

Basierend auf den Ergebnissen der genannten Studien können Schlussfolgerungen gezogen werden. Es ist anzunehmen, dass in niedrigdimensionalen Szenarien k-NN und SVM´s besser performen als LogR. Jedoch ist zu vermuten, dass die Wahl des Kerns bei SVM´s einen großen Einfluss auf die Performance hat.
In hochdimensionalen Szenarien zeigt vermutlich die SVM mit polynomialem oder radialem Kern eine gute Performance, unabhängig von der Form der Entscheidungsgrenze, während die lineare SVM voraussichtlich weniger gut abschneiden wird. Es scheint so, dass auch k-NN und LogR in hochdimensionalen Szenarien zumindest mittelmäßig abschneiden. Hier ist jedoch zu beachten, dass nur eine lineare Entscheidungsgrenze betrachtet wurde und in den Szenarien S8 und S9 andere Ergebnisse möglich sind.

$$H1 : Hypothese1 \tag{1}$$

Literatur

Entezari-Maleki, R., Rezaei, A., & Minaei-Bidgoli, B. (2009). Comparison of Classification Methods Based on the Type of Attributes and Sample Size. *Journal of Convergence Information Technology*, *4*(3), 94–102. https://doi.org/10.4156/jcit.vol4.issue3.14

Fávero, L. P., Belfiore, P., Santos, H. P., dos Santos, M., de Araújo Costa, I. P., & Junior, W. T. (2022). Classification Performance Evaluation from Multilevel Logistic and Support Vector Machine Algorithms through Simulated Data in Python. *Procedia Computer Science*, *214*, 511–519. https://doi.org/10.1016/j.procs.2022.11.206

Scholz, M., & Wimmer, T. (2021). A Comparison of Classification Methods across Different Data Complexity Scenarios and Datasets. *Expert Systems with Applications*, *168*, 114217. https://doi.org/10.1016/j.eswa.2020.114217