Im Folgenden werden Studien hinzugezogen, um eine Einschätzung der Performance in den verschiedenen Szenarien vorzunehmen und Hypothesen abzuleiten. Vorab ist zu erwähnen, dass die Evaluation von Klassifikationsmethoden anhand synthetischer Datensätze in der Literatur begrenzt ist. Da für diese Arbeit die Form der Entscheidungsgrenze entscheidend ist, werden dennoch auschließlich Arbeiten mit synthetischen Datensätzen zu Rate gezogen.
Aufgrund dessen, dass der Datengenerierende Prozess hier so ausgearbeitet wurde, dass er mit den Annahmen der SVMs arbeitet, erwarten wir zuerst einmal eine bessere Performance der SVM Classifier im Vergleich zu den anderen Methoden.

H1: Die SVM Classifier Performen über alle Datensituationen im Durchschnitt besser als die anderen Classifier

Des Weiteren wurden in den einzelnen Kategorien des daten generierenden Prozesses die Entscheidungsgrenzen speziell auf verschiedene Kernels der SVMs zugeschnitten. Daher sollten SVMs mit linearem Kernel im Setting mit linearer Entscheidungsgrenze mindestens so gut oder besser als die restlichen Classifier performen. Gleiches gilt für SVMs mit polynomialen Kernel im Setting mit einer quadratischen Entscheidungsgrenze und radiale Kernel bei einer Hypershäre als Entscheidungsgrenze.

H2: Die SVM Classifier mit dem Kernel, der für die jeweiligen DGP zugeschnitten ist sollten mindestens genauso gut oder besser Performen als die restlichen Classifier

Es konnte weiterhin 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 logistischer Regression 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. Die logistische Regression 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 logistischen Regression eine um 34% höhere Genauigkeit (Fávero et al., 2022).

Die Szenarien S4 bis S6 finden in der Literatur kaum Beachtung, weshalb hier keine Studien herangezogen werden können. Liegt jedoch 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 logistische Regression und SVM mit radialem Kern mittelmäßig ab. Hierbei ist wichtig zu erwähnen, dass in der Studie keine Ergebnisse über die genaue Performance präsentiert wurden, sondern lediglich die Ränge der 25 behandelten Klassifikationsmethoden. Somit können nur eingeschränkte Schlussfolgerungen gezogen werden.

Basierend auf den Ergebnissen der genannten Studien können folgende Schlussfolgerungen gezogen werden.

H3: In niedrigdimensionalen Szenarien performen k-NN und SVM´s besser als eine logistische Regression.

Jedoch ist zu vermuten, dass die Wahl des Kerns bei SVM´s einen großen Einfluss auf die Performance hat. So ist, basierend auf den mathematischen Grundlagen, anzunehmen, dass die SVM mit dem jeweils passenden Kern zu der vorliegenden Datensituation am besten performt. Da die SVM mit polynomialem und radialem Kern weitaus flexibler sind, werden diese voraussichtlich insgesamt betrachtet besser abschneiden als die SVM mit linearem Kern.

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 logistische Regression in hochdimensionalen Szenarien zumindest mittelmäßig abschneiden. Es ist aber auch bekannt, dass gerade die k-NN Methode in hochdimensionalen Settings schlechter performt (James et al., 2021). Hier ist jedoch zu beachten, dass

nur eine lineare Entscheidungsgrenze betrachtet wurde und in den Szenarien S8 und S9 andere Ergebnisse möglich sind. Wir schließen Final daraus:

H4: In hochdimensionalen Settings performen v.a. SVMs mit radialen und polynomialen Kernel besser als die anderen Klassifikationsmethoden

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

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