Das Ziel dieser Arbeit ist es, in verschiedenen Datensituationen die performance von SVM Algorithmen für die binäre Klassifikation zu evaluieren. Dafür wollen wir eine Reihe von Datensätzen mit verschiedenen Charakteristiken synthetisch erstellen und entsprechen als Trainigsdaten verwenden. Die Datensätze unterscheiden sich in zwei zentralen Eigenschaften. Das erste sind die Dimensionen. Es soll unterschieden werden in drei Kategorien. Die erste ist, dass es deutlich mehr Beobachtungen als Variablen gibt, also $n \gg p$. Ein solches Datenszenario kann z.B. im Kontext des Zensus auftreten, in dem eine große Anzahl an Personen befragt wird, aber die Vorgabe besteht, dass die Bürger nicht zu stark belastet werden sollen, weshalb nur einige wenige Kernfragen gestellt werden. Das zweite Szenario stellt Datensätze vor die etwa gleich viele Beobachtung wie Variablen haben $n \approx p$. Ein solches Szenario kann in vielen Kontexten auftreten. Das letzte Szenario behandelt dann entsprechend den Fall $n \ll p$. Dies tritt oft im Kontext von Datenerhebungen im medizinischen Bereich auf, da sehr viele Erhebungen zu kostspielig wären. Auch im Bereich des Natural Language Processing sind solche Datensätze häufiger anzutreffen (Scholz & Wimmer, 2021).

Die zweite Charakteristik ist der Datengenerierende Prozess. Da in dieser Arbeit SVMs im Vordergrund stehen und wir hier vor allem Zeigen wollen, wie SVMs funktionieren ist die Idee den DGP so aufzubauen, dass er der grundlegenden Idee der SVMs am ehesten Entspricht. Die Grundlegendeidee ist, im ersten Schritt eine Hyperplane, im p-Dimensionalen Raum, in einer bestimmte Form zu erstellen und anschließend auf jeweils einer Seite dieser Hyperplane $n/2$ zufällige Punkte zu samplen, die die jeweilige Ausprägung in der Zielvariable repräsentieren.

Insgesammt gibt es auch hier wieder 3 Kategorien. Die erste sind linear getrennte Daten. Dafür wird eine lineare Hyperplane der Form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0$$

mit zufälligen Koeffizienten erzeugt. Diese Daten sollen also den Annahmen entsprechen, die für SVMs mit linearem Kernel gelten.
In der zweiten Situation hat die Hyperplane eine quadratische Form:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + ... + \beta_{2p-1} X_p + \beta_{2p} X_p^2 = 0$$

diese Form der Trennung stellt also eine Merkmalserweiterung um quadratische Terme dar und funktioniert damit ähnlich wie eine SVM mit polynomialen Kernel mit $d = 2$. Der letzte DGP geht von einer noch Komplexeren Entscheidungsgrenzen aus. Es wird hier ein Hypershpäre im p dimensionalen Raum erstellt und einmal innerhalb und einmal außerhalb dieser gesampelt. Dafür wurde für eine Beobachtung j, $p-1$ Winkel θ zufällig erstellt, ein Radius r festgelegt und anschließend die einzelnen Werte $X_{1,j}, X_{2,j}, ..., X_{p,j}$ berechnet. Die berechnung erfolgt dabei über die Definition von spärischen Koordinaten:

$$\begin{aligned}
X_{i,1} &= r\cos(\theta_1) \\
X_{i,2} &= r\sin(\theta_1)\cos(\theta_2) \\
X_{i,3} &= r\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\
&\vdots \\
X_{i,p-1} &= r\sin(\theta_1)\ldots\sin(\theta_{p-2})\cos(\theta_{p-1}) \\
x_{i,p} &= r\sin(\theta_1)\ldots\sin(\theta_{p-2})\sin(\theta_{p-1})
\end{aligned}$$

Dieser Vorgang wird dann $n/2$ mal, mit einem zufälligen Radius mit Mittelwert μ_r und einer Varianz σ_r^2 für die eine Ausprägung der Zielvariable wiederholt. Für die andere Ausprägung wurde das gleiche dann durchgeführt mit einem neuen Mittelwert $\mu_r + k, k \in \mathbb{R}$. je nachdem wie k und σ_r^2 gewählt werden kann die Trennbarkeit der Daten angepasst werden.

Es ergeben sich daher 9 unterschiedliche Datensituationen, welche in ihren Dimensionen und Komplexität der Entscheidungsgrenze variieren. Die Kürzel für die Situationen sind in Tabelle 1 abgetragen

Zusätzlich soll nicht nur ein Vergleich zwischen der Performance der SVMS mit verschiedenen Kernel gemacht werden, sondern auch die Klassifikationsgüte der SVMs im Vergleich zu anderen gängigen Klassifikationsmethoden gezogen werden. Dafür werden Logistic Regression und k-nearest Neighbour als Vergleichsalgorithmen hinzugezogen.

	linear	polynomial	radial
$p \ll n$	S1	S2	S3
$p = n$	S4	S5	S6
$p \gg n$	S7	S8	S9

Tabelle 1: Datensituationen

Literatur

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