Homework 3

Jizheng Chen 520021911182

May 18, 2023

[1]

1 SVM vs. Neural Networks

1.1 Datasets

In this experiment three datasets with different sample scales, feature dimensions and number of classes are used:

- **Iris** The Iris dataset is a popular and widely used dataset in machine learning and statistics. It consists of measurements of various attributes of three different species of Iris flowers: setosa, versicolor, and virginica. The dataset was first introduced by the British statistician and biologist Ronald Fisher in 1936 and has since become a standard benchmark for classification algorithms.

  The dataset contains 150 samples, with 50 samples for each Iris species. For each sample, four features are measured: sepal length, sepal width, petal length, and petal width. These measurements are recorded in centimeters.

- **Glass** The Glass dataset is another well-known dataset often used in machine learning and pattern recognition tasks. It contains information about the chemical composition of different types of glass, along with their corresponding classification labels. The dataset is commonly used for classification and pattern recognition algorithms.

  The Glass dataset consists of 214 samples, each representing a different type of glass. For each sample, there are nine attributes or features that describe the chemical composition of the glass, such as the percentages of various elements like sodium, magnesium, aluminum, silicon, potassium, calcium, barium, iron, and the refractive index. The classification label indicates the type of glass, which can fall into one of seven categories, including float processed building windows, vehicle windows, containers, tableware, headlamps, and more.

- **AwA2** The AwA2 (Animals with Attributes 2) dataset is a widely used dataset in computer vision and machine learning research, specifically in the field of visual recognition and attribute-based classification. It is designed to facilitate the study of the relationship between visual features and semantic attributes of animals.

  The AwA2 dataset consists of images of 50 different animal classes, with each class having a varying number of images. In total, the dataset contains approximately 37,000 images. The animals span a wide range of species, including mammals, birds, reptiles, and more.

  Each image in the AwA2 dataset is associated with a set of 85 semantic attributes, which describe various characteristics of the animals. These attributes include information such as color, texture, body size, habitat, and behavior. The attribute annotations provide rich semantic information about the animals, enabling researchers to explore and develop algorithms for attribute-based classification and understanding.

  In my experiments I use the feature extracted by Resnet101, which is a 2048-dim vector for each data sample.

[1] Data and code at: https://github.com/Otsuts/CS420_ML_projects/tree/main/project3

1.2 Implementation Details

In my experiments, three models are used to reveal the performance difference of svm and mlp: *MLPSmall*, *MLPBig* and *SVM*, where MLPSmall is a three-layer neural network with hidden size (5,3), MLPBig is a three-layer neural network with hidden size (256,64), and svm is a multi-classification svm with C=1 and linear kernel. The reason why I use a small mlp is that in some dataset e.g. iris, the sample number is small, and a very big neural network may lead to overfitting.

Different test sizes are tried, with testset ratio = 0.2, 0.4, 0.5, 0.6 and 0.8.

1.3 Experimental Results

Table 1 and Figure 1 2 3 demonstrate the model performance on different datasets with varying test sizes, from which several conclusions can be drawn:

- SVM performs well on small scale datasets, outperforming the other two mlp models, showing that when training samples are not sufficient, traditional models can prevent overfitting and get better results.
- On large datasets, MLPSmall model can't get satisfying results due to its simple structure, but on small datasets, it perform better than large mlp under certain test size. Also, MLPLarge model performs slightly better on large scaled datasets, glass and AwA2.
- When test size varies, MLPSmall' s accuracy changes a lot, while other two model's performance doesn't get influenced much, showing that when model gets more complicated, size of test size don't matter that much. Also, results shows that best result is obtained when test size is between 0.4 and 0.6.

Table 1: Model Performance on Different Datasets with Varying Test Sizes

Model	**Test Size**				
	0.2	**0.4**	**0.5**	**0.6**	**0.8**
mlp_big	0.9667	0.9667	0.9600	0.9778	0.9833
	0.6279	0.6471	0.6604	0.6535	0.5976
	0.9286	0.9277	0.9259	0.9229	0.9166
mlp_small	1.0000	0.9833	0.6933	0.6778	0.9500
	0.5581	0.6353	0.6226	0.4803	0.4379
	0.5113	0.4960	0.5112	0.4528	0.4006
svm	1.0000	1.0000	0.9867	0.9889	0.9833
	0.6235	0.6235	0.6038	0.6063	0.5680
	0.9172	0.9134	0.9127	0.9079	0.8937

2 Causal discovery algorithms

2.1 Dataset

Twins dataset is used in this task. The Twin Database of the Global Data Lab provides information on 70,000+ twins in 76 low and middle income countries. It contains three files: twin_pairs_X_3years_samesex.csv includes 50 covariates for the twin pair such as mother and father age and education, health complications and so on. The features which are different between the pair such as sex and birth order are denoted with _0 and _1 for the lighter and heavier twin, respectively. twin_pairs_T_3years_samesex.csv includes the birth weights in grams of both twins in the pair, dbirt_0 and dbirt_1. The lightest always first. I removed all pairs with exactly the same weight. twin_pairs_Y_3years_samesex.csv includes the mortality outcome for both twins, mort_0 and mort_1.

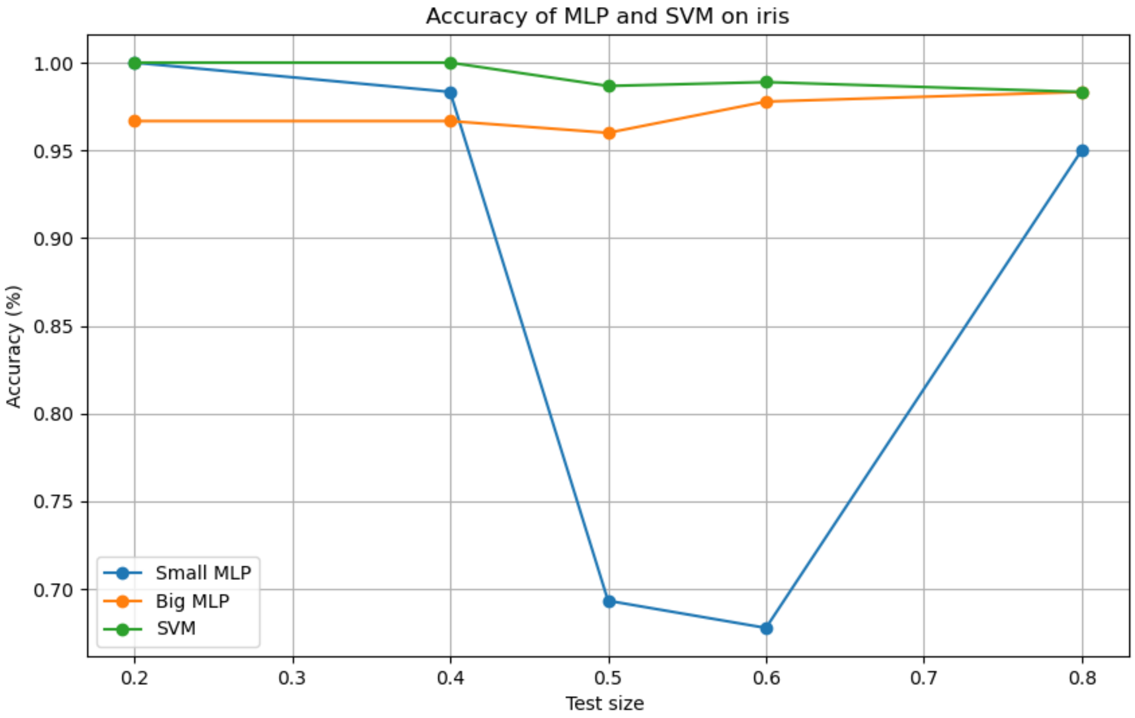

Figure 1: Accuracy of different model on iris dataset

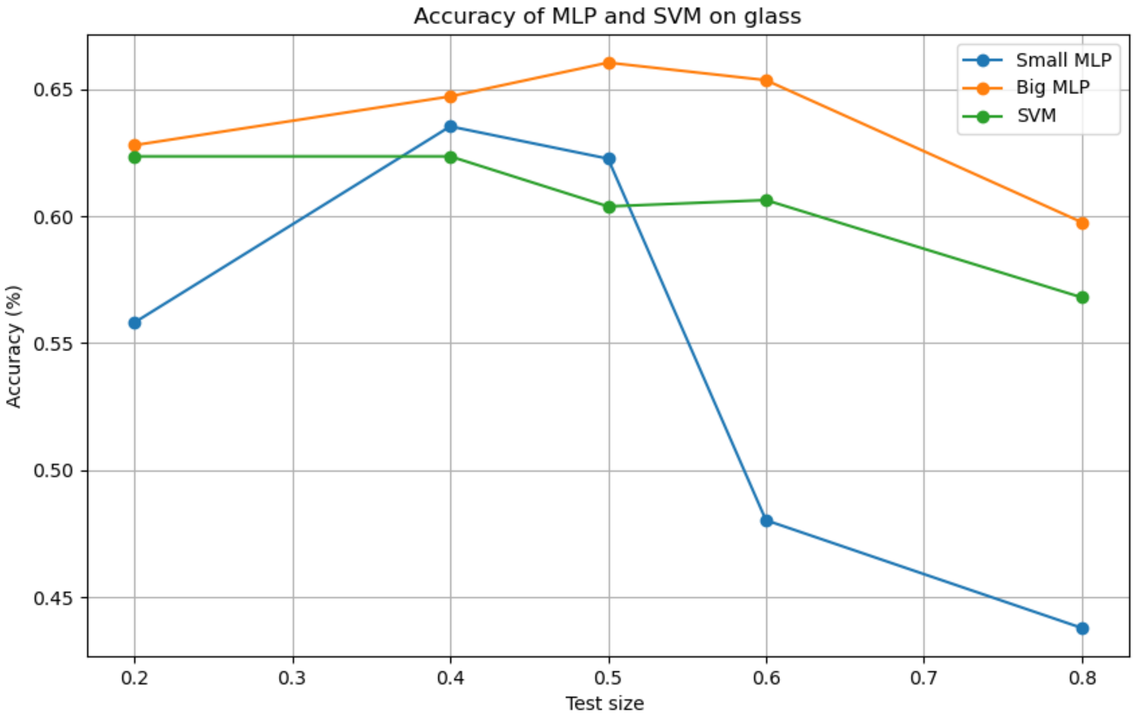

Figure 2: Accuracy of different model on glass dataset

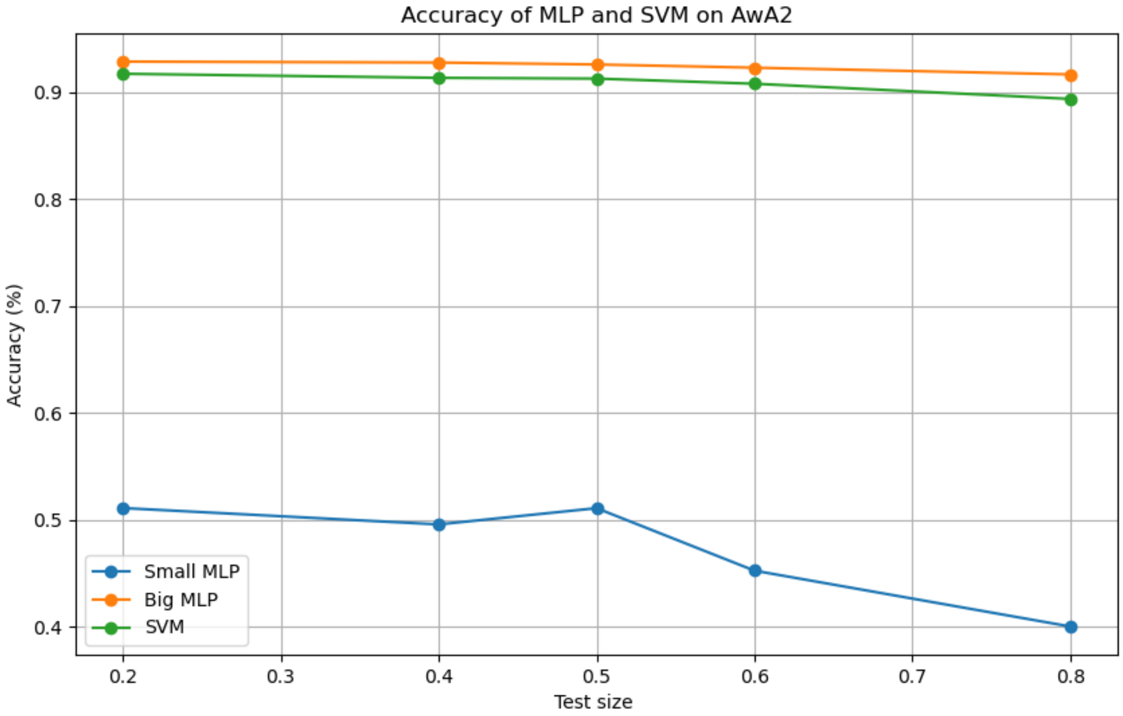

Figure 3: Accuracy of different model on AwA2 dataset

2.2 Algorithm Details

In the twins dataset I tried to investigate the causal relationship between certain features of baby mother and the average twins weights, chosen features and the description is listed below in Table 2, and an example scatter plot of mom age and race and twins weight is shown at Figure 4 and 5

Table 2: Feature choosen for twins dataset

Name	Description
birmon	birth month Jan-Dec
brstate_reg	US census region of brstate
brstate	state of residence NCHS
crace	race of child
csex	sex of child
data_year	year: 1989, 1990 or 1991d
dmar	married
gestat10	gestation 10 categories
mager8	mom age
mrace	mom race
stoccfipb_reg	US census region of stoccfipb
stoccfipb	state of occurence FIPB

I use **lingam** algorithm to measure the causal factor between x and y mentioned above. Lingam is a causal discovery algorithm used to infer causal relationships from observational data. It aims to identify the underlying causal structure among variables by assuming that the relationships are linear and that the variables are non-Gaussian.

The Lingam algorithm is designed to handle scenarios where only observational data is available, meaning that no interventions or randomized experiments have been conducted. It assumes that the

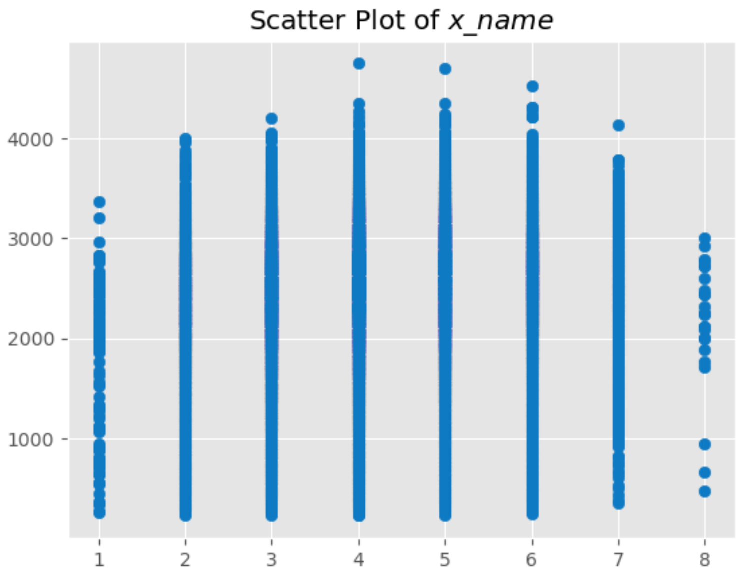

Figure 4: Scatter plot of mom age and twin weight

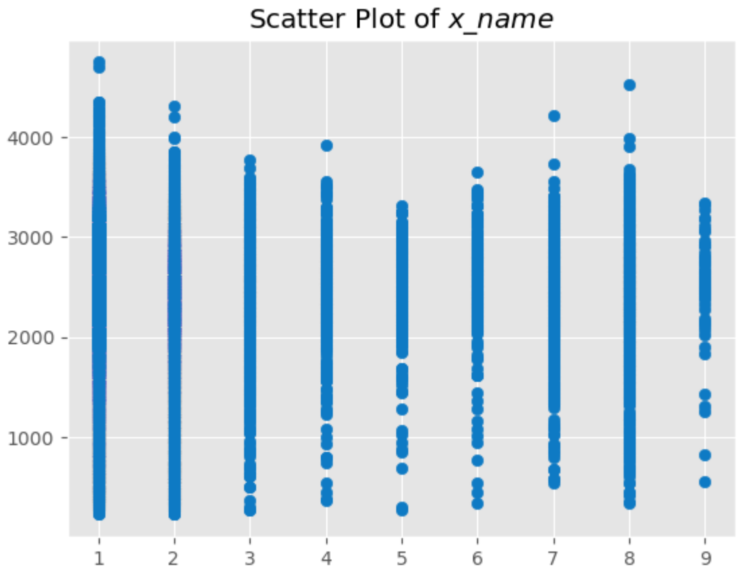

Figure 5: Scatter plot of child race and twin weight

causal relationships can be represented as a directed acyclic graph (DAG), where the arrows indicate the direction of causality between variables.

The Lingam algorithm estimates the causal relationships by combining two main steps: identification and estimation. In the identification step, it determines the causal ordering among the variables based on a series of statistical tests. In the estimation step, Lingam estimates the parameters of the linear structural equation models that best fit the data, given the causal ordering.

2.3 Experimental Results

Experiments are listed as follows at Table 3, from which several interesting conclusions can be drawn: first, **gestat10** has the highest causal scores, showing that the gestation process really matters a lot to baby weights. Next is **dmar**, which is interesting, showing that shotgun marriage influence the situation of mother quite a bit. Then comes **baby sex** and **mother age**, which are reasonable because different baby sex and mother age really lead to baby weights. Other factors may not matter that much, such as the year that the data is obtained, and birth month.

Table 3: Feature choosen and Causal Score

Feature Name	**Score**
birmon	-0.55
brstate_reg	-0.630
brstate	-0.415
crace	-28.984
csex	68.879
data_year	-6.281
dmar	241.818
gestat10	259.056
mager8	73.506
mrace	-32.621
stoccfipb_reg	-0.728
stoccfipb	-0.519