Kunskapskontroll

Rapport MNIST

Markus Ekman

Machine Learning

EC Utbildning

Data Scientist Göteborg

06-04-2023

## 1.0 Introduktion

### 1.1 Bakgrund

Inom området av Machine Learning vill man ta data, analysera den, och skapa en modell som kan förutspå framtida data. Arthur Samuel gav definitionen att “Machine Learning är ett studieområde som ger datorer förmågan att lära sig utan att bli direkt programmerade”. Inom Machine Learning används ofta datasetet MNIST för att utvärdera nya modeller. MNIST är skapat av Yann LeCun, Corinna Cortes och Christopher J.C. Burges.

### 1.2 Syfte och Frågeställning

Syftet med denna rapport är att undersöka vilken modell som ger högst prediktionsförmåga utifrån returnerad precision, recall, och accuracy.

## 2.0 Databeskrivning/EDA

Databasen består av handskrivna siffror från 0 till 9. Databasen består av 60000 träningsexemplar och 10000 testexemplar. Varje bild innehåller 784 pixlar i storleken 28 x 28.

A picture containing text, electronics, calculator, keyboard

Description automatically generated

Bild 2.1 - Exempel på de handskrivna siffrorna. Källa (<https://www.scirp.org/journal/paperinformation.aspx?paperid=80066#:~:text=The%20MNIST%20dataset%20contains%2060%2C000,with%20size%2028%20%C3%97%2028>.)

A screenshot of a computer

Description automatically generated with low confidence

Bild 2.2 – Förekomst av varje siffra i MNIST datasetet. Se jupyter notebook för kod.

## 3.0 Metoder och modeller

För att skynda på processen med att hitta vilken model som ger bättre resultat väljer jag att enbart använda 10% av datan vilket motsvarar 6000 i train och 1000 i test. Eftersom att MNIST redan är uppdelat i 60000 train och 10000 test kommer jag att plocka de första 6000 från de tänkta 60000 och 1000 från de tänkta 10000 test. Det kommer nödvändigtvis inte ge ett helt korrekt svar om man jämför med hela datasetet men det bör ta fram den bästa modellen som man sedan kan använda för att träna hela datasetet med de bästa parametrarna.

De modeller jag har valt att utvärdera är Support Vector Classification, DecisionTrees och RandomForest. Support Vector Classification (SVC) är en av de mest populära modellerna inom Machine Learning. SVC tar vanligtvis väldigt lång tid att bearbeta på stora dataset men eftersom att datan är reducerad går det mycket snabbare att hitta vilka hyperparametrar som bör användas. SVC är en av de mest populära modellerna inom Machine Learning. DecisionTrees är en modell som bryter ner datan kring dess features. Nedbrytningen skapar ett ”djup” (depth) i modellen, där varje djup är mer detaljerade regler för att skilja på datan. RandomForest är, enkelt förklarat, flera DecisionTrees som tillsammans skapar ett bättre resultat. Anledningen till att jag valde att göra både DecisionTrees och RandomForest var att jag hade testat att modellera ett DecisionTree, så var resultaten väldigt låga. Med RandomForest var min hypotes att jag möjligtvis skulle lyckas få fram en bättre prediktionsförmåga.

För att optimera hyperparametrarna har jag använt mig av GridSearch. GridSearch gör att man kan välja några hyperparametrar och mata in värden, för att sedan låta datorn söka igenom vilken kombination som ger bäst resultat. I min rapport har jag valt att göra två gridsearches för både SVC och RandomForest för att finjustera vilka parametrar som ska användas. Vid DecisionTree var accuracy, precision och recall resultaten så pass låga efter den första sökningen att jag valde att släppa modellen helt.

Precision, recall och accuracy används för att utvärdera modellens prediktiv förmåga. Accuracy ger en helhets bild för att visa de korrekta prediktionerna delat med de totala antalet prediktioner som görs. Precision defineras som antalet korrekta positiva (true positive) prediktioner delat med det totala antalet positiva prediktioner (både true och false). Recall är antalet korrekta positiva prediktioner delat med det totala antalet faktiska positiva fall (alltså true positives och false negatives).

## 4.0 Projekt Resultat och Analys

### 4.1 SVC

Följande tabeller visar resultatet från de olika modellerna när det kommer till accuracy score, precision, recall.

|  |  |
| --- | --- |
| Modell | Resultat |
| SVC | 0.918 |
| SVC Grid #1 | 0.947 |
| SVC Grid #2 | 0.947 |
| SVC Fin | 0.978 |

4.1.1 – Accuracy score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Klass | SVC | SVC Grid #1 | SVC Grid #2 | SVC Fin |
| 0 | 0.92 | 0.97 | 0.97 | 0.99 |
| 1 | 0.99 | 0.98 | 0.98 | 0.99 |
| 2 | 0.86 | 1.00 | 1.00 | 0.97 |
| 3 | 0.92 | 0.95 | 0.95 | 0.98 |
| 4 | 0.94 | 0.90 | 0.90 | 0.97 |
| 5 | 0.96 | 0.99 | 0.99 | 0.98 |
| 6 | 0.95 | 0.97 | 0.97 | 0.98 |
| 7 | 0.90 | 0.96 | 0.96 | 0.98 |
| 8 | 0.85 | 0.84 | 0.84 | 0.96 |
| 9 | 0.89 | 0.92 | 0.92 | 0.97 |
| Average | 0.918 | 0.948 | 0.948 | 0.977 |

4.1.2 – Precision Score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Klass | SVC | SVC Grid #1 | SVC Grid #2 | SVC Fin |
| 0 | 0.98 | 0.98 | 0.98 | 0.99 |
| 1 | 0.98 | 0.99 | 0.99 | 1.00 |
| 2 | 0.94 | 0.89 | 0.89 | 0.98 |
| 3 | 0.89 | 0.94 | 0.94 | 0.98 |
| 4 | 0.92 | 0.98 | 0.98 | 0.98 |
| 5 | 0.89 | 0.95 | 0.95 | 0.97 |
| 6 | 0.91 | 0.97 | 0.97 | 0.98 |
| 7 | 0.91 | 0.92 | 0.92 | 0.97 |
| 8 | 0.87 | 0.93 | 0.93 | 0.97 |
| 9 | 0.88 | 0.91 | 0.91 | 0.96 |
| Average | 0.917 | 0.946 | 0.946 | 0.978 |

4.1.3 – Recall Score

### 4.2 DecisionTree

Efter gridsearch identifierades de bästa parametrarna som:

Chart

Description automatically generated with low confidence

|  |  |
| --- | --- |
| Modell | Resultat |
| DecisionTree | 0.758 |
| DecisionTree Grid | 0.763 |

4.2.1 – Accuracy Score

|  |  |  |
| --- | --- | --- |
| Klass | DecisionTree | DecisionTree Grid |
| 0 | 0.78 | 0.85 |
| 1 | 0.88 | 0.88 |
| 2 | 0.77 | 0.79 |
| 3 | 0.71 | 0.70 |
| 4 | 0.83 | 0.79 |
| 5 | 0.67 | 0.68 |
| 6 | 0.81 | 0.87 |
| 7 | 0.70 | 0.71 |
| 8 | 0.65 | 0.66 |
| 9 | 0.72 | 0.71 |
| Average | 0.752 | 0.764 |

|  |  |  |
| --- | --- | --- |
| Klass | DecisionTree | DecisionTree Grid |
| 0 | 0.88 | 0.91 |
| 1 | 0.91 | 0.87 |
| 2 | 0.68 | 0.71 |
| 3 | 0.65 | 0.63 |
| 4 | 0.75 | 0.73 |
| 5 | 0.64 | 0.68 |
| 6 | 0.82 | 0.78 |
| 7 | 0.77 | 0.82 |
| 8 | 0.70 | 0.75 |
| 9 | 0.76 | 0.78 |
| Average | 0.756 | 0.766 |

4.2.2 – Precision score 4.2.3 – Recall score

### 4.3 RandomForest

Hyperparametrarna för RF är default med undantaget för random\_state som är satt till 40.

Efter gridsearch #1 är bästa parametrarna:

A picture containing graphical user interface

Description automatically generated

Efter gridsearch #2 är bästa parametrarna:

Graphical user interface, text

Description automatically generated with medium confidence

RF Fin #1 och RF Fin #2 skiljer sig i att max\_depth är begränsad till 10 i RF Fin #2.

|  |  |
| --- | --- |
| Modell | Resultat |
| RF | 0.932 |
| RF Grid #1 | 0.933 |
| RF Grid #2 | 0.936 |
| RF Fin #1 | 0.970 |
| RF Fin #2 | 0.970 |

4.3.1 – accuracy score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Klass | RF | RF Grid #1 | RF Grid #2 | RF Fin #1 | RF Fin #2 |
| 0 | 0.94 | 0.95 | 0.95 | 0.97 | 0.97 |
| 1 | 1.00 | 0.99 | 0.99 | 0.99 | 0.97 |
| 2 | 0.95 | 0.93 | 0.96 | 0.96 | 0.94 |
| 3 | 0.94 | 0.92 | 0.93 | 0.97 | 0.94 |
| 4 | 0.93 | 0.94 | 0.94 | 0.98 | 0.96 |
| 5 | 0.92 | 0.93 | 0.91 | 0.98 | 0.97 |
| 6 | 0.93 | 0.95 | 0.93 | 0.98 | 0.96 |
| 7 | 0.94 | 0.92 | 0.93 | 0.97 | 0.96 |
| 8 | 0.90 | 0.87 | 0.90 | 0.96 | 0.94 |
| 9 | 0.85 | 0.90 | 0.89 | 0.95 | 0.89 |
| Average | 0.93 | 0.93 | 0.933 | 0.971 | 0.95 |

4.3.2 – Precision score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Klass | RF | RF Grid #1 | RF Grid #2 | RF Fin #1 | RF Fin #2 |
| 0 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| 1 | 0.99 | 0.98 | 0.98 | 0.99 | 0.99 |
| 2 | 0.94 | 0.91 | 0.93 | 0.97 | 0.94 |
| 3 | 0.91 | 0.90 | 0.90 | 0.97 | 0.94 |
| 4 | 0.89 | 0.94 | 0.93 | 0.97 | 0.93 |
| 5 | 0.94 | 0.94 | 0.93 | 0.96 | 0.93 |
| 6 | 0.95 | 0.95 | 0.94 | 0.98 | 0.97 |
| 7 | 0.89 | 0.93 | 0.94 | 0.96 | 0.93 |
| 8 | 0.88 | 0.85 | 0.85 | 0.96 | 0.93 |
| 9 | 0.94 | 0.94 | 0.96 | 0.95 | 0.94 |
| Average | 0.932 | 0.933 | 0.935 | 0.97 | 0.949 |

4.3.3 – Recall score

### 4.4 Analys

Det blir snabbt väldigt tydligt att DecisionTree modellen inte kommer kunna ge en tillräckligt bra prediktiv förmåga om man jämför med SVC och RandomForest. Om man kollar på resultatet av RF Fin #1 som representerar den första RandomForest modellen med hela datasetet, så går det att se att den gör bäst ifrån sig när det kommer till precision och recall och har samma accuracy score som RF Fin #2. Däremot om man jämför RF Fin #1 med SVC Fin så presterar SVC modellen bättre (om även bara med någon decimalpunkt). Något som också är oroande för RandomForest modellen är risken för overfitting när max\_depth ökas till 17.

## 5.0 Slutsats och vidareutveckling

SVC modellen bör anses som den bättre modellen när det kommer till att prediktera och korrekt identifiera siffran i en bild. Detta då modellen har högre accuracy score, precision samt recall. Det finns fortfarande mer att göra för att vidareutveckla denna rapport. Fler hyperparametrar hade kunnat finjusteras i RandomForest, för att säkerställa att ingen overfitting sker. Detta hade potentiellt kunnat ge en verkligare bild av hur väl RandomForest modellen fungerar. .

**1.Utmaningar du har haft under arbetet samt hur de är hanterade**

Jag har i stundvis varit lite osäker på hur jag ska gå tillväga med uppgiften. Men efter att ha repiterat visa delar av materialet (föreläsningar och läshänvisningar) samt ställt frågor både till Antonio och kurskamrater har jag fått mer klarhet. Jag har dessutom googlat fram svar på mina egna frågor eller när jag blir osäker på hur vissa hyperparametrar fungerar.

**2.Vilket betyg du anser att du skall ha och varför.**

Jag anser att det arbete som jag har gjort uppnår i alla fall ett G. Jag tycker jag har löst uppgiften på G nivå och försökt att nå VG nivån. Jag tror att vissa steg som jag gjort är lite för grundliga för att det skulle kunna nå VG nivån.

**3.Tips du hade “gett dig själv” i början av kursen**

Jobba mer på repetition av kursmaterialet och att bli mer effektiv med tiden som är utanför föreläsningarna.