Uppgift 1: Maskininlärning

Läs nedanstående text om maskininlärning och svara på frågorna utgående från texten och dina egna kunskaper.

Maskininlärning är det område inom artificiell intelligens, med vars metoder vi kan lösa uppgifter som har att göra med implementation av informationssystem, utan att vi i systemet direkt programmerar regler för att lösa uppgiften. I maskininlärning strävar man till att på basen av **inlärningsdata** producera en **modell**, som man kan dra nytta av för att utföra en uppgift. Till exempel inom bildigenkänning kan inlärningsdatan bestå av en mängd bilder i vilka det objekt som ska identifieras förekommer och en mängd där objektet inte finns. Med hjälp av maskininlärningsmetoden som utvecklats för bildigenkänning kan man från dylika data skapa en modell, med vilken man kan försöka skilja mängderna från varandra.

Maskininlärningsmetoderna kan indelas enligt vilka inlärningsdata de använder. **Övervakade** metoder utgör en central kategori, vilkas inlärningsdata består av **indata** (input) och motsvarande **utdata** (output). Målet för övervakad maskininlärning är att från inlärningsdatan skapa en modell, som från nya indata skapar korrekta utdata. Till exempel skräppostens filtrering kan man utföra med övervakad maskininlärning så, att indatan består av e-postmeddelanden och utdatan av informationen om vilka meddelanden som är skräppost. En modell som lärt sig från sådana data kan till exempel innehålla information om hur ofta varje ord förekommer i skräppost jämfört med övrig e-post. Övervakade metoder kan ännu grupperas på basen av värdena på deras utdata. **Klassificering** är övervakad maskininlärning, där värdena förutspår en ändlig mängd på förhand bestämda kategorier (klasser). Filtrering av skräppost är ett exempel på klassificering, där värdena på utdatan representerar klasserna `skräp` och `icke-skräp`. **Regression** igen är övervakad maskininlärning, där värdena på utdatan är kontinuerliga. Att bestämma temperaturen är till exempel en regressionsuppgift: temperaturens möjliga värden finns på en kontinuerlig skala.

I **oövervakad** maskininlärning innehåller inte träningsdatan utdata och målet är att modellera indatans struktur till exempel på basen av vilka indata som är sinsemellan lika. **Klustring** är ett oövervakat inlärningsuppdrag, som motsvarar övervakad klassificering, men där målet är att gruppera ihop liknande indata. Genom att utföra klustring på data som består av nyhetstexter kan det till exempel vara möjligt att hitta grupper, som behandlar politik, sport och väder. Å andra sidan, eftersom det önskade resultatet inte är förutbestämt i oövervakad inlärning, kan klustring även skapa oväntade grupper: klustringsmetoden kan till exempel i stället för att gruppera nyheter enligt ämne, gruppera dem på basen av om nyhetstexterna är långa eller korta.

På grund av att maskininlärningsmetoderna inte direkt kan behandla alla möjliga indataformer, är datarepresentationen för maskininlärning en central fråga för tillämpningen av dessa metoder. I stället för att maskininlärningsmetoden direkt ges indata till exempel i form av textdokument skrivna på svenska, representeras indatan typiskt i form av **särdrag**: varje särdrag representerar någon egenskap hos indatan och värdet på särdraget anger till vilken grad denna egenskap beskriver indatan. Till exempel för att klassificera en text är det möjligt att för varje ord i språket bestämma ett särdrag och att sätta dess värde på basen av hur många gånger varje ord förekommer i vart och ett av dokumenten. I stället för att särdragen bestäms så här i förväg, kan man även lära sig indatans representation av datan med hjälp av **inlärning av särdrag**.

I motsats till hur det är i människornas inlärningsprocess, är det inte en utmaning i maskininlärning att den inlärda kunskapen glöms: eftersom maskinen kan lära sig inlärningsdatan utantill, är det lätt att skriva ett program som alltid returnerar rätta utdata för sådana indata som

förekommer i inlärningsdatan. Ett sådant program är ändå inte användbart, om det inte kan producera önskade utdata även för sådana indata som inte förekommer i inlärningsdatan. Till exempel ett skräppostfilter, som endast identifierar sådana meddelanden som skräp, vilka redan tidigare har klassificerats som skräppost, skulle missta sig på varje nytt skräppostmeddelande, som skiljer sig ens med ett tecken från inlärningsdatans meddelanden. Maskininlärningens centrala mål är ju skapandet av sådana modeller, som förutspår önskade resultat för nya indata utöver inlärningsdatan, eller som kan **generalisera** inlärningsdatan från exemplen. För utvärderingen används **testdata**, som är fristående från inlärningsdatan och som inte har utnyttjats vid utvecklingen av metoden. För att den skapade modellen ska vara användbar för praktiska uppgifter och dess utvärdering ska ge ett realistiskt resultat angående dess noggrannhet, är det viktigt att såväl inlärningsdata som testdata representerar de data på vilka den färdiga modellen kommer att tillämpas.

Maskininlärning ger inte alltid önskat resultat och det finns flera utmaningar i tillämpningen av dess metoder. Många av de här utmaningarna har att göra med datan. Om inte mängden inlärningsdata är tillräckligt stor, kan det vara omöjligt att på basen av den lära sig modeller, som ger önskade resultat: till exempel för att utveckla en klassificeringsmodell för text behövs ofta hundratals eller tusentals exempel på varje klass. Om datan igen inte representerar uppgiften för vilken metoden utvecklas, kan modellen bli förvrängd och den kan misslyckas i praktiken. Om exempelvis endast klart uttalat högspråk används för att utveckla en taligenkänningsmodell, kan modellen misslyckas i en bullrig omgivning och med att känna igen dialekter.

Det är möjligt att tillämpa många olika metoder på maskininlärningsuppgifter och för att förverkliga dessa metoder finns det verktyg på många olika programmeringsspråk. Många maskininlärningsmetoder baserar sig på statistiska metoder. Till exempel kan en enkel klassificeringsmetod för skräppost skapas genom att räkna hur ofta varje ord förekommer i meddelandena i inlärningsdatan för varje klass (`skräp`/ `icke-skräp`) och genom att på basen av sannolikheterna per ord beräkna en uppskattning av om det nya meddelandet är skräp. Många maskininlärningsmetoder som har utvecklats under de senaste åren baserar sig på **neuronnät**, som modellerar sammankopplade enkla beräkningsenheter. Av dessa modellerar **djupinlärningsmetoderna**, som uppnått goda resultat bl.a. inom bildigenkänning, stora neuronnät bestående av till och med miljontals beräkningsenheter. Dessa metoder kan lära sig mycket komplicerade avbildningar från indata till utdata, men för att träna dem kan det å andra sidan krävas en mycket stor mängd data.

Frågorna

Vilka maskininlärningsmetoder används i följande system? *Välj för varje beskrivet system i frågorna 1.1 – 1.4 de rätta alternativen. Välj endast de alternativ som är motiverade direkt på basen av beskrivningen och märk att det kan finnas flera rätta alternativ.*

Poängsättning: Exakt rätt svar ger maximipoäng. För varje saknat eller överflödigt alternativ dras poäng av. Varje frågas minimipoäng är 0.

Fråga 1.1. (2 p.)

Meningen med systemet A är att känna igen författaren till texten på basen av dess innehåll. Systemet baserar sig på en statistisk modell, för vilken man från ett omfattande textmaterial har sammanställt information om hur ofta varje författare använder ett visst ord eller en viss ordkombination. När ny text matas in i systemet, avgör det på basen av modellen vem av

personerna representerade i materialet som mest sannolikt har skrivit texten. Begränsningen med systemet och materialet är, att varje text bör ha exakt en författare.

- ☐ Övervakad inlärning
- ☐ Oövervakad inlärning
- ☐ Klassificering
- ☐ Regression
- ☐ Klustring
- ☐ Ingen av de ovanstående

Fråga 1.2. (2 p.)

Systemet B är utvecklat för att stöda aktiehandel. Systemet innehåller en modell, som försöker förutspå det framtida värdet för varje aktie på marknaden på basen av hur aktiernas värden har förändrats under det senaste året. Modellen baserar sig på en matematisk funktion, vars parametrar har anpassats genom att använda data från en databas som representerar den historiska utvecklingen av aktiernas värden.

- ☐ Övervakad inlärning
- ☐ Oövervakad inlärning
- ☐ Klassificering
- ☐ Regression
- ☐ Klustring
- ☐ Ingen av de ovanstående

Fråga 1.3. (2 p.)

Med hjälp av systemet C kan man från mängden e-post söka alla meddelanden i vilka ett av användaren definierat ord förekommer. Systemet baserar sig på en databas, i vilken man har sammanställt alla ord som förekommer i mängden och för varje ord alla meddelanden i vilka ordet förekommer. Systemet stöder också sökformer, där man förutom ordet man fått som indata även söker ord från databasen vilka systemets utvecklare på basen av en digital ordbok har bestämt att betyder samma sak.

- ☐ Övervakad inlärning
- ☐ Oövervakad inlärning
- ☐ Klassificering
- ☐ Regression
- ☐ Klustring
- ☐ Ingen av de ovanstående

Fråga 1.4. (2 p.)

Systemet D är utvecklat för att organisera textmaterial. Systemet presenterar för varje text i dess databas fördelningen av de ord som förekommer i texten och skapar från dem grupper av texter med sinsemellan liknande fördelningar. Systemet namnger varje grupp enligt de ord som förekommer oftare i dess texter än i andra gruppers texter.

- ☐ Övervakad inlärning
- ☐ Oövervakad inlärning
- ☐ Klassificering
- ☐ Regression
- ☐ Klustring
- ☐ Ingen av de ovanstående

Av vilka orsaker relaterade till maskininlärning ger följande system inte de önskade resultaten? *Välj för varje beskrivet system i frågorna 1.5 – 1.7 lämpliga alternativ, definiera problemet eller problemen noggrannare med egna ord och berätta hur de kunde lösas. Svara kort på varje fråga med högst 40 ord. Välj endast alternativ som kan motiveras direkt från beskrivningen, och märk att det kan finnas flera rätta alternativ.*

Modellsvar: Fråga 1.x: Val av inlärningsdata – Inlärningsdatan för e-postens klassificeringsmetod består endast av skräppostmeddelanden. Inlärningsdatan bör utvidgas så att de också innehåller exempel på meddelanden, som inte är skräppost.

Poängsättning: Exakt rätt svar ger maximipoäng. För varje saknat eller överflödigt alternativ dras poäng av. Du bör motivera ditt svar. Om motiveringen saknas får du automatiskt 0 poäng för svaret. Varje frågas minimipoäng är 0.

Fråga 1.5. (2 p.)

E utvecklar ett system att känna igen djurarter från fotografier genom att använda språket Python och de nyaste djupinlärningsmetoderna. För utvecklingen av systemet sammanställs inlärningsdata, i vilka det finns ett fotografi av varje djurart som systemet försöker känna igen, samt testdata sammanställda på motsvarande sätt. Systemet tränas att klassificera bilden som används som indata till den klass som djurarten på bilden motsvarar. Systemet misslyckas ändå att ge önskade resultat oberoende av hur det förverkligas och tränas.

- ☐ Val av inlärningsdata
- ☐ Val av testdata
- ☐ Val av programmeringsspråk
- ☐ Val av maskininlärningsmetod
- ☐ Val av särdrag

Fråga 1.6. (2 p.)

F äger ett omfattande fotografimaterial, som är ordnat enligt motiv, så att varje bild har exakt ett motiv. F håller på att utvidga materialet med bilder tagna med samma motiv och hoppas att systemet baserat på maskininlärning ska hjälpa att ordna de nya bilderna enligt motiv. För att stöda systemets utveckling sammanställer F för varje motiv en samling fotografier som föreställer motivet i fråga och skickar den till en konsultfirma som specialiserat sig på programutveckling. Firman skapar ett system som använder klustringsmetoder från maskininlärningsbiblioteket, vilket förverkligats med språket Lua. När systemet tas i bruk, konstateras ändå att den gruppering som det skapar avviker helt från det på förhand önskade.

- ☐ Val av inlärningsdata
- ☐ Val av testdata
- ☐ Val av programmeringsspråk
- ☐ Val av maskininlärningsmetod
- ☐ Val av särdrag

Fråga 1.7. (3 p.)

G utvecklar en ansiktsigenkänningsmetod som stöd för en mobil-applikation genom att använda Java och ett maskininlärningsbibliotek baserat på neuronnät. För att samla data tas en omfattande mängd bilder i studio-omgivning, på vilka många olika ansikten förekommer, och en motsvarande mängd, där ansikten inte förekommer. Materialet indelas slumpmässigt i inlärnings- samt testdata och det på särdragsinlärning baserade systemet, som tränats på inlärningsdatan, konstateras klassificera testdatans material med hög noggrannhet. Trots detta är systemets ansiktsigenkänningsfunktion mycket opålitlig enligt mobilapplikationens användarfeedback.

- ☐ Val av inlärningsdata
- ☐ Val av testdata
- ☐ Val av programmeringsspråk
- ☐ Val av maskininlärningsmetod
- ☐ Val av särdrag