Master´s Programme in Statistics and Data Mining

120 hp

Master´s Programme in Statistics and Data Mining

F7MSG

Gäller från:

Fastställd av
Fakultetsstyrelsen för filosofiska fakulteten

Fastställandedatum
2015-06-16

Syfte

Som en följd av en ständigt ökande dator- och lagringskapacitet blir databaserna allt större och mer komplexa. Detta program syftar på att omvandla stora eller komplexa datamängder till kunskap. Statistisk modellering och analys är integrerade med maskinlära, data mining och visualisering för att forma en grundlig bas för ett professionellt arbete med datahantering och dataanalys eller för att göra en forskningskarriär.

Mål

Efter genomgången utbildning ska den studerande kunna

- utvinna och strukturera stora och komplexa datamängder
- utforska, sammanfatta och presentera stora och komplexa datamaterial med hjälp av statiska, interaktiva och dynamiska visualiseringsredskap
- välja en lämplig modell för ett givet statistiskt problem och en datamängd
- upptäcka och statistiskt granska tidigare okända mallarna och trenderna i en datamängd
- använda avancerade statistiska och dataminingsmjukvaror för att analysera stora datavolymer
- implementera dataanalysmodeller i något programmeringsspråk
- kombinera datamaterialet och andra källor av prioriinformation för att förbättra inferensen och prediktionsförmågan
- ge exempel på tillämpningsområden där analys av stora och komplexa datamängder behövs
- presentera en uppsats som innehåller en teoretisk eller en tillämpad studie av en komplex datamängd

Kunskap och förståelse

Efter genomgången utbildning ska den studerande

- visa kunskap och förståelse inom huvudområdet statistik, inbegripet såväl brett kunnande inom området som väsentligt fördjupade kunskaper inom dess bransch, data mining, samt fördjupad insikt i aktuellt forsknings- och utvecklingsarbete, och
- visa fördjupad metodkunskap inom statistik.

Fördjupade kunskaper inom data mining skall inkludera moderna kraftiga metoder för prediktion, klassificering, klustring, Bayesianska metoder och associationsanalys.

Färdighet och förmåga

Efter genomgången utbildning ska den studerande

- visa förmåga att kritiskt och systematiskt integrera kunskap och att analysera, bedöma och hantera komplexa företeelser, frågeställningar och situationer även med begränsad information,
- visa förmåga att kritiskt, självständigt och kreativt identifiera och formulera frågeställningar, att planera och med adekvata metoder genomföra

kvalificerade uppgifter inom givna tidsramar och därigenom bidra till kunskapsutvecklingen samt att utvärdera detta arbete,
- visa förmåga att i såväl nationella som internationella sammanhang muntligt och skriftligt klart redogöra för och diskutera sina slutsatser och den kunskap och de argument som ligger till grund för dessa i dialog med olika grupper, och
- visa sådan färdighet som fordras för att delta i forsknings- och utvecklingsarbete eller för att självständigt arbeta i annan kvalificerad verksamhet.

Värderingsförmåga och förhållningssätt

Efter genomgången utbildning ska den studerande

- visa förmåga att inom huvudområdet statistik göra bedömningar med hänsyn till relevanta vetenskapliga, samhälleliga och etiska aspekter samt visa medvetenhet om etiska aspekter på forsknings- och utvecklingsarbete,
- visa insikt om vetenskapens möjligheter och begränsningar, och speciellt möjligheter och begränsningar av statistik och data mining, dess roll i samhället och människors ansvar för hur den används, och
- visa förmåga att identifiera sitt behov av ytterligare kunskap och att ta ansvar för sin kunskapsutveckling.

Innehåll

Programmet kombinerar kurser i statistik, datavetenskap och matematik. Obligatoriska kurser, inledande kurser och en masteruppsats på 30 hp främjar progression och djupet av förståelsen. Inledande kurser erbjuds för att fylla i brister i studenternas kunskaper och för att se till att studenterna är ordentligt förberedda till programmets kurser.

Obligatoriska kurser

AKADEMISKA STUDIER PÅ AVANCERAD NIVÅ, 3 HP (ges termin 1)
Målet för denna kurs är att förbereda studenterna för akademiska studier på avancerad nivå, samt att lära ut ett akademiskt förhållningssätt i stort. En grundläggande ambition är att tillhandahålla väsentliga redskap för att studera på avancerad nivå i Sverige. Dessutom kommer programspecifika moment diskuteras.

INTRODUKTION TILL MASKININLÄRNING (ges termin 1)
Grundläggande koncept inom maskininlärning och data mining. Bayesiansk och frekventastes modellering, modelval. Linjär regression och regularisering. Linjär diskriminantanalys och logistisk regression. Bagging och boosting. Splines, generaliserade additiva modeller, beslutsträd och random forest. Kernel utjämning och stödvektormaskiner. Gaussiansk process.

DATA MINING: KLUSTRING OCH ASSOCIATIONSANALYS, 15 HP (ges termin 2)
Statistiska principer och redskap för uppdelning av objekt i grupper och utvinning av samband som är gömda i stora datamängder. Partitionell och hierarkisk klustring. Klustervärdering. Associationsanalys med hjälp av enhetsmängder och associationsregler. Utvärdering av associationsregler.

VETENSKAPSFILOSOFI, 3 HP (ges termin 2)
Naturlagarna och vetenskapliga modeller. Samband mellan teorier och observationer. Krafter som påverkar vetenskapliga förändringar.

BAYESIANSKA METODER, 6 HP (ges termin 2)
Bayes sats för att kombinera datamängder med prioriinformation. Bayesiansk analys av konjugerade modeller. Markov Kedjor Monte Karlo för Bayesianska beräkningar. Bayesianska modellvalet.

DATORINTENSIVA STATISTISKA METODER, 6 HP (ges termin 2)
Datorernas aritmetik. Slumptalgenererings- och simuleringsmetoder. Markov Kedjor Monte Karlo. Numerisk linjär algebra. Optimeringsmetoder i statistik.

Profilkurser

VISUALISERING, 6 HP (ges termin 1 för studenter antagna jämna år och termin 3 för studenter antagna ojämna år)
Avancerade visualiseringsmetoder för stora och komplexa datamängder. Interaktiva och dynamiska statistiska diagram. Visualisering av spatial information.

AVANCERAD MASKININLÄRNING, 6 HP (ges termin 3)
Bayesianska nätverk och dolda Markovmodeller. State-space modeller och slumpfält. Neurala nätverk. Principer av djupinlärning och dess redskap: djupa neurala nätverk, Boltzman maskiner.

TIDSSERIEANALYS, 6 HP (ges termin 1 för studenter antagna ojämna år och termin 3 för studenter antagna jämna år)
Tidsseriedekomposition. Autokorrelation och partiell autokorrelation. Prognoser med hjälp av regression av tidsserier, ARIMA modeller och transferfunktioner. Interventionsanalys. Trendutvinning.

MULTIVARIATA STATISTISKA METODER, 6 HP (ges termin 1 för studenter antagna ojämna år och termin 3 för studenter antagna jämna år)
Analys av korrelation- och kovariansstrukturer, inklusive principalkomponennter, faktoranalys, och kanonisk korrelation. Klassificering- och diskrimineringsmetoder. Flerdimensionell inferens.

SANNOLIKHETSLÄRA, 6 HP (ges termin 3)
Flerdimensionella slumpvariabler och betingade sannolikheter. Fördelningar av största och minsta värden i ett stickprov. Karaktäristiska funktioner och transformer. Multivariat normalfördelning. Sannolikhetsrelaterade konvergensbegrepp.

STATISTISK RESULTATVÄRDERING, 6 HP (ges termin 3)
Resonemang med sannolikheter och likelihood-teori. Bayesiansk hypotesprövning. Bayesianska nätverk. Grunder i statistisk beslutsteori och influensdiagram. Värdering av forensiska resultat. Känslighetsanalys.

Kompletterande kurser

WEBBPROGRAMMERING OCH INTERAKTIVITET, 4 HP (ges termin 2)
En översikt av de allmänna webbredskap, inklusive HTML, DHTML, XML, PHP.

NEURALA NÄTVERK OCH INLÄRNINGSSYSTEM, 6 HP (ges termin 2)
Oövervakad inlärning: principalkomponentanalys, analys avoberoende komponenter, vektorkvantifiering. Övervakad inlärning: neurala nätverk, radiala basfunktioner, support vector machines. Förstärkningslära: Markovprocesser, Q-

learning, genetiska algoritmer.

DATA MINING PROJEKT, 6 HP (ges termin 3)
En projektkurs där den studerande anger, implementerar och redovisar en data mining-algoritm för ett valt data-mining problem.

TEXTMINING, 6 HP (ges termin 3)
Utvinning av textinformation från olika källor. Textbearbetning med hjälp av beräkningslingvistiska metoder. Statistiska modeller för textklassificering och textprediktion.

OPTIMERINGSLÄRA, 6 HP (ges termin 3)
Grundläggande principer inom optimering såsom matematisk modellering, optimeringsvillkor, konvexitet, känslighetsanalys och Lagrange avspänning. Grundläggande principer för linjär och icke-linjär optimering, och heltals- och nätverksoptimering.

DATABASMETODER, 6 HP (ges termin 3)
En databashanterare (DBMS). Metoder för datamodellering och databasdesign. ER-diagram, relationsdatabaser och datasturkturer för databaser. Dataarkitektur och urvalsspråk för relationsdatabaser. Relationsalgebra and urvalsoptimering.

BIOINFORMATIK, 6 HP (ges termin 1 för studenter antagna jämna år och termin 3 för studenter antagna ojämna år)
Grundläggande molekylär biologi. Grundläggande statistik för biologi. Hidden Markov modeller. QTL-mapping. Sequence alignment. Analys av Nästa Generationens sekventiella data. Microarray analys.

Inledande kurser

STATISTISKA METODER, 6 HP (ges termin 1)
Sannolikhetsbegreppet. Slumpvariabel, vanliga statistiska fördelningar och deras egenskaper. Punkt- och intervallskattning. Hypotesprövning. Enkel och multipel linjär regression. Sampling.

AVANCERAD R PROGRAMMERING, 6 HP (ges termin 1)
R miljö. Grundläggande programmeringsmetoder. Språkelement i R: variabler, vektorer, dataramar. Redskap i R: operatorer, loopar, villkor, funktioner. Importing data från text och webben. Debuggning, parallell programmering och prestandaförbättringsredskap. Statistiska och datamining redskap i R. Grafiska funktioner.Objektorienterad programmering. Effektivitetsförbättring och parallell programmering.

MASTERUPPSATS, 30 HP
En teoretisk eller en tillämpad studie av en komplex datamängd med hjälp av statistiska och dataminingsmetoder.

Undervisnings- och arbetsformer

Programmets kurser består av föreläsningar, datorlaborationer och seminarier. Föreläsningarna ägnas åt teorier, begrepp och metoder. Datorlaborationer ger praktisk erfarenhet av dataanalys och andra metoder. Seminarier ägnas åt studentpresentationer och diskussioner av uppgifter. Kurserna vars namn innehåller ”projekt” har endast handledning.

Examination

Kurser som omfattar minst 4,5 hp examineras genom en eller flera uppgifter och en skriftlig examination. Projektkurser och masteruppsatsen examineras genom en skriftlig rapport och genom muntligt försvar av densamma.

Betyg

Betyg på kurs anges i respektive kursplan.

Förkunskapskrav

För behörighet till programmet krävs en kandidatexamen i något av följande ämnen: statistik, matematik, tillämpad matematik, datavetenskap, teknik eller motsvarande examen. Utöver detta, erfordras kurser i kalkyl, linjär algebra, statistik och programmering.

Den som söker till utbildningen måste skicka in ett motiveringsbrev på engelska, som visar varför hon eller han är intresserad av att följa utbildningen. Motiveringsbrevet ska dessutom innehålla en sammanfattning av den sökandes examensuppsats eller -projekt. Sökande vilka har en examen som inte omfattar ett examensarbete ska i motiveringen beskriva tidigare studier och akademisk verksamhet av relevans för den utbildning som söks.

Engelska B/Engelska 6 eller motsvarande.

Tillträdeskrav till högre termin eller kurser

För att bli behörig till termin 2 skall den studerande ha uppnått minst 6 hp på de kurser som ingår i termin 1.

För att bli behörig till termin 3 skall den studerande ha uppnått minst 40 hp på de kurser som ingår i termin 1 och 2

För att bli behörig till termin 4 skall den studerande ha uppnått minst 65 hp på de kurser som ingår i termin 1, 2 och 3, inklusive alla programmets obligatoriska kurser.

Examenskrav

En student inom programmet kan erhålla ett examensbevis med beteckningen Filosofie masterexamen med huvudområdet Statistik givet att studenten har avslutat kurser motsvarande 90 högskolepoäng som inkluderar obligatoriska kurser motsvarande 42 högskolepoäng, inledande kurser motsvarande minst 6 högskolepoäng, profilkurser motsvarande minst 12 högskolepoäng och eventuellt några kompletterande kurser. Studenten skall ytterligare ha avslutat den obligatoriska masteruppsatskursen som omfattar 30 högskolepoäng.

Examensbevis utfärdas av Filosofiska fakultetsstyrelsen, efter begäran av den studerande.

Examensbenämning på svenska

Filosofie masterexamen i huvudområdet statistik

Examensbenämning på engelska

Master of Science (120 Credits) with a major in Statistics

Särskild information

Meriter

De särskilda behörighetskrav anses som icke uppfyllda om examens medelbetyg kan hänföras till den lägre tredjedelen enligt den betygsskala som används i det land där examen avlagts, det vill säga att examensbetyget måste motsvara det svenska betyget Godkänd.

Tillgodoräknande

Filosofiska fakultetsstyrelsen eller person som utsetts av styrelsen beslutar huruvida tidigare utbildning kan överföras till programmet.

Antagningsförfarande

Studerande antas till programmet i dess helhet.

Programplan

Termin 1 (HT 2016)

Kurskod	Kursnamn	Hp	Nivå	Veckor	VOF
732A34	Tidsserieanalys	6	A1X	v201634-201641	V
732A93	Statistical Methods	6	A1X	v201634-201641	V
732A94	Avancerad programmering i R	6	A1X	v201634-201641	V
732A42	Introduktion till akademiska studier på avancerad nivå	3	A1X	v201634-201702	O
732A37	Multivariata Statistiska Metoder	6	A1X	v201643-201651	V
732A95	Introduktion till maskininlärning	9	A1X	v201643-201652	O

Termin 2 (VT 2017)

Kurskod	Kursnamn	Hp	Nivå	Veckor	VOF
732A90	Datorintensiva statistiska metoder	6	A1X	v201704-201711	O
732A31	Data Mining - Clustering and Association Analysis	15	A1X	v201704-201723	O
720A04	Philosophy of Science	3	A1X	v201718-201723	O

Termin 3 (HT 2017)

Kurskod	Kursnamn	Hp	Nivå	Veckor	VOF
732A57	Databasteknik	6	A1X		V
732A62	Tidserieanalys	6	A1X	v201734-201741	V
732A63	Sannolikhetsteori	6	A1X	v201734-201741	V
732A66	Beslutsteori	6	A1X	v201734-201751	V
732A96	Avancerad maskininlärning	6	A1X	v201735-201741	V
732A65	Data Mining Project	6	A1N	v201735-201751	V
732A92	Text Mining	6	A1X	v201744-201751	V
732A97	Multivariata statistiska metoder	6	A1X	v201744-201751	V

Termin 4 (VT 2018)

Kurskod	Kursnamn	Hp	Nivå	Veckor	VOF
732A64	Masteruppsats i statistik	30	A1X	v201803-201823	O

Hp = Högskolepoäng
VOF = Valbar / Obligatorisk / Frivillig
*Kursen läses över flera terminer