

STAN45, Statistics: Data Mining and Visualization, 7,5 högskolepoäng

Statistics: Data Mining and Visualization, 7.5 credits

Avancerad nivå / Second Cycle

Fastställande

Kursplanen är fastställd av Institutionsstyrelsen vid Statistiska institutionen 2015-06-08 och senast reviderad 2015-06-08. Den reviderade kursplanen gäller från och med 2015-06-08, höstterminen 2015.

Allmänna uppgifter

Undervisningsspråk: Engelska

Huvudområde

Statistik

Fördjupning

A1N, Avancerad nivå, har endast kurs/er på grundnivå som förkunskapskrav

Kursens mål

Kunskap och förståelse

Ett godkänt betyg på kursen ges till de studenter som visar kunskap om

- analysmetoder för stora mängder data som är svåra att hantera utan moderna analysmetoder som involverar effektiva beräkningsverktyg (Big Data- problemet).

Färdighet och förmåga

Ett godkänt betyg på kursen ges till de studenter som visar förmåga att

- applicera en data mining-lösning på ett praktiskt problem som involverar komplexa, stora datamängder med fokus på ekonomi och affärsorienterade applikationer
- välja en algoritm så att den kan effektivt hantera mycket storskaliga datamängder men också bedöma algoritmens skalbarhet.

Värderingsförmåga och förhållningssätt

Ett godkänt betyg på kursen ges till de studenter som visar

- förtrogenhet med statistiska metoder som är användbara för att urskilja olika mönster från mångdimensionella data och som kan användas för att förbättra beslutsfattandet.

Kursens innehåll

Med den snabba utvecklingen inom informationstekniken har vi under det senaste decenniet bevittnat en explosiv tillväxt i vår förmåga att generera och samla in data. Inom affärsvärlden har man samlat mycket stora databaser med kommersiella och finansiella transaktioner. Inom vetenskapen har enorma mängder av vetenskapliga data genererats inom olika områden. Till exempel har kartläggningen av det mänskliga genomet genererat en databas om många gigabyte. Ett annat exempel är klimat- och miljödata som samlats med hjälp av satelliter. Internet är ytterligare ett exempel med miljarder sidor text- och multimedia-sekvenser som används dagligen av miljontals människor. Hur man skall analysera stora mängder data, så att de kan förstås och användas på ett effektivt sätt, är ett utmanande problem.

Data mining är en samling av flera metoder som tillhandahåller teknik och programvara för att automatisera analys och utforskande av stora komplexa datamängder. Utvecklande av data mining har bedrivits av forskare inom en mängd olika områden, bland annat statistik, maskininlärning, databashantering och datavisualisering. Detta är ett växande och snabbt växande område som kräver förståelse för både etablerad metod och nyligen framtagna tekniker.

Denna kurs presenterar metoder, de mest använda programmeringsverktygen och applikationer inom detta område. Genom att introducera huvudprinciper i statistiskt lärande bidrar kursen till att eleverna kan förstå grunderna i data mining. Betydande fokus kommer också att läggas på beräkningstekniker för använda algoritmer. Speciellt fokus kommer att läggas på att analysera vad som gör att en algoritm effektivt kan hantera mycket storskaliga datamängder men också studera algoritmens skalbarhet. Data mining och inlärningstekniker som har utvecklats i andra områden än statistik, t.ex. maskininlärning och signalbehandling, kommer också att studeras. Kursen tar också upp frågan om vad visualisering är, och varför man ska använda visualiseringar för kvantitativa data.

I kursen kommer studenterna att arbeta med egna projekt där olika standardprogram skall användas men där de också kommer att i viss utsträckning utveckla egna algoritmer. Diskussioner kring olika projekt kommer att ge studenterna möjlighet att dela och jämföra ideer med varandra och få särskild vägledning från instruktörerna. Insatser kommer att göras för att hjälpa eleverna att formulera verkliga problem i matematiska modeller, så att lämpliga algoritmer kan tillämpas med hänsyn till beräkningsbegränsningar.

Genom att kartlägga särskilda ämnen kommer eleverna att utsättas denna massiva litteratur och bli mer medvetna om den senaste forskningen. I synnerhet grunderna för klassificering och klustring, t.ex. linjära klassificeringsmetoder, prototypmetoder, beslutsträd, och dolda Markov-modeller, kommer att introduceras. Ungefär fem kursprojekt kommer att ingå med betoning på att förstå och använda inlärningsalgoritmer. Eleverna kommer att uppmuntras att lägga fram egna forskningsproblem där data mining modeller kan användas. Möjliga teman är bildsegmentering och bildhämtning; textsökning, länkanalys, och microarray dataanalys. I den laborativa delen av kursen kommer ett antal vanliga digitala visualiseringsmjukvaruverktyg att presenteras. Laborationerna kommer att fokusera

på att ge erfarenhet med hjälp av verkliga data.

Kursens genomförande

Kursen är utformad som en serie föreläsningar, självstudier och laborationer med rapporter. Undervisningen ges i tre former: föreläsning, projektdiskussion och en fördjupad undersökning.

Kursens examination

Betyg bygger på individuella prestationer, via skriftliga inlämningsuppgifter, muntlig presentation samt gruppaktiviteter.

Examinationen består dels av inlämningsuppgifter och dels av en datorbaserad tentamen.

Anmärkning

Universitetet betraktar plagiat mycket allvarligt, och kommer att vidta disciplinära åtgärder mot studenter för varje form av försök till fusk i undersökningar och bedömningar. Plagiering anses vara ett mycket allvarligt akademiskt brott. Straffet som kan utdömas för detta, och andra otillbörliga metoder i undersökningar eller bedömningar, inkluderar avstängning från universitetet.

Prov/moment för denna kurs finns i en bilaga i slutet av dokumentet.

Betyg

Betygsskalan omfattar betygsgraderna Underkänt, E, D, C, B, A.

A (Utmärkt) 85-100 poäng/procent. Ett framstående resultat som är utmärkt vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

B (Mycket bra) 75-84 poäng/procent. Ett mycket bra resultat som karakteriseras av mycket bra teoretiskt djup, praktisk relevans, analytisk förmåga samt självständighet.

C (Bra) 65-74 poäng/procent. Ett bra resultat som karakteriseras av bra teoretiskt djup, praktisk relevans, analytisk förmåga samt självständighet.

D (Tillfredsställande) 55-64 poäng/procent. Ett resultat som är tillfredsställande vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

E (Tillräckligt) 50-54 poäng/procent. Ett resultat som möter minimikraven vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet, men inte mer.

U (Otillräckligt/Underkänt) 0-49 poäng/procent. Ett resultat som är otillräckligt vad gäller teoretiskt djup, praktisk relevans, analytisk förmåga och självständighet.

För att få godkänt på en kurs måste studenten få betyg E eller högre.

Förkunskapskrav

Allmänna förkunskaper för masterprogrammet i statistik.

Prov/moment för kursen STAN45, Statistics: Data Mining and Visualization

Gäller från H15

- 1301 Data Mining and Visualization, exam, 7,5 hp
Betygsskala: Underkänt, E, D, C, B, A
- 1303 Data Mining and Visualization, lab, 0,0 hp
Betygsskala: Underkänd, Godkänd

Gäller från H13

- 1301 Data Mining and Visualization, exam, 7,5 hp
Betygsskala: Underkänt, E, D, C, B, A
- 1302 Data Mining and Visualization, lab, 0,0 hp
Betygsskala: Underkänt, E, D, C, B, A