ECUTBILDNING

Studieplanering

Kurs 5. Prediktiv analys

30 Yhp
2022-02-07 -2022-03-18

Kursens huvudsakliga innehåll

Kursen syftar till att ge den studerande specialiserade kunskaper i att förstå och använda olika statistiska metoder för att analysera historisk information för att på så sätt förutspå framtida händelser samt få färdigheter i att använda enkel prediktiva analys, regression och klassificering.

Kursen omfattar följande moment

- Regression
- Klassificering
- Prediktiva analys
- Machine learning metoder
- Visualisering av prediktiva analyser

Kursens mål/läranderesultat

Målet med kursen är att den studerande genom teori och praktiska övningar utvecklar specialiserade kunskaper för regression och klassificering.
Den studerande ska få utveckla sina färdigheter i att kunna tillämpa prediktiva analysalgoritmer och modeller samt skapa visualisering av prediktiva analyser
Den studerande har efter avslutad kurs kompetens för att självständigt kunna utföra prediktiv analys från dataset.

Efter genomförd kurs ska den studerande kunna:

Kunskaper:

1. Förklara regression och klassificering

Färdigheter:

2. Tillämpa prediktiva analysalgoritmer och modeller
3. Tillämpa Machine learning metoder
4. Skapa visualisering av prediktiva analyser

Kompetenser:

5. Utföra prediktiv analys från dataset

Former för undervisning

Kursen kommer att genomföras med traditionell undervisning i form av föreläsningar varvat med tid för praktisk träning på övningsuppgifter, med handledning av läraren. I kursen ingår också att genomföra övningsuppgifter på självstudietiden samt göra en individuell inlämningsuppgift.

Former för kunskapskontroll

Examination kommer att ske genom:
1 individuell inlämningsuppgift (IG/G/VG)

Betygsskala

Följande betygsskala tillämpas:
VG = Väl Godkänd, G = Godkänd, IG = Icke Godkänd

Läranderesultat	Inlämningsuppgift (IG/G/VG)
1	x
2	x
3	x
4	x
5	x

Principer för betygssättning

För betyget Godkänd ska den studerande

- Kunna på ett grundläggande sätt förklara regression och klassificering
- Kunna på ett grundläggande sätt tillämpa prediktiva analysalgoritmer och modeller
- Kunna på ett grundläggande sätt tillämpa Machine learning metoder
- Kunna på ett grundläggande sätt skapa visualisering av prediktiva analyser
- Kunna på ett grundläggande sätt utföra prediktiv analys från dataset

För betyget Väl Godkänd ska den studerande:

- Uppnått kraven för betyget Godkänd
- Kunna på ett fördjupat sätt tillämpa prediktiva analysalgoritmer och modeller
- Kunna på ett detaljerat sätt tillämpa Machine learning metoder
- Kunna på ett kreativt sätt skapa visualisering av prediktiva analyser
- Kunna på ett självständigt sätt utföra prediktiv analys från dataset

Icke Godkänd ges till studerande som har fullföljt kursen men inte nått alla mål för kursen.

Kunskapskontroll 1: Individuell inlämningsuppgift 1, prediktera från data, deadline 2022-03-18

Detta projekt syftar till att uppfylla följande kunskapskrav:
Kunskaper: 1
Färdigheter: 2, 3, 4
Kompetenser: 5

Bedömningskriterier:

- G: Ha uppnått samtliga lärandemål för uppgiften
- VG: Ha uppnått samtliga lärandemål för uppgiften samt självständigt reflekterat över, och motiverat, de valda teknikerna och lösningarna i inlämningsuppgiften

Övriga uppgifter i kursen

Övningsuppgifter presenteras på Teams/GitHub

Utbildare

Namn: Eva Hegnar
E-post: eva.hegnar@codic.se
Tfn: 073-805 91 60
Tillgänglighet: tisdag - torsdag 8.00-16.30, svarar **ibland** på direkta meddelanden måndag och fredag

Gästföreläsare

Kursmaterial

Typ av material	Kommentar
An Introduction to Statistical Learning	PDF Teams
The Elements of Statistical Learning	PDF Teams
Data Camp	Kurser på datacamp.com

Schema v.6

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-02-07	9.00-16.00	Egenstudier	
Tisdag 22-02-08 Föreläsning 1	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Intro till kursen och kursplanering Etablering av koncept: prediktiv analys, algoritm, statistik, machine learning, neuralt närverk, data mining, business intelligence, regression, klassificering, datarensning Installera Conda, VSC, Jupyter Notebook, Virtual Environment, GitHub Desktop och kursens repository	
Onsdag 22-02-09 Föreläsning 2	9.00-16.00	Föreläsning Sthlm Vi går igenom grundläggande koncept i prediktiv analys. Supervised VS unsupervised learning. Regression och klassificering. Modeller och algoritmer.	Vidare läsning och videor, pdf. An introduction to statistical learning (IntroStat) s.15-29
Torsdag 22-02-10 Föreläsning 2	9.00-16.00	Föreläsning Gbg Vi går igenom grundläggande koncept i prediktiv analys. Supervised VS unsupervised learning. Regression och klassificering. Modeller och algoritmer.	Vidare läsning och videor, pdf. An introduction to statistical learning (IntroStat) s.15-29
Fredag 22-02-11	9.00-16.00	Egenstudier	

Schema v.7

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-02-14	9.00-16.00	Egenstudier	
Tisdag 22-02-15 Föreläsning 3	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Viktiga python bibliotek för prediktiv analys. viktigaste stegen för prediktiv analys: 1. Förbereda data 2. Importera estimeringsobjektet 3. Skapa instans av modellen 4. Träna modellen 5. Utvärdera modellen 6. Prediktera	
Onsdag 22-02-16 Föreläsning 4	9.00-16.00	Föreläsning Sthlm Regression. The multiple regression model Ordinary least square Error metrics för regression: Mean squared error, Root mean squared error, Mean absolute error, R-squared, Explained variance. Predicting crime	IntroStat: Model accuracy (MSE) s.29-33 MLR s.71-75 The elements of statistical learning (ElemStat): MLR s.43-
Torsdag 22-02-17 Föreläsning 4	9.00-16.00	Föreläsning Gbg Regression. The multiple regression model Ordinary least square Error metrics för regression: Mean squared error, Root mean squared error, Mean absolute error, R-squared, Explained variance. Predicting crime	IntroStat: Model accuracy (MSE) s.29-33 MLR s.71-75 The elements of statistical learning (ElemStat): MLR s.43-
Fredag 22-02-18	9.00-16.00	Egenstudier	

Schema v.8

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-02-21	9.00-16.00	Egenstudier	
Tisdag 22-02-22 Föreläsning 5	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Regression i Python. KNN och Lasso regression. Model evaluation för regression: Cross validation, Overfitting, Bias-variance tradeoff, Overfitting, Regularization, Feature selection Predicting Diamond prices och Post popularity	IntroStat: Bias-var tradeoff s.33-36 KNN s.104-109 Cross-validation s175-186 Lasso s.219-227 ElemStat: Lasso s.68 ElemStat: Overfitting s.228-230 Regularization s.167-176
Onsdag 22-02-23 Föreläsning 6	9.00-16.00	Föreläsning Sthlm Klassificering. Olika typer av klassificering metoder, vad sannolikhet är och hur det används i klassificering. Klassificeringsmodellerna: Logistic Regression, Classification Trees och Naive Bayes.	IntroStat: Classification s.127- Logistic regression s.131-137 Classification tree s.311-314 ElemStat: Naive Bayes s.210-211
Torsdag 22-02-24 Föreläsning 6	9.00-16.00	Föreläsning Gbg Klassificering. Olika typer av klassificering metoder, vad sannolikhet är och hur det används i klassificering. Klassificeringsmodellerna: Logistic Regression, Classification Trees och Naive Bayes.	IntroStat: Classification s.127- Logistic regression s.131-137 Classification tree s.311-314 ElemStat: Naive Bayes s.210-211
Fredag 22-02-25	9.00-16.00	Egenstudier	

Schema v.9

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-02-28	9.00-16.00	Egenstudier	
Tisdag 22-03-01 Föreläsning 7	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Klassificering. Evaluering av klassifikationsmodeller med error metrics, confusion matrix och threshold value. När vi har kategoriska input features behövs de ofta konverteras till numeriska värden för att användas i modeller. Vi går igenom tre sätt att göra detta på: Ordinal Encoding, One-Hot Encoding och Dummy Variable. Predikterar Credit Card default.	IntroStat: Classification s.127- Dummy variable s.130-134 ElemStat: Classification s.101-
Onsdag 22-03-02 Föreläsning 8	9.00-16.00	Föreläsning Sthlm Klassificering. Om klasserna i output target är obalanserad (imbalanced data) kan vi behöva använda oss av Stratified Train-Test Split, Oversampling eller Undersampling. Vi predikterar Bankruptcy i Python.	IntroStat: Classification s.127- ElemStat: Classification s.101-
Torsdag 22-03-03 Föreläsning 8	9.00-16.00	Föreläsning Gbg Klassificering. Om klasserna i output target är obalanserad (imbalanced data) kan vi behöva använda oss av Stratified Train-Test Split, Oversampling eller Undersampling. Vi predikterar Bankruptcy i Python.	IntroStat: Classification s.127- ElemStat: Classification s.101-
Fredag 22-03-04	9.00-16.00	Egenstudier	

Schema v.10

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-03-07	9.00-16.00	Egenstudier	
Tisdag 22-03-08 Föreläsning 9	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Data pre-processing som är när man undersöka och fixar datan innan man modellerar. De teman som ingår är: Hur man hanterat NULL värden. Feature scaling med normalisering, standardisering och robust scaler. Outliers – upptäcka och hantera.	
Onsdag 22-03-09 Föreläsning 10	9.00-16.00	Föreläsning Sthlm Random Forest som är Ensemble methods. Det vill säga att det är metoder som kombinerar flera algoritmer för att bättre prediktiva prestanda. Feature selection. Vilka features man ska välja och hur ta reda på vilka som är viktiga.	IntroStat: Random forest s. 316-321 Feature selection s. 204 ElemStat: Random Forest s.587-
Torsdag 22-03-10 Föreläsning 10	9.00-16.00	Föreläsning Gbg Random Forest som är Ensemble methods. Det vill säga att det är metoder som kombinerar flera algoritmer för att bättre prediktiva prestanda. Feature selection. Vilka features man ska välja och hur ta reda på vilka som är viktiga.	IntroStat: Random forest s. 316-321 Feature selection s. 204 ElemStat: Random Forest s.587-
Fredag 22-03-11	9.00-16.00	Egenstudier	

Schema v.11

Datum	Tid	Lektionens innehåll	Att läsa till lektionen
Måndag 22-03-14	9.00-16.00	Egenstudier	
Tisdag 22-03-15 Föreläsning 11	9.00-16.00	Gemensam föreläsning Gbg/Sthlm Hyperparameter tuning. Hur välja bäst parameter till modellerna? Multicollinearity – när features är starkt beroende av varandra. Andra tema vi inte hunnit igenom, men är kul att kunna!	
Onsdag 22-03-16 Föreläsning 12	9.00-16.00	Föreläsning Sthlm Repetition och jobba med inlämning	
Torsdag 22-03-17 Föreläsning 12	9.00-16.00	Föreläsning Gbg Repetition och jobba med inlämning	
Fredag 22-03-18 Deadline Inlämning kl 23.55	9.00-16.00	Egenstudier	