Màster en Ciència de Dades

Continguts provisionals

Curs 2021/2022

Universitat de Girona
Escola Politècnica Superior

Presentació i agraïments

Aquest document descriu les diferents assignatures del Màster en Ciència de Dades. Els continguts són provisionals en el sentit que s'aniran configuran i ajustant a les necessitats que es vagin identificant per part de l'equip de treball format per professors del màster i professionals experts en el camp de la ciència de dades. En concret, es treballarà en la coordinació de les assignatures i els seus continguts per tal de completar-los, eliminar redundàncies i ajustar temps i càrregues de treball.

Aprofito per agraïr la comissió de definició del màster i les persones que han participat en el disseny de les assignatures: Esteve del Acebo, Anton Bardera, Miquel Bofill, Marc Comas, Ignacio Martín, Santiago Thió, Josep Soler, Marta Fort, Adrià Arnau, Òscar Galera, Juan González, David Juher, Beatriz López, Xavier Lladó, Glòria Mateu, Joan Saldaña, Xaquín Veira.

Mateu Villaret
Coordinador del Màster en Ciència de Dades

Índex

1. Introducció

El Màster en Ciència de Dades consta de 60 crèdits i estan distribuïts a parts iguals en dos quatrimestres.

Al primer quatrimestre s'imparteixen les assignatures:

- Estadística per a Ciència de Dades (EstCD - 6)
- Adquisició i preparació de dades (APC - 6)
- Machine Learning (ML - 9)
- Visualització de la informació (VI - 3)
- Desenvolupament, gestió i casos pràctics de projectes de ciència de dades (DGDCP-6)

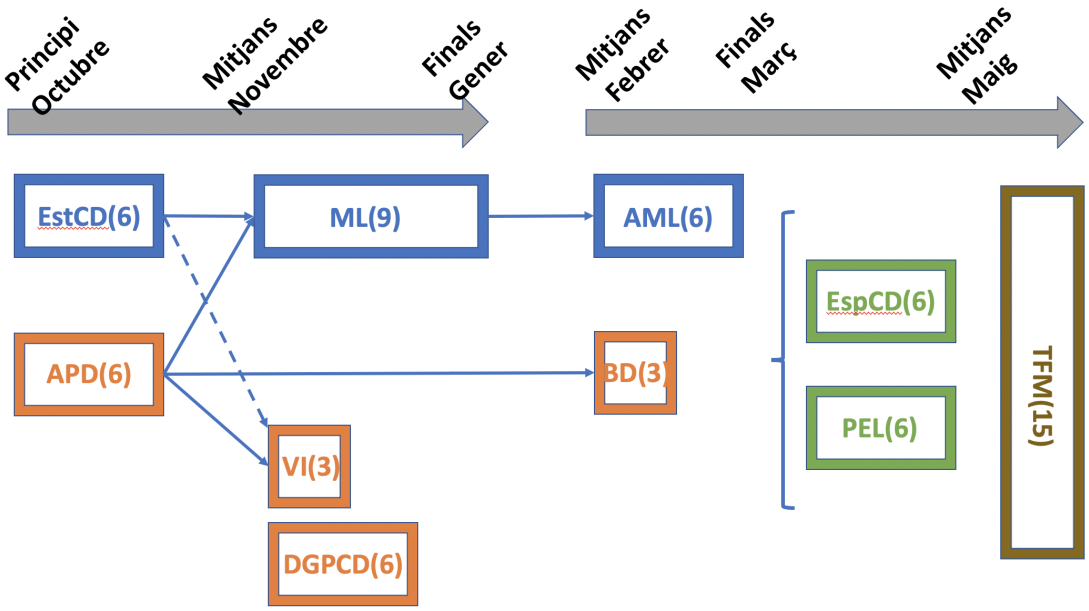
Al segon quatrimestre s'imparteixen les assignatures:

- Tècniques Avançades de Machine Learning (AML - 6)
- Big Data (BD - 3)
- Especialitzacions de ciència de dades (EspCD - 6)
- Pràctiques en entorn laboral (PEL - 6)

i també es realitzaria el Treball Final de Màster (TFM de 15 crèdits). Totes les assignatures són obligatòries llevat d'Especialitzacions de ciència de dades i Pràctiques en entorn laboral que són optatives i se n'ha de triar una de les dues.

També s'ofereix l'opció de fer el màster en dos anys, cursant 30 crèdits a cada any. El primer any s'haurien de cursar: Estadística per a Ciència de Dades, Adquisició i preparació de dades, Machine Learning, Tècniques Avançades de Machine Learning i Big Data. Al segon any s'haurien de cursar Desenvolupament, gestió i casos pràctics de projectes de ciència de dades, Visualització de la informació, Especialitzacions de ciència de dades o bé Pràctiques en entorn laboral i el Treball Final de Màster.

La impartició de les assignatures seguirà una estructura de prerrequisits, de manera que, per exemple, abans de fer ML, s'haurà cursat EstCD i APD, imprescindibles per tal de poder aprendre amb solvència els conceptes i els mètodes que es veuran a ML.



Les següents seccions corresponen a la descripció de contingut i funcionament de cada assignatura.

2. Estadística per a Ciència de Dades (EstCD)

L'Estadística és la pràctica de desenvolupar coneixement humà mitjançant l'ús de dades empíriques, i com a tal és una de les potes en la que se sustenta la Ciència de Dades.

En aquesta assignatura es repassarà els conceptes d'estadística descriptiva, probabilitat i inferència i s'aprofundirà en la modelització d'una variable resposta a partir d'un conjunt de variables explicatives que tant poden ser quantitatives com qualitatives. Es veuran els casos en que la variable resposta és numèrica, binària o un comptatge.

Continguts

Estadística descriptiva [1,5 / 6 crèdits]

Tipus de mostreig.

Descripció de dades multivariants.

- Tipus de variables.
- Matrius de dades.
- Mesures de centralitat i variabilitat.
- Mesures de dependència lineal.

Correlació i causalitat.

Anàlisi de components principals.

Anàlisi de correspondències.

Probabilitat [0,5 / 6 crèdits]

Conceptes de probabilitat.

- Què és una probabilitat? Interpretació freqüentista i bayesiana.
- Variables aleatòries.
- El principi d'inclusió-exclusió.
- Probabilitat conjunta.
- Probabilitat condicionada.
- La regla de Bayes.
- Independència i independència condicionada.

Principals distribucions contínues i discretes.

Distribució de probabilitat conjunta.

Transformació de variables aleatòries.

Inferència estadística [0,5 / 6 crèdits]

Estimació amb mètodes de remostreig.

Estimació amb resultats asimptòtics.

Estimació per màxima versemblança.

Models lineals generalitzats [3,5 / 6 crèdits]

Models de regressió lineal.

- Definició i notació dels models de regressió.
- Estimació i inferència dels models de regressió.
- Assumpcions dels models lineals generalitzats.
- Relaxació de l'assumpció de linealitat dels predictors continus: categorització, utilització d'splines.
- Remostreig, validació i simplificació del model.

Models lineals generalitzats.

- Especificació i estimació dels models lineals generalitzats.
- Comparació de models.

Modelització d'una resposta binària. Regressió logística.

- La funció logística.
- Interpretació dels models logístics.

Modelització de comptatges. Regressió de Poisson.

- Interpretació dels models de Poisson.

Pràctiques

El contingut pràctic de l'assignatura, que es farà principalment amb R, s'alternarà amb la teoria dins una mateixa sessió al llarg de tota l'assignatura. La raó entre teoria i pràctica serà d'1:1.

Prerrequisits

Assignatura de grau d'Estadística.

Assignatura de grau de Càlcul.

Avaluació

L'avaluació de l'assignatura serà contínua, amb forma d'activitats al llarg del curs.

Bibliografia

[1] Frank Harrell (2015). Regression Modeling Strategies. With Applications to Linear Models, Logistics and Ordinal Regression and Survival Analysis. Second edition. Springer Series in Statistics.

[2] Alan Agresti (2015). Foundations of Linear and Generalized Linear Models. Wiley Series in Probability and Statistics.

[3] Andrew Gelman, Jennifer Hill (2006). Data Analysis Using Regression and Multilevel-Hierechical Models. Cambridge University Press.

[4] Daniel Peña (2002). Análisis de Datos Multivariantes. McGraw-Hill.

3. Adquisició i Preparació de Dades (APD)

L'adquisió i preparació de les dades és un pas fonamental en tot projecte de Machine Learning, ja que dependrà d'aquestes la qualitat dels resultats generats pels models.

Inicialment es farà una introducció a Python on es donaran els coneixements suficients perque els alumnes siguin capaços d'executar les pràctiques posteriors. Es veuran els diferents tipus de fonts de dades així com les seves integracions, com extreure informació de les dades i com manipular-les.

Continguts

Introducció a l'Adquisió i Preparació de dades [0,5 / 6 crèdits]

- Fases d'un projecte de Machine Learning
- Perquè és tant important una bona preparació de les dades?

Introducció a Python per a Machine Learning [0,5 / 6 crèdits]

Teoria

- Introducció a Python
- Variables i tipus de dades
- Sintaxis
- Condicionals
- Funcions
- Seqüències
- Iteracions

Pràctiques

- Instal·lació de Python i exercicis bàsics

Introducció a Dataframes [1 / 6 crèdits]

Teoria

- Creació i parsing
- Select i filter
- Add
- Delete
- Rename
- Format
- Groupby
- Merge

- Iteració
- Exportació

Pràctiques

- Exercicis bàsics de Dataframes

Fonts i adquisió de dades [2 / 6 crèdits]

Teoria

- Fonts de dades i els seus tipus (API, BBDD, altres)
- Viabilitat de les dades en producció
- Integracions (API, Wrappers, drivers, scrapping)
- Preprocessament (errors de format, parsing, nulls)

Pràctiques

- Captura de dades a través de fitxer i API
- Captura de dades a través de BBDD relacional i no relacional

Anàlisi exploratòria de dades [2 / 6 crèdits]

Teoria

- Exploració de les dades
- Eines de visualització
- Manipulació de dades (constants, duplicades, dates, dades temporals)

Pràctiques

- Exploració i visualització de les dades
- Exploració, visualització i manipulació de dades

Pràctiques

El contingut pràctic de l'assignatura s'alternarà amb la teoria dins una mateixa sessió al llarg de tota l'assignatura. La raó entre teoria i pràctica serà d'1:1.

Prerrequisits

Coneixements bàsics de programació i algorísmica.

Avaluació

L'avaluació de l'assignatura tindrà dues parts: una avaluació continua amb forma d'activitats al llarg del curs, i una activitat final on s'avaluaran els coneixements adquirits en un problema real.

Bibliografia

[1] Alice Zheng, Amanda Casari *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*,O'Reilly 2018.

[2] Jacqueline Kazil, Katharine Jarmul *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*,O'Reilly 2016.

4. Machine Learning (ML)

L'aprenentatge automàtic o machine learning és la part de l'anàlisi de dades que es dedica a la creació d'algoritmes capaços de resoldre una certa tasca, i millorar els seus resultats en incrementar l'experiència.

En aquesta assignatura es definiran les principals tasques de l'aprenentatge automàtic, els principals mètodes per resoldre aquestes tasques i aprendrem a avaluar el rendiment dels resultats obtinguts amb aquests mètodes.

Continguts

Introducció i conceptes d'aprenentatge automàtic [0,5 / 9 crèdits]

- Què és l'aprenentatge automàtic?
- Tipus d'aprenentatge automàtic. Supervisat, no supervisat, semi-supervisat, per reforç.
- Metodologia de l'aprenentatge automàtic. De l'extracció de característiques per acomplir una tasca fins a l'avaluació d'un mètode per realitzar-la.

Aprenentatge supervisat [5,5 / 9 crèdits]

Introducció

- Com funciona l'aprenentatge supervisat?
- Regressió amb splines lineals: un cas simple d'aprenentatge supervisat.
- Tipus de tasques d'aprenentatge supervisat: classificació i regressió.
- Mesures de rendiment segons els tipus de tasques d'aprenentatge.
- Biaix i variància.
- Conjunts d'entrenament, conjunts de validació i el conjunt de test.
- Selecció d'hiperparàmetres. Estratègies de validació. Validació creuada, validació per bootstrapping.

Tasques

- Regressió. Predicció d'una variable numèrica.
- Classificació. Predicció d'una variable categòrica.

Mètodes

- Regressió lineal i regressió logística[1].
- *K* nearest-neighbours.
  - Definicions i objectius del mètode.
  - Mètrica i nombre de veïns.

[1] Ja vist a l'assignatura d'Estadística per DS.

- Arbres de decisió.
  - Definicions i objectius del mètode.
  - Mètodes/mètriques de separació.
  - Profunditat, mida màxima de les fulles, complexitat.
- Màquines de suport vectorial.
  - Definicions i objectius del mètode.
  - Aspectes geomètrics del mètode.
  - Aplicació del mètode per dades no separables.
  - Equilibri entre la mida de la frontera i les dades no separades.
  - Aplicació no lineal del mètode. Redefinició del producte escalar (kernel trick).
- Xarxes neuronals.
  - Definicions i objectius del mètode.
  - Conceptes: funció d'activació, nombre de capes, perceptró multicapa.
- Assemblatge d'algoritmes.
  - Definicions i objectius del mètode.
  - Conceptes: Weak learners.
  - Paradigmes d'assemblatge: bagging i boosting.
  - Exemples més importants: Random Forest i gradient boosting.

Miscel·lània

- Elecció del mètode d'aprenentatge.
  - Interpretabilitat dels mètodes.
  - Flexibilitat dels mètodes.
  - Temps i consum de memòria dels mètodes.
- Explicabilitat dels algoritmes.
  - Objectius del mètode.
  - Local surrogate models.
  - Shapley values.

Aprenentatge no supervisat [3 / 9 crèdits]

Introducció

- Objectiu de l'aprenentatge no supervisat.
- Mesures del rendiment d'un algoritme d'aprenentatge no supervisat.

Tasques

- Clustering (agrupació).
  - Creació d'una partició.

    - Creació d'una jerarquia.
  - Reducció de la dimensionalitat.
  - Modelització de la funció de densitat de probabilitat.
  - Detecció d'outliers.
  - Descobriment de patrons.

Mètodes

  - K-means.
  - DBSCAN.
  - Mixtures finites de distribucions.
    - Definició i paràmetres del mètode.
    - Suavització de la funció de densitat d'unes dades.
    - Agrupació de dades.
  - Anàlisi de components principals.
  - *t*-SNE i UMAP.
  - Mapes autoorganitzats.
  - Regles d'associació.
  - Sequence pattern mining.

Miscel·lània

  - Selecció del nombre grups.
    - Existeix més d'un grup? El test de Duda-Hart.
    - Mètodes visuals per la tria del nombre d'agrupacions.
    - Índex per la tria del nombre d'agrupacions: Calinski-Harabasz, silhouette width.
    - Mètodes basats en bootstrapping.

Pràctiques

El contingut pràctic de l'assignatura s'alternarà amb la teoria dins una mateixa sessió al llarg de tota l'assignatura. La raó entre teoria i pràctica serà d'1:1.

Prerrequisits

Estadística per a ciència de dades. Conceptes necessaris de l'assignatura: models lineals, components principals, tècniques de mostreig, avaluació dels models predictius. Programació amb R.

Adquisició i preparació de dades. Conceptes necessaris de l'assignatura: adquisició de dades, preprocessament de dades. Programació amb Python.

Avaluació

L'avaluació de l'assignatura tindrà dues parts: una avaluació continua amb forma d'activitats al llarg del curs, i una activitat final on s'avaluaran els coneixements adquirits en un problema real.

Bibliografia

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer Publishing Company, 2014.

[2] Andriy Burkov. *The hundred-page machine learning book.* 2019.

[3] Trevor Hastie, Robert Tibshirani, and Jerome H.Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* Springer, 2001.

[4] Kevin P. Murphy. *Machine Learning. A probabilistic Perspective.* MIT Press, 2012.

5. Visualització de la Informació (VI)

Per tal d'obtenir l'informació rellevant i entendre els patrons subjacents de grans col·leccions de dades, cal una complexa interacció entre l'anàlisi de dades, la mineria de dades i la visualització.

Aquesta assignatura pretén proporcionar als estudiants els conceptes teòrics i pràctics de visualització de l'informació, percepció aplicada a la visualització, mètodes de visualització avançats i de visualització interactiva, així com les eines bàsiques per al seu desenvolupament.

L'objectiu és preparar els estudiants per treballar en projectes complexos de ciència de dades que requereixen el desenvolupament d'interfícies visuals interactives per a l'anàlisi de dades.

Continguts

Introducció a la visualització de dades (0.5 cr.)

- Definicions de visualització de l'informació
- Per què visualitzar i per a què?
- Cicle de vida de la visualització

Estratègies per a la visualització (0.5 cr.)

- Abstraccions i models de dades i estratègies per mapejar les dades a gráfics
- Percepció i processament visual: atributs preatentius i la seva efectivitat, canals visuals

Tipus de gràfics (1 cr.)

- Capes de la visualització: visual, contextual, anotacions i interacció
- Univariate plots: enumeracions, distribucions, i sèries temporals
- Bivariate plots i multivariate plots
- Mapes, xarxes i grafs
- Facetes i múltiples petits

Disseny de l'interacció i lògica narrativa (1 cr.)

- Interacció i tasques visuals: selecció, agregació, filtres...
- Interfícies d'usuari, experiències i patrons d'interacció per a la visualització
- Notebooks, dashboards i vistes coordenadas
- Estratègies narratives per a la comunicació de dades

Eines per a la visualització: d3, ObservableHQ i Svelte (Una hora per classe, a totes les sessions)

Pràctiques

Mitjançant exercicis pràctics progressius, aprendrem Svelte en combinació amb d3 per a la visualització de dades. Svelte és un framework JavaScript basat en components, ràpid, modern i cada vegada més popular entre els desenvolupadors per a la producció d'eines visuals interactives.

Prerrequisits

- Assignatura: "Adquisició i preparació de dades"
- Assignatura "Estadística per a Ciència de Dades"
- Coneixements de tecnologies web (HTML, JavaScript, CSS, SVG, Canvas...)
- Coneixements bàsics de eines col·laboratives (GitHub)

Avaluació

L'avaluació es basarà en els exercicis setmanals, la participació als canals oberts de la classe i en un treball final. El pes de cada part serà:

- Exercicis setmanals: 50%
- Treball final: 20% (10% avaluació en grup, l'altre 10% per part del professor)
- Participació a classe i als canals de Slack: 30%

Bibliografia

[1] A. Wattenberger *Fullstack D3 and Data Visualization: Build Beautiful Data Visualizations with D3*. Fullstack.io, 2019.

[2] N.H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-Driven Storytelling*. AK Peters Visualization Series. CRC Press, 2018.

[3] S. Murray. *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. Number 3. O'Reilly Media, 2013.

[4] I. Meirelles. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. EBSCO ebook academic collection. Rockport Publishers, 2013.

[5] B. Fry. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media, 2007.

[6] C. Ware. *Information Visualization: Perception for Design*. Interactive Technologies. Elsevier Science, 2004.

6. Desenvolupament, Gestió i Casos Pràctics de Projectes de Ciència de Dades (DGPCD)

La implementació de projectes en ciència de dades involucra equips multifuncionals, focalitzats en àrees de negoci, governança i extracció de dades, modelatge i validació d'algoritmes, així com arquitectura i desplegament de software. Tenir una visió holística del procés, i entendre els principals reptes i beneficis de diferents metodologies és clau per a maximitzar les probabilitats d'èxit en el desenvolupament de solucions.

L'objectiu d'aquesta assignatura és conèixer les principals metodologies de desenvolupament de projectes de ciència de dades així com la implementació completa de projectes amb l'ajut de l'estudi de casos reals.

Continguts

Cicle de vida i viabilitat d'un projecte en ciència de dades [1/6 crèdits]

Introducció al desenvolupament de projectes

- El cicle de vida d'un projecte en ciència de dades.
- Rols al cicle: científic/enginyer/arquitecte de dades/ML etc.
- La importància de les metodologies híbrides: TDSP, CRISP-DM, Agile.

Relació amb el valor de negoci

- Coneixement de domini i equips multifuncionals.
- El valor primer: definicions, metodologies, PoC and MVPs.
- Traducció analítica de problemes de negoci.
- Comunicació per audiències no tècniques.

Dades, modelatge i validació [2/6 crèdits]

Privacitat en el tractament de dades

- El marc legal: intro al GDPR.
- Reptes tècnics: accessibilitat, (pseudo)anonimització, privacitat diferencial, aprenentatge federat.

Principis fonamentals del modelatge

- Metodologia científica: hipòtesis, reproductibilitat, replicabilitat i comunicació de resultats.
- El valor de la simplicitat.

Ètica en el tractament de dades

- El temut bias: impacte en la ciència de dades.
- Deures del científic de dades: interpretabilitat i transparència.

Esquemes de validació

- Validació offline: decàleg de bones pràctiques.
- Validació online: experimentació i field tests.

A producció [1/6 crèdits]

Desplegament de software

- Reptes associats a la posada en producció de models.
- Conceptes d'actualitat: SaaS, APIs, containers, MLOps etc
- Sumari d'eines i funcionalitats.

El cicle de vida d'un projecte en perspectiva

- Punts clau i oportunitats d'aprofundiment.
- Revisió en detall d'un cas real.

Casos pràctics d'ús de ciència de dades [2 / 6 crèdits]

En aquesta part de curs es presentaran diferents casos reals d'ús de la ciència de dades. Es preten que es pugui observar i assimilar la transversalitat d'aquesta ciència mitjançant una varietat de projectes reals en sectors de tots els àmbits.

Es demanarà i s'avaluarà als alumnes segons la comprensió dels projectes presentats demanant-los una implicació en les presentacions així com identificació de punts febles, punts forts i possibles extensions dels casos que es vegin.]

Pràctiques

Pels primers 4 crèdits de l'assignatura el contingut pràctic s'alternarà amb la part teòrica en una relació 1:1. Tota l'assignatura està enfocada en donar una visió pràctica de com és el desenvolupament de projectes en casos reals. La part pràctica guiada es centrarà en l'aprofundiment dels conceptes teòrics mitjançant la preparació, discussió i presentació de casos d'ús per part dels estudiants.

Prerrequisits

L'assignatura requereix de la part introductòria als conceptes bàsics impartits a les assignatures:

- Visualització de la Informació
- Estadística per la Ciència de Dades
- Machine Learning
- Adquisició i preparació de Dades

Avaluació

L'avaluació de l'assignatura serà contínua, es basarà en la participació i la discussió dels estudiants durant les parts pràctiques de l'assignatura (inclosos els casos pràctics d'ús de ciència de dades). Es demanarà una recol·lecció de les aportacions fetes per cada estudiant al final del curs per a completar l'avaluació.

Bibliografia

[1] Robert de Graaf. *Managing Your Data Science Projects: Learn Salesmanship, Presentation, and Maintenance of Completed Models.* Apress, ISBN: 9781484249079, (2019).

7. Tècniques Avançades de Machine Learning (AML)

L'objectiu d'aquesta assignatura és aprofundir en el coneixement de l'aprenentatge automàtic estudiant diferents tècniques i aplicacions avançades de ciència de dades.

Les tècniques de Deep Learning basades en xarxes neuronals han creat un gran impacte en nombroses aplicacions. En aquesta assignatura es descriuran els elements principals d'aquestes tècniques, s'analitzaran les arquitectures més rellevants i s'avaluaran els seus resultats en diferents aplicacions.

Continguts

Introducció al Deep Learning [0,4 / 6 crèdits]

- Història del Deep Learning
- Aplicacions del Deep Learning
- Conceptes bàsics i terminologia
- Introduint datasets: MNIST i IMAGENET

Deep Learning [1,3 / 6 crèdits]

- Terminologia per Xarxes Neuronals
- La regla del perceptró
- Representant components de les xarxes amb vectors i matrius
- Concepte de Backpropagation
- Feedforward Neural Network i entrenament
- Regularització
- Learning Rate, Momentum i Dropout
- Stochastic Gradient Descent

Convolutional Neural Networks [1,3 / 6 crèdits]

- Introducció al processat d'imatges
- Definició de filtres i convolució
- Terminologia de les xarxes convolucionals
- Convolutional i fully connected layers
- Mapes de característiques, funcions d'activació, pooling i normalització
- Arquitectures Deep i Shallow
- Arquitectures per classificació i per segmentació d'imatges
- Data augmentation

- Aplicacions de les xarxes convolucionals

Transfer Learning i aprenentatge per reforç [1 / 6 crèdits]

- Concepte de transfer learning
- Fine-tuning d'una xarxa neuronal
- Exemples pràctics de transfer learning
- Aprenentatge per reforç

Sistemes recomanadors [1 / 6 crèdits]

- Introducció. Recomanadors no personalitzats
- Recomanadors basats en el filtrat de continguts
- Recomanadors basats en filtratge col.laboratiu ítem-íten i usuari-usuari
- Avaluació online i offline de sistemes recomanadors. Mètriques

Mineria de text [1 / 6 crèdits]

- Introducció. Obtenció de dades. web crawling i web scraping
- Processament del lenguatge natural. Natural Language Toolkit NLTK
- Mètodes d'aprenetatge automàtic per classificació de textes.
- Models generatius, Latent Dirichlet Allocation.

Mètodes

- Xarxes convolucionals
  - Arquitectures 2D i 3D
  - Arquitectures multicanal
  - Arquitectures Unet o patch based
- Tècniques de transfer learning (fine tuning)

Tasques

- Programació de xarxes neuronals convolucionals
- Classificació d'imatges
- Aplicació de tècniques de transfer learning
- Implementació de sistemes recomanadors
- Anàlisi d'opinions i de sentiments. Identificació de temes (subject modelling).

Pràctiques

El contingut pràctic de l'assignatura s'alternarà amb la teoria dins una mateixa sessió al llarg de tota l'assignatura. La proporció entre teoria i pràctica serà de 1 a 1.

Prerrequisits

Conceptes necessaris de l'assignatura: fonaments de Machine Learning, adquisició de dades, preprocessament de dades. Programació amb Python.

Avaluació

L'avaluació de l'assignatura tindrà dues parts: una avaluació continua amb forma d'activitats i pràctiques al llarg del curs, i una prova final on s'avaluaran els coneixements adquirits en tot el curs.

Bibliografia

[1] J. Howard, S. Gugger. *Deep Learning for Coders with Fastai and PyTorch*. 2020.

[2] S. Skansi. *Introduction to deep learning. From logical calculus to artificial intelligence*. Springer 2018.

[3] I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*. MIT Press. 2016.

[4] Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer 2016.

[5] R. Banik. *Hands-On Recommendation Systems with Python*. Packt Publishing 2018.

[6] A. Kedia. *Hands-On Python Natural Language Processing*. Packt Publishing 2020.

8. Big Data (BD)

Les dades massives o més conegut com el fenomen *Big Data*, és un moviment relativament nou que ha anat a l'alça en els últims anys, gràcies a l'avanç tecnològic que comporta l'aparició de grans quantitats de dispositius capaços de generar fluxos continus de dades, i a la voluntat d'analitzar aquestes dades per tal d'extreure coneixement.

Com es veurà en aquesta assignatura, aquest augment en la quantitat, velocitat de processament i tipologia de les dades, suposa nous reptes que caldrà afrontar mitjançant l'ús de noves tècniques i tecnologies principalment distribuïdes.

Continguts

Introducció al *Big Data* [0,5 / 3 crèdits]

- Què és?
- Quines dimensions té?
- Quins reptes nous suposa?
- Algorismes en paral·lel

Sistemes de fitxers distribuïts [0,5 / 3 crèdits]

- Què és *HDFS*?
- Arquitectura
  - *NameNode* i *DataNode*
  - Espai de noms (*namespace*)
- Fiabilitat
  - *HDFS heartbeats*
  - Reajustament de blocs de dades
  - Permisos d'usuaris, fitxers i directoris

Bases de dades NoSQL [1 / 3 crèdits]

- Què vol dir *NoSQL*?
  - Diferència amb les bases de dades relacionals
- Característiques de les bases de dades *NoSQL*
- Què és la persistència políglota?
- Modelització de les dades
  - Agregació per clau-valor
  - Agregació per columna
  - Agregació per document

  - Model en graf
- Què és una base de dades distribuïda?
  - Tipus
    * Client / Servidor
    * Sistema en 3 capes
    * Sistema *peer-to-peer*
  - Consistència [2]
    * Què és una transacció?
      · Transaccions *ACID*
      · Transaccions *BASE*
    * Tipus d'inconsistències
      · Actualització perduda
      · Lectura no confirmada
      · Lectura no repetible
      · Anàlisi inconsistent (fantasma)

Tecnologies i eines *Big Data* [1 / 3 crèdits]

- *Apache Hadoop*
- *Apache Spark*
- *Apache Flink*
- *Apache Kafka*

Pràctiques

El contingut pràctic de l'assignatura es durà a terme després de cada una de les unitats didàctiques i es desenvoluparà parcialment a l'aula i a casa. La raó entre teoria i pràctica serà aproximadament de 1 : 1.

Prerequisits

Per cursar aquesta assignatura cal haver superat la matèria **Adquisició i preparació de dades**, així com tenir coneixements de programació amb el llenguatge *Python* o *Scala* i conceptes bàsics sobre bases de dades relacionals i sistemes distribuïts.

Avaluació

L'avaluació serà contínua i comptarà amb quatre pràctiques on es provaran els conceptes adquirits a cada un dels blocs didàctics.

[2] D'acord a l'evolució de l'assignatura i al temps disponible, es parlarà del teorema *CAP*

Bibliografia

[1] Alejandro Corbellini, Cristian Mateos, Alejandro Zunino, Daniela Godoy and Silvia Schiaffino. *Persisting big-data: The NoSQL landscape* ISISTAN Research Institute, 2017.

[2] Pramod J. Sadalage and Martin Fowler. *NoSQL Distilled* Addison-Wesley, 2013.

[3] Neha Narkhede, Gwen Shapira and Todd Palino. *Kafka The Definitive Guide* OReally, 2017.

[4] Bill Chambers and Matei Zaharia. *Spark The Definitive Guide* OReally, 2018.

[5] Tom White. *Hadoop The Definitive Guide* OReally, 2015.

9. Especialitzacions de Ciència de Dades (EspCD)

La recerca en l'àmbit de la ciència de dades permet expandir els límits d'aquesta ciència cada cop més lluny.

En aquesta assignatura es presentaran diferents temes de recerca relacionats amb la ciència de dades que són objecte d'estudi de diferents grups de recerca a l'EPS i que ofereixen la possibilitat de fer TFM i/o Tesis doctorals amb els respectius grups.

Continguts

Restriccions i Optimització per a ciència de dades [1 / 6 crèdits]

Es presentarà com utilitzar eines de resolució de problemes combinatoris durs per a la ciència de dades. Grup de recerca de Lògica i Intel·ligència Artificial (GRCT0038).

Modelització amb Restriccions

Anàlisi Lògic de Dades

[1] Chikalov I. et al. (2013) Logical Analysis of Data: Theory, Methodology and Applications. In: Three Approaches to Data Analysis. Intelligent Systems Reference Library, vol 41. Springer, Berlin, Heidelberg.

[2] Chambon, A., Boureau, T., Lardeux, F. et al. Logical characterization of groups of data: a comparative study. Appl Intell 48, 2284–2303 (2018).

Introducció a l'anàlisi de xarxes complexes. [1 / 6 crèdits]

Es posarà un especial èmfasi en el problema de la detecció de comunitats en aquestes xarxes. Grup de recerca d'Equacions diferencials, modelització i aplicacions (GRCT0068).

Descripció estadística de les xarxes complexes

- distribució i correlació de graus, clustering.

Algorismes de generació de xarxes complexes.

Algorismes per a la detecció de comunitats en xarxes complexes.

[1] A-L, Barabási: Network Science, Cambridge University Press, 2016.

[2] S.N. Dorogovtsev: Lecture notes on complex networks, Oxford University Press, 2010.

Dades Espacials. [1 / 6 crèdits]

Es treballarà amb dades espacials que identifiquen la posició o forma geomètrica d'objectes georeferenciables i possiblement també les relacions entre ells. Es veuran alguns dels algorismes i àmbits d'aplicació bàsics de la geometria computacional.

Àrea de recerca de Processat de Dades espacials i espai-temporals del grup de recerca Graphics and Imaging Laboratory (GRCT0081).

- Subdivisions planars, triangulacions i estructures de dades espacials.
- Problemes de proximitat.
- Problemes de visibilitat.

[1] M.de Berg, O. Cheong, M van Kreveld, M. Overmars, Computational Geometry: Algorithms and Applications. Ed. Springer-Verlag, 3rd ed. 2008. ISBN: 978-3-540-77973-5

[2] J.-R. Sack, J. Urrutia, Handbook of Computational Geometry. Ed. Elsevier 2000. ISBIN 978-0-444-82537-7

Anàlisi estadística de dades composicionals. [1 / 6 crèdits]

Es treballarà la naturalesa de les dades composicionals, l'estructura particular del seu espai mostral i principis bàsics i les eines específiques per la seva anàlisi estadística. S'utilitzaran exemples de motivació i s'analitzaran conjunts de dades reals. Grup de recerca en estadística i anàlisi de dades composicionals (GR-EADC).

- Dades composicionals, espai mostral. Exemples. Dificultats en l'anàlisi estadístic.
- Principis de l'anàlisi estadístic de dades composicionals
- La solució: anàlisi estadística a través dels log-quocients i balanços.
- Preprocessament i eines composicionals per anàlisis descriptives i inferencials.
- Exemple d'aplicació utilitzant el software específic CoDaPack: microbioma (dades òmiques).

[1] Aitchison ,J. (1986) The statistical analysis of compositional data, Monographs on Statistics and Applied Probability. Chapman and Hall Ltd., London (UK) ISBN: 978-1930665781

[2] Filzmoser P, Hron K, Templ M. (2018) Applied compositional data analysis: with worked examples in R. Springer Series in Statistics book series, New York, NY. ISBN: 978-3319964201.

[3] Pawlowsky-Glahn V., Egozcue JJ., Tolosana-Delgado R. (2015) Modeling and Analysis of Compositional Data. Wiley. ISBN: 978-1-118-44306-4

[4] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017 Nov 15;8:2224. doi: 10.3389/fmicb.2017.02224.

Els següents dos temes són objecte de recerca del Grup de recerca d'Enginyeria de control i sistemes intel·ligents (GRCT 41).

Processament del senyal [1 / 6 crèdits]

- Filtrat i tractament de soroll
- Aplicacions de tècniques de ML per a la classificació de senyals
- Aplicacions de tècniques de ML per a la predicció de valors
- Mètodes basats en l'espai freqüencial.
- Aplicacions en senyals fisiològics (EEG, ECG, etc.)

[1] Abdulhamit Subasi, Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques A MATLAB® Based Approach Elsevier, 2019
[2] Palit, Ajoy K., Popovic, Dobrivoje. Computational Intelligence in Time Series Forecasting. Theory and Engineering Applications. AIC 2005

Aprenentatge automàtic en entorns distribuïts[1/6 ECTS]

- Formulació de l'aprenentatge en un entorn distribuït
- Aprenentatge federat
- Aplicació a mètodes de classificació – Adaboost
- Aplicació a mètodes de clustering
- Aplicacions in IIoT

Bibliografia

[1] Amita Kapoor. Hands-On Artificial Intelligence for IoT: Expert machine learning and deep learning techniques for developing smarter IoT Systems. Packt Publishing (January 31, 2019)
[2] Gerhard Weiss. Multiagent Systems: A modern Approach to Distributed Artificial Intelligence. MIT Press, 1999. Chapter 6. Learning in Multiagent Systems.

Pràctiques

El contingut pràctic de l'assignatura s'alternarà amb la teoria dins una mateixa sessió al llarg de tota l'assignatura. La raó entre teoria i pràctica serà d'1:1.

Prerrequisits

Estadística per a Ciència de Dades i Machine Learning.

Avaluació

L'avaluació de l'assignatura serà contínua, amb forma d'activitats al llarg del curs.

10. Pràctiques en Entorn Laboral (PEL)

Acció formativa desenvolupada per un estudiant en qualsevol empresa col·laboradora, pública o privada, nacional o estrangera, o en unitats de la pròpia universitat, amb l'objectiu d'aplicar i complementar la formació adquirida en la seva formació acadèmica, apropar a l'estudiant a la realitat de l'àmbit professional en el qual exercirà la seva activitat professional i desenvolupar competències que afavoreixin la seva incorporació al mercat de treball.

11. Treball Final de Màster

El Treball Final de Màster permetrà posar de manifest la maduresa i nivell cientificotècnic aconseguits durant el procés formatiu. Es presentarà una memòria per escrit i l'alumne també haurà de defensar el treball davant d'un tribunal format per professors del màster.