DIARIO PROGETTO DI STAGE ZUCCHETTI S.P.A

1 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/06/08	- Installazione software Orange Canvas; - Inizio studio documentazione/tutorial di Orange Canvas (https://orange.biolab.si/docs/); - Inizio attività di analisi dei dati operando su alcuni dataset semplici (es. visualizzazione dei dati in modo che siano significativi, suddivisione dei dati in sottogruppi - clustering ecc.) - Inizio studio tecniche di previsione con confronto (albero decisionale e regressione logaritmica sembrano dare previsioni molto simili).	Nessuna.
2020/06/09	- Continuazione studio documentazione/tutorial Orange Canvas; - Studio di alcune nozioni necessarie per l'analisi e la valutazione dei diversi algoritmi di classificazione (curva roc, cross validation, confusion matrix ec.); - Continuazione studio metodi di classificazione: osservazione e valutazione tramite Orange Canvas dei vari risultati di apprendimento che si possono ottenere dai vari algoritmi (regressione logistica, regressione lineare, Tree, SVM ...) - Test & Score; - Studio tecniche di visualizzazione per grandi moli di dati e individuazione delle classi/funzionalità più significative: tecnica PCA, analisi dei dati in base al grado che ciascuna funzionalità presenta, relativa distribuzione e box-plot (sempre meglio trovare una classe che abbia valori ben separati per ciascuna tipologia che la compone); - Inizio studio k-Means, sempre tramite Orange Canvas.	Nessuna.
2020/06/10	- Continuazione studio documentazione/tutorial Orange Canvas; - Continuazione studio k-Means: analisi di come appaiono i dati, individuazione numero cluster migliori, posizionamento dei centroidi ecc.; - Studio tecnica di addestramento: impostazione di differenti percentuali per la parte di train e di test, impiego di differenti tipologie di algoritmi e osservazione critica degli esiti ottenuti (precisione, confusion matrix, curva ROC...). Il medesimo procedimento e' stato poi fatto impiegando direttamente il widget Test & Score, confrontando sempre differenti tipologie di	Nessuna.

	algoritmi fra loro (SVM, Random Forest, Regressione Logistica...) e valutandone i risultati ottenuti al fine di comprenderne la bontà; - Analisi con l'impiego sempre del sw Orange Canvas, di quelli che possono essere i pesi di ciascun predittore e di cosa accade modificando la variabile target e le varie altre esplicative coinvolte. A tal fine sono stati impiegati differenti tipologie di algoritmi (sempre distinguendo fra i dati di test e di train), in modo da poter fare delle valutazioni prima "ragionate" attraverso l'osservazione dei risultati ottenuti (confusion matrix, curva ROC, tabelle coi risultati...) e poi in modo più puntuale attraverso il widget "Rank". In tal modo si sono così potute dare delle basi verificate alle conclusioni fatte in precedenza; - Inizio studio, anche se non ancora terminato dell'utilità e dei modi di impiego della Distance Map nell'analisi dei dati.	
2020/06/11	- Continuato studio documentazione/tutorial Orange Canvas; - Continuato studio dei concetti di PCA e Silhouette con successiva loro applicazione; - Continuato studio algoritmo non supervisionato k-Means e perfezionati concetti di cluster, medioide, centroide, Hierarchical Clustering e Distance Map; - Analizzato dataset Wine (facente parti dei vari database che Orange Canvas dispone). Nelle giornate precedenti infatti si era data precedenza ad Iris, considerandolo abbastanza capiente (150 entry) ma non troppo complesso (solo 4 var.indipendenti: predittori), per poter approcciarsi all'attività di analisi dei dati (gli altri erano stati comunque un po' guardati ma non in modo molto approfondito). Su questo nuovo set di dati sono dunque state applicate le tecniche di PCA e k-Means; i risultati sono stati poi analizzati in modo da poter fare delle considerazioni sull'efficienza dell'algoritmo (il widget Distributions e' stato molto utile per capire se i punti nei vari cluster erano stati considerati giusti o meno).	Nessuna.
2020/06/12	- Continuazione studio documentazione/tutorial Orange Canvas; - Continuazione studio dell'algoritmo non supervisionato k-Means, PCA, Hierarchical Clustering, Distance Map e e Silhouette con relative applicazioni; - Continuazione analisi del dataset Wine, e relativa redazione della documentazione contenente il resoconto di quanto fatto e dei risultati ottenuti.	Nessuna.

2 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/06/15	- Studio delle origini della regressione lineare; - Studio principio dei minimi quadrati; - Studio concetti di R^2, devianza totale, devianza spiegata e devianza residua (d.tot = d.spiegata + d.residua e R^2 = d.spiegata/d.totale = 1 - d.residua/d.totale); - Iniziata a guardare (leggermente) libreria relativa; - Iniziata documentazione di quanto appreso sulla regressione lineare.	Nessuna.
2020/06/16	- Iniziato studio libreria regressione lineare, aiutandosi anche con JSMLT (http://visualml.io/#/) e https://github.com/Tom-Alexander/regression-js.	Devo ancora capire esattamente come funziona la cosa dal punto di vista pratico.
2020/06/17	- Continuato studio teorico della regressione lineare con integrazione della documentazione relativa; - Continuato studio librerie regressione lineare.	Sebbene la situazione sia un po' più chiara di ieri non si e' ancora riusciti a impostare il lavoro.
2020/06/18	- Continuato studio libreria sulla regressione lineare: sono stati riprodotti in un file .html alcuni dati di prova per comprenderne il comportamento predittivo e il modo di calcolo di intercetta e coefficiente angolare; ci si e' aiutati riproducendo la cosa anche tramite Orange Canvas.	Come i giorni scorsi ci sono state alcune difficoltà nella comprensione dei metodi impiegati dalla libreria. Ora in teoria risolte.
2020/06/19	- Iniziato studio linguaggio R: esaminazione e documentazione dei risultati ottenuti dall'analisi del dataset Iris (da terminare perché ci stanno un sacco di cose da provare volendo); - Ripresi alcuni concetti come omoschedasticita', eteroschedasticita', p-value, statistica-test, RSE, t-value ecc. cc.., così da avere una comprensione dettagliata dei risultati ottenuti, e poter fare delle valutazioni in merito.	Nessuna.

3 SETTIMANA

Data	Descrizione	Eventuali criticità

2020/06/22	- Continuato studio R; - Studiati i concetti di ho e h1 (ipotesi nulla e ipotesi alternativa), e capito il nesso fra test t, t value, statistica F e statistica T...; - Compreso come rifiutare o accettare l'ipotesi nulla in base ai risultati mostrati da R; - Viste tecniche di selezione dei modelli (forward selection, backward selection e stepwise regression); - Varie: applicata trasformata logaritmica così da osservarne i risultati. - Documentato tutto man mano.	- Devo ancora comprendere bene come funziona anova(): mi permetterebbe di confrontare fra loro più modelli; - Devo capire come far venire fuori il grafico della regressione in modo sensato quando ho un modello composto da più predittori.
2020/06/23	- Continuato studio R; - Confrontati vari modelli di regressione lineare selezionandone poi il migliore (impiegando anche anova()..); - Inclusi nei modelli di regressione anche i valori discreti, osservandone l'incidenza anche a livello di significatività dei vari predittori. Si sono inoltre costruiti dei grafici che mostrassero la distribuzione per classe di appartenenza; - Dove possibile tutti i risultati sono stati confrontati con quelli dati da Orange Canvas in modo da avere certezza di quanto osservato; - Iniziato studio algoritmo Box-Cox; - Tutto documentano in modo da tenere traccia di osservazioni -> teorie.	Nessuna.
2020/06/24	- Continuato studio R; - Compreso algoritmo Box-Cox e sperimentato in un dataset idoneo: trasformazione modello dataset Cars; - Appreso calcolo indice AIC e confutati gli esiti anche in base a quanto fatto nei giorni precedenti su Iris; - Documentato tutto; - Iniziato studio SVM con R.	Nessuna.
2020/06/25	- Continuato studio R; - Continuato studio SVM; - implementato programmino che dati 20 valori random riesce a individuare il piano che taglia meglio lo spazio. Inoltre partendo dalle osservazioni iniziali riesce a predire quali sono i vettori di supporto; - Ripassati alcuni concetti statistici sulle variabili fattore.	Nessuna.
2020/06/26	- Continuato studio R; - Ottimizzato il programmino di ieri; - Iniziata analisi dataset Iris con SVM; - Ripresi alcuni concetti statistici (es.	Devo capire bene come fare in modo che mi vengano visivamente segnati i vettori di supporto nelle

	variabile fattore...)	quantità/posizioni che mi aspetto.

4 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/06/29	- Continuato studio SVM con R, e analisi dataset Iris; - Imparato a dividere il set di dati per fare prima l'addestramento e poi la predizione; - Imparato ad osservare il plot che mostra i risultati di SVM; - Implementate diverse tipologie di svm (lineare, radiale, polinomiale, e sigmoide), sperimentando anche cosa succede quando aumento o diminuisco il costo: confrontati i risultati ottenuti attraverso la confusion matrix e il calcolo della bontà della previsione; - Confrontati differenti tipologie di modelli per SVM lineare (uno senza vincoli sugli attributi e uno con vincoli), selezionandone poi il migliore; - I risultati ottenuti sono stati riportati sotto forma di commento nel programma R (in qualche frangente lo riporto in forma più formale... appena finisco di studiare la teoria del SVM); - Iniziato studio di alcuni concetti di Statistica Learning (statistica parametrica e non, maledizione della dimensionalità, teorema del limite centrale…).	Devo capire come inquadrare in modo sensato, le varie teorie che sono state sviluppate nel tempo, in modo che mi aiutino nella comprensione del mio lavoro.
2020/06/30	- Continuato studio SVM: sia con R che concetti teorici (VC dimension, margine, +1/-1…); - Studio metodo k-fold, attraverso anche R in modo da avere consapevolezza dei risultati ottenuti. Confrontato col metodo k-fold con ripetizione; - Iniziata annotazione di alcuni concetti chiave di SVM (brainstorming).	Nessuna.
2020/07/01	- Terminato studio teorico svm; - Iniziato a documentare bene il tutto.	Nessuna.
2020/07/02	- Proseguita documentazione SVM; - Comprensione / tentativi con libreria RL.	Devo ancora capire come implementare R^2.

2020/07/03	- Implementati R^2 aggiustato, RSE, statistica F, falsi positivi e veri positivi nella libreria di regressione; - Studio libreria di regressione; - Implementato CSS, in modo che tutto fosse omogeneo e carino.	Nessuna (in teoria).

5 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/07/06	- Continuato studio libreria RL; - Implementati metodi di varianza e covarianza; - Ricontrollati i metodi fatti in precedenza.	Non sono completamente sicura che i metodi siano corretti.
2020/07/07	- Continuato a lavorare sulla libreria RL; - Iniziato studio e lavoro libreria classificazione.	Devo ancora capire benissimo come incastrare bene le cose nel SVM, ma ci sto lavorando.
2020/07/08	- Continuazione studio e lavoro libreria algoritmi di classificazione: in particolare la regressione logistica; - Implementati: confusion-matrix, calcolo della probabilità di un evento e metodo odds.	In teoria ho implementato un metodo per valutare la bontà del modello di classificazione: (err. no modello + err. col modello) / err. no modello; ma non sono convintissima.
2020/07/09	- Continuato lavoro su libreria classificazione; - Implementati funzionalità recall, precision, accuratezza...	Devo ancora capire il modo in cui viene fatto il train perché allo stato attuale le labels sono sempre -1 per ogni dato; di conseguenza la previsione e' sempre -1...quindi se ci sono, ci sono sempre e solo tutti veri negativi. A me questa cosa sembra strana.
2020/07/10	- Implementati un altro po' di metodi per la classificazione (error rate, f-score ecc); - Continuato studio libreria classificazione.	Non sono sicurissima dei risultati, infatti ho fatto due versioni: una fa i conti sui dati di train, l'alta usa invece dei dati di test. Non sono convintissima in nessuno dei due casi.

6 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/07/13	- Continuato studio libreria SVM; - Sistemate cose varie; - Implementata sia per SVM che per RL funzionalità clean desk che pulisce il piano di lavoro e permette di riiniziare da capo; - Implementati tutti i metodi per la predizione del modello (su dati di train).	In teoria ho risolto tutti i problemi; non sono sicurissima su tutto ma ci dovrei essere (spero).
2020/07/14	- Continuato studio libreria di classificazione; - Sistemate cose varie nelle librerie di RL e classificazione; - implementate un po' di funzionalità di contorno.	Nessuna (in teoria).
2020/07/15	- Sistemate un po' di cose nelle interfacce per RL ed SVM (arrotondamento decimali, slider per kernel rbf….); - iniziato studio libreria random forest e impostato lavoro.	Nessuna.
2020/07/16	- Proseguito lavoro sulle interfacce (rl, svm, rf).	Ho trovato un'errore nel knn che devo ancora capire bene come risolvere.
2020/07/17	- Proseguito studio e lavoro nelle librerie rf, rl e svm; - Studiato metodo del kernel trick; - Implementate nuove funzionalità (probabilità di classificazione errata, rimozione di un singolo punto dal grafico …) e sistemate alcune imprecisioni nelle interfacce; - Reso un po' più fluido il css.	Nessuna (in teoria).

7 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/07/20 +1	- Iniziato studio/lavoro su libreria clustering; - iniziata creazione interfaccia.	Devo capire se quello che ho fatto oggi e' giusto.
2020/07/21	- Continuazione lavoro/studio interfaccia/libreria clustering;	Devo ancora capire bene come si fa il train.

+1	- Studio di metodi di valutazione del modello di clustering da includere nell'interfaccia; - Riordino documentazione stage relativa alla parte di analisi dei dati precedentemente generata.	
2020/07/22 +1	- Andata un po' avanti con la documentazione; - Proseguito studio libreria k-means e lavoro relativo.	Devo ancora risolvere il problema di ieri, ci sto lavorando.
2020/07/23 +1	- Andata "leggermente" avanti con la documentazione; - Proseguito estensione e aggiunta di funzionalità alle interfaccie; - Sistemati problemi vari con clustering e altro; - Introdotti oltre al k-means anche Dbscan e Optics; - Cominciata l'introduzione nel clustering, dei criteri di valutazione del modello (non ancora terminato).	Nessuna.
2020/07/24 +1	- Proseguita documentazione; - Fatta area di influenza nel clustering + altre cose sulle interfacce.	Nessuna.

8 SETTIMANA

Data	Descrizione	Eventuali criticità
2020/07/27 +1	- Continuata documentazione; - Continuata estensione interfaccia clustering + sistemate altre cose; - Continuato studio modelli di clustering/modi di valutazione dei modelli.	Nessuna.

2020/07/28 +1	- Continuata documentazione; - Corretto errore in interfaccia clustering.	Nessuna.
2020/07/29 +1	- Terminata Documentazione; - Ricontrollate interfacce/ lavoro fatto e corretti errori sparsi.	Ho provato a fare in modo che gli slider per kernel rbf che regolano la C ed il Sigma cambino effettivamente il grafico, ma la cosa mi deve ancora riuscire.
2020/07/30	- Conclusi lavori vari e fatti controlli sulle interfacce..	Nessuna.