Software Visuali per Analisi Avanzate

Corso sostitutivo di Tirocinio

UNIVERSITÀ DEGLI STUDI DEL SANNIO
DEMM

Calendario del corso

Giorno	Durata	Argomenti
Venerdì 27/3 16.00 – 18.00	2 ore	Presentazione del Corso Il contesto dei Big Data Analitica Avanzata Visual Analytics
Sabato 28/3 9.30 - 13.30	4 ore	Visual Analytic, Il Quadrante Gartner per la Visual Analytics Tableau
Martedì 31/3 16.00 – 18.00	2 ore	Tableau
Sabato 4/4 9.30 - 13.30	4 ore	Tableau Percezione e rappresentazione funzionale delle informazioni
Martedì 7/4 14.00 – 18.00	4 ore	Il Quadrante di Gartner per l'Advanced Analytics Ripasso concetti di Machine Learning Sas: Introduzione, ambiente operativo, Visual Analytic
Martedì 14/4	4 ore	SaS – Ospite SaS Italia
Sabato 18/4	4 ore	SaS Conclusione del Modulo 1

UNIVERSITÀ DEGLI STUDI DEL SANNIO
DEMM

Sommario

3

DSML – Data Science & Machine Learning

- Una piattaforma DSML è un prodotto principale ed un portafoglio coerente ed integrato di prodotti, componenti, librerie e framework (proprietarie, di partner e open source) a supporto di una pipeline analitica.
- Gli utenti principali sono pofessionisti delle data science:
  - expert data scientists
  - citizen data scientists (*)
  - data engineers (**)
  - machine learning (ML) engineers/specialists.

[Magic Quadrant for Data Science and Machine Learning Platforms]

(*) https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist/

(**) https://cognitiveclass.ai/blog/data-scientist-vs-data-engineer

DSML – Data Science & Machine Learning

- Le piattaforme DSML offrono un mix di funzioni di base ed avanzate per la costruzione di soluzioni DSML (principalmente modelli predittivi e prescrittivi).
- Le piattaforme supportano anche l'incorporazione di queste soluzioni in processi di business, infrastrutture, prodotti ed applicazioni.
- Esse devono supportare i task:
  - Data ingestion
  - Data preparation
  - Data exploration
  - Feature engineering
  - Model creation and training
  - Model testing
  - Deployment
  - Monitoring
  - Maintenance
  - Collaboration

DSML – Data Science & Machine Learning

- Non tutte le organizzazioni creano i modelli da zero od in modo completamente autonomo. In alcuni casi è necessario ricorrere a qualche tipo di assistenza
- E' quindi importante, ed il Magic Quadrant DSML lo prende in esame, la disponibilità di template ed esempi già pronti nelle piattaforme, perché semplificano l'avvio dei progetti.
- Questo aspetto è molto importante: oggi avere progetti di analitica avanzata permette di avere vantaggio competitivo!

DSML – Data Science & Machine Learning

- E' molto importante che le piattaforme supportino non solo la costruzione di un modello, ma anche la sua messa in esercizio. Infatti i benefici di un progetto DSML non potranno essere ottenuti senza che essi:
  - vengano "embeddati" nei processi di business e negli ambienti decisionali
  - Vengano mantenuti, monitorati e gestiti nel tempo
- Nonostante vi siano stati numerosi avanzamenti tecnologici nell'ambito dell'AI e del ML, una percentuale molto alta di progetti non sono resi operativi.
- Uno dei motivi cruciali è la scarsità di tool che facilitano l'operazionalizzazione dei modelli.

Gartner Magic Quadrant

- **Magic Quadrant (MQ)** è una serie di ricerche di mercato pubblicate dalla società di consulenza **Gartner** basati su metodi di analisi proprietari dei dati per mostrare le tendenze del mercato.
- Le analisi sono condotte per diversi settori tecnologici specifici e vengono periodicamente aggiornate.
- Gartner valuta i fornitori in base a due criteri: **completeness of vision** (completezza della visione) e **ability to execute** (capacità di esecuzione)
- Il report di Garnter si divide in 4 quadranti

Gartner Magic Quadrant

- **Leaders** – In questo quadrante sono presenti i vendor che hanno puntaggio più alto per Completeness of Vision e Ability to Execute. Hanno quote di mercato, credibilità e le capacità di marketing e di vendita necessarie a guidare la tecnolgia al successo. Questi vendor dimostrano chiara comprensione delle necessità del mercato, hanno un pensiero innovativo e piani ben definiti.
- **Challengers** – I vendor di questo quadrante sono presenti nel mercato ed hanno una Ability to Execute buona, tanto da costituire una seria minaccia per i venditori nel quadrante Leader. Hanno prodotti validi, posizione di mercato e risorse sufficientemente credibili per sostenere la crescita continua. Hanno buoan redditività finanziaria, ma non hanno le dimensioni e l'influenza dei venditori nel quadrante Leader.

Gartner Magic Quadrant

- **Visionaries** – Un fornitore nel quadrante Visionaries offre prodotti innovativi che affrontano bene i problemi degli utenti finali a livello operativo o finanziario, ma non ha ancora dimostrato la capacità di acquisire quote di mercato o redditività sostenibile. I venditori visionari sono spesso aziende private e obiettivi di acquisizione per aziende più grandi e affermate. La probabilità di essere acquisiti spesso riduce i rischi associati all'adozione dei loro sistemi.
- **Niche Players** – I vendors di questo quadrante sono spesso focalizzati su segmenti di mercato o esigenze verticali specifiche. Questo quadrante può anche includere venditori che stanno riadattando loro prodotti esistenti per entrare nel mercato, o venditori più grandi che hanno difficoltà a far affermare la propria visione.

Magic Quadrant for Data Science and Machine Learning Platforms

Expert data scientists and other professionals working in data science roles require capabilities to source data, build models and operationalize machine learning insights. **Significant vendor growth, product development and myriad competing visions reflect a healthy market that is maturing rapidly**.

https://www.gartner.com/doc/reprints?id=1-1YCR6NY7&ct=200213&st=sb

11

Osservazioni

- Quadrante dei Leader più «affollato» rispetto a quello dell'Analytics and Business Intelligence
- Molte delle aziende presenti in questo Magic Qaudrant erano presenti anche in quello dell'Analytics and Business Intelligence
- Alcune aziende di particolare rilevanza:
  - Alteryx
  - Databricks
  - Tibco
  - H2O.ai
  - Microsoft
  - Google

UNIVERSITÀ DEGLI STUDI DEL SANNIO DEMM

12

Magic Quadrant for Analytics and Business Intelligence Platforms

Augmented capabilities are becoming key differentiators for analytics and BI platforms, at a time when cloud ecosystems are also influencing selection decisions. This Magic Quadrant will help data and analytics leaders evolve their analytics and BI technology portfolios in light of these changes.

https://www.gartner.com/doc/reprints?id=1-1YAE9AY1&ct=200206&st=sb&signin=640614711143974a616dfab8ab01ef7d

13

SAS

- SAS offers a variety of software products for analytics and data science supporting statistics, ML, text analytics, forecasting, time series analysis, econometrics and optimization.
- SAS Visual Data Mining and Machine Learning (VDMML) was the core product evaluated for this Magic Quadrant. It incorporates multiple products including Visual Analytics and Visual Statistics.
- SAS provides a dedicated resource center for analytics that offers a series of webinars, events, fact sheets, webcasts, white papers, and more.
- The SAS Resource Center acts as both a library and a support community for users looking for feature information, tips and tricks and case studies in analytics.
- SAS is again positioned as a Leader this year. Its DSML products have a high degree of enterprise readiness and consistently deliver high business value to customers.
- SAS's Ability to Execute continues to be impacted by high license costs, which cause existing and prospective customers to explore other options. It recently launched life cycle product bundles called Unified Insights to reduce licensing complexity.

SaS Viya

- SAS Viya è un'architettura unificata con un environment analitico centralizzato per consentire la gestione end-to-end del dato, dall'esplorazione al risultato di business
- SAS è Leader nel Magic Quadrant 2020 di Gartner per le piattaforme di Data Science e Machine Learning.
- SAS Visual Data Mining and Machine Learning fornisce un'interfaccia intuitiva per velocizzare la costruzione di modelli e la generazione di codici.
- Ha funzioni di ricerca dati, feature engineering, riduzione della dimensione, analisi esplorativa, modellizzazione, training, tuning e implementazione dei modelli nei processi produttivi.

SaS Viya

- Durante il corso utilizzeremo la piattaforma per esplorare le funzionalità di advanced analytics, con particolare riferimento al machine learning supervisionato
- Piattaforma e-learning di SaS per il corso " Machine Learning Using SAS® Viya"
- Badge pubblicabile su Linkedin dopo 14 h di lezione e-learning
- Un contributo direttamente da SaS Italia
- Possibilità di certificarsi SaS Certified Specialist: Machine Learning Using SAS® Viya

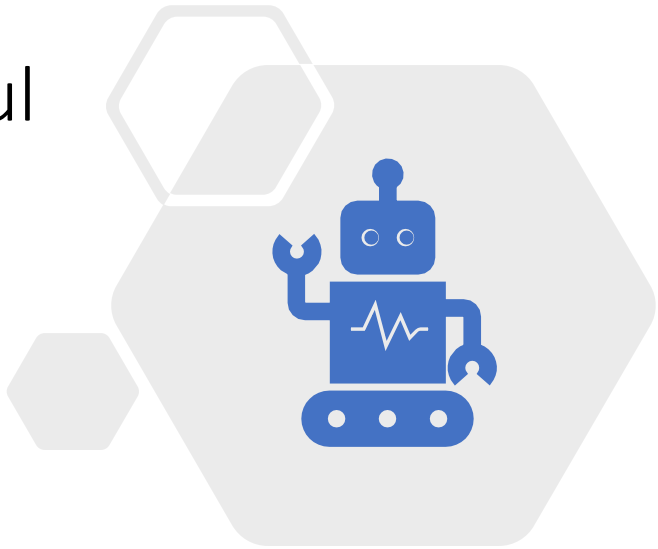
Lezione SaS del 17 Aprile

- Il giorno 17 Aprile avremo ospite un docente di SaS
- E' importante avere l'ambiente SaS installato
- Consigliato seguire la lezione 1 dell'e-learning SaS

- A tutti i partecipanti della lezione del 17 Aprile SaS rilascerà un certificato di partecipazione, previo rilascio del conenso.

- Comunicare il consenso a comunicare a SaS nome, cognome ed e-mail nella chat

Panoramica sul Machine Learning



Panoramica sul Machine Learning

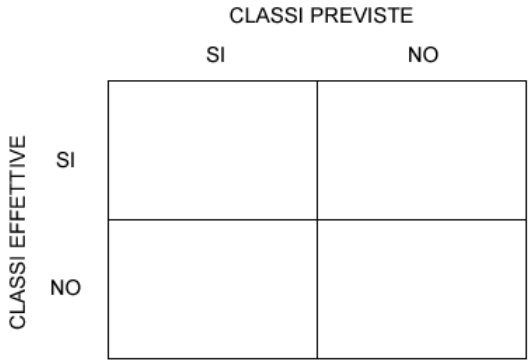
- https://www.slideshare.net/LucaNaso/panoramica-sul-machine-learning
- http://www.andreaminini.com/ai/machine-learning/matrice-di-confusione
- https://en.wikipedia.org/wiki/Confusion_matrix

19

20

Matrice di confusione

- La matrice di confusione (confusion matrix) è uno strumento per analizzare gli errori compiuti da un modello di machine learning.
- Prendo il semplice caso di un classificatore binario. Le classi sono due: SI o NO
- Elenco le classi del problema nelle righe e nelle colonne. Nelle righe indico le classi effettive(classi delle risposte corrette). Nelle colonne indico le classi di previsione (le classi predette dal modello).

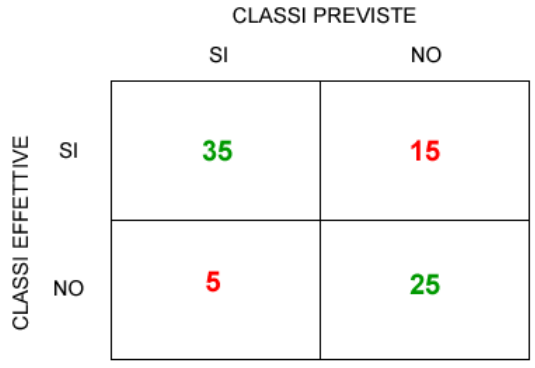


WWW.ANDREAMININI.COM

21

Matrice di confusione

- Ad esempio, il modello analizza 80 email e le classifica spam/no spam.
- In 60 casi il modello classifica correttamente mentre in 20 sbaglia.



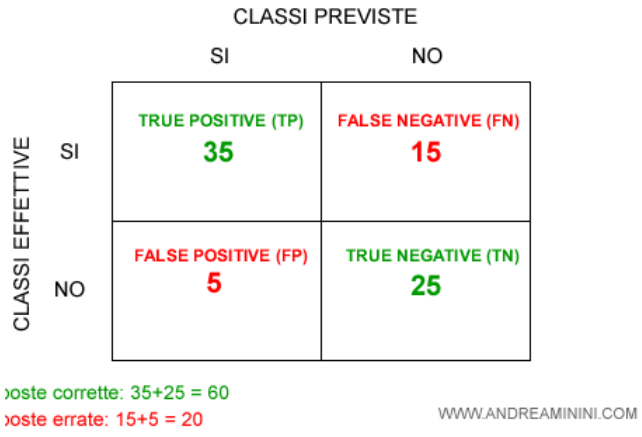
oste corrette: 35+25 = 60
oste errate: 15+5 = 20

WWW.ANDREAMININI.COM

22

Matrice di confusione

- **True positive (TP)** Se la classe prevista è SI ed è uguale alla classe effettiva. Il modello ha predetto correttamente SI.
- **True negative (TN)** Se la classe prevista è NO ed è uguale alla classe effettiva. Il modello ha risposto correttamente NO.
- **False positive (FP)** Se la classe prevista è SI ma è diversa dalla classe effettiva. Il modello ha sbagliato a rispondere SI.
- **False negative (FN)** Se la classe prevista è NO ma è diversa dalla classe effettiva. Il modello ha sbagliato a rispondere NO.



23

Tasso di errore

- Il tasso di errore (error rate) misura la percentuale di errore delle previsione sul totale delle istanze.
- Varia da 0 (peggiore) a 1 (migliore).

$$ERR = \frac{FP + FN}{TP + TN + FP + FN}$$

24

Accuratezza

- L'accuratezza (accuracy) misura la percentuale delle previsioni esatte sul totale delle istanze. E' l'inverso del tasso di errore.
- Varia da 0 (peggiore) a 1 (migliore).

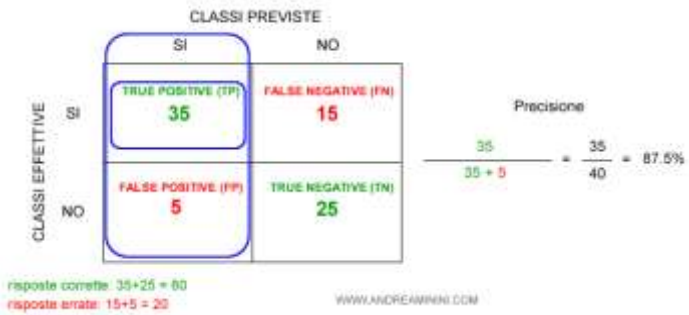
$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = 1 - ERR$$

25

Precisione

- La precisione (precision) è la percentuale delle previsioni positive corrette (TP) sul totale delle previsioni positive del modello (giuste TP o sbagliate FP).

$$PR = \frac{TP}{TP + FP}$$



26

Richiamo o sensitività

- Il richiamo (o recall) o sensitività (sensitivity) è la percentuale delle previsioni positive corrette (TP) sul totale delle istanze positive.
- Varia da 0 (peggiore) a 1 (migliore).

$$Recall = \frac{TP}{TP + FN}$$

27

Specificità

- La specificità (specificity) è la percentuale delle previsioni negative corrette (TN) sul totale delle istanze negative.
- Varia da 0 (peggiore) a 1 (migliore).

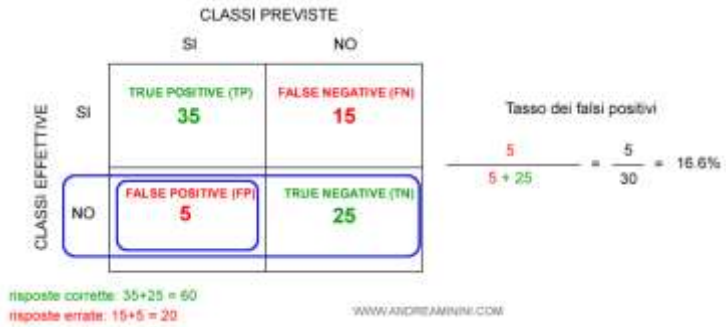
$$SP = \frac{TN}{TN + FP}$$

28

Tasso dei falsi positivi

- Il tasso dei falsi positivi (False Positive Rate) è la percentuale delle previsioni positive errate (FP) sul totale delle istanze negative.
- Varia da 0 (migliore) a 1 (peggiore).

$$FPR = \frac{FP}{TN + FP}$$



29

F Score

- Il punteggio F (F-Score) o F1 score è la media armonica delle metriche Precision e Recall.
- Varia da 0 (peggiore) a 1 (migliore).

$$FS = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

- Molto utile in particolare nel caso di dataset molto sbilanciati (es. altissimo numero di casi negativi e pochi positivi)

$$FS = \frac{2 \cdot 0.70 \cdot 0.875}{0.70 + 0.875} = \frac{1.225}{1.575} = 0.77$$

30

Esempio di utilizzo di F Score

- Se in un dataset di tamponi ho 99.900 casi negativi e 100 casi positivi
- Se l'algoritmo predice sempre «negativo»

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = 1 - ERR$$

- ACC = 0 + 99.900 / 0 + 99.900 + 0 + 100 = 0,999 => **99,9% accuratezza**

CLASSI EFFETTIVE \ CLASSI PREVISTE	SI	NO
SI	0	0
NO	100	99.900

WWW.ANDREAMININI.COM

- F-score = 2*0 / 0 + 100 + 99.900 = **0 F1 score!**

31

Hands on!

32

Link utili alla documentazione

- Home del Virtual Leraning Environment per SCYP Machine Learning with SAS: https://vle.sas.com/course/view.php?id=3436
- Cliccare su: LEARN SAS per accedere all'e-learning (link diretto: https://vle.sas.com/course/view.php?id=3376)
- Cliccare su Access SaS Software per accedere a Sas Viya (link diretto: https://vle.sas.com/course/view.php?id=3436&section=1)

La piattaforma SaS

- Machine Learning con SaS
- Il problema della Churn Prediction
- Accesso alla piattaforma SaS Viya
- Creazione di un nuovo progetto
- La pipeline di Machine Learning
- Visual Analytics con SaS

Primi passi su SaS Viya

- https://vle.sas.com/mod/scorm/view.php?id=78169
  - Create a Project
  - Modify the Data Partition
  - Build a pipeline from a Basic Template
- Visual Analytics con Sas Viya

35

Primi passi su SaS Viya

- https://v4e001.vfe.sas.com/SASDrive/
- Andare su menù delle applicazioni
- Esplirazione e visualizzazione dei dati
- Cliccare su Dati
- Selezionare COMMSDATA
- Creare grafico a Barre
  - Aggiungere una categoria (osservare cardinalità)
  - Osservare misura di default
- Creare grafico a torta
  - Aggiungere una categoria (osservare cardinalità)
  - Aggiungere una misura
- Dashboard
- Esporta Dati
- Heatmap
- Mappa Ad albero
- Cluster

36

Grazie per l'attenzione


https://it.linkedin.com/in/valerio-morfino


vmorfino@unisannio.it

37