Orange (autore: Vittorio Albertoni)

Premessa

Nel 1996, presso l'Università di Lubiana, viene lanciato un progetto per un frame di machine learning, denominato appunto ML, da realizzarsi in linguaggio C++.

L'anno dopo al linguaggio C++ si associa il linguaggio Python.

Nel 2003 viene ridisegnata l'interfaccia grafica utilizzando la libreria PyQt, il binding Python del framework Qt.

Nel frattempo il progetto viene ribattezzato con il nome Orange e nel 2009 vengono superati i 100 strumenti di elaborazione messi a disposizione, chiamati widgets.

Nel 2015 viene rilasciata la versione 3, Orange3, che nel momento in cui scrivo (dicembre 2020) è la 3.27.

Orange è una libreria Python che può essere importata per programmare script di elaborazione dati in linguaggio python. La libreria è concepita per semplificare lo scripting e potenziare alcuni comandi ma si appoggia alle solite librerie del mondo Python per il data mining (numpy, scipy, matplotlib, pandas, sklearn, ecc.) nulla aggiungendo alla capacità elaborativa di queste.

I suoi comandi semplificati o potenziati sono stati organizzati in modo da poter essere richiamati visualmente attraverso icone in un canvas dove, opportunamente collegati tra loro attraverso il mouse, vadano a svolgere elaborazioni anche complesse, senza che all'utente sia richiesto di scrivere una sola riga di codice.

Gli oggetti richiamabili vengono chiamati widgets e gli widgets presenti e tra loro collegati nel canvas costituiscono un workflow.

In questo manualetto, ad uso di principianti dilettanti, non guarderemo a Orange come libreria richiamabile per scrivere programmi ma ci dedicheremo a capire i procedimenti attraverso i quali sia possibile compiere elaborazioni, anche non banali, utilizzando il canvas visuale senza scrivere codice.

Indice

[1] Questo esercizio è stato affrontato nel Capitolo 3 dell'allegato in formato PDF «knime» al mio articolo «KNIME: l'alternativa a Python per i data scientists». In quella sede ho mostrato come si possa risolvere il problema con un foglio di calcolo, con uno script Python e con il software KNIME. KNIME, scritto in linguaggio Java, è un'ottima alternativa a Orange per fare le cose che stiamo vedendo senza scrivere codice.

[2] Questo esercizio è stato affrontato nel Capitolo 5 dell'allegato in formato PDF «knime» al mio articolo «KNIME: l'alternativa a Python per i data scientists». In quella sede ho mostrato come si possa risolvere il problema con uno script Python e con il software KNIME. KNIME, scritto in linguaggio Java, è un'ottima alternativa a Orange per fare le cose che stiamo vedendo senza scrivere codice.

1 Installazione

Troviamo Orange all'indirizzo *https://orange.biolab.si/*.

Su questo sito è presente, in lingua inglese, tutta la documentazione esistente su Orange, anche in forma di simpatici video dimostrativi, sempre in lingua inglese, ma facile e sorretta da didascalie. Visitando le varie pagine possiamo leggere e vedere molte cose.

Per quanto riguarda l'installazione, il miglior modo di installare Orange è farlo da Anaconda o Miniconda[3]con i comandi

```
conda config --add channels conda-forge
conda install orange3.
```

Si può installare il software anche dal navigatore di Anaconda cliccando sull'icona

Altro modo che può funzionare è quello di ricorrere a `pip`, a patto che non vi sia troppo divario temporale tra l'epoca di installazione di Python e dei moduli per il data mining e quella di installazione di Orange[4].

Dalla pagina DOWNLOAD del sito internet di Orange possiamo comunque scaricare gli installer per Windows e macOS.

Per Windows è disponibile anche una versione portable in formato zip. Basta estrarla su disco o su pennetta USB ed è pronta all'uso.

Per Linux è disponibile il tarball del source su GitHub.

Con la prima installazione abbiamo disponibili praticamente tutti gli strumenti per il data mining numerico, salvo alcune finezze, come, per esempio, la regressione polinomiale, e non vengono installati strumenti per l'analisi del testo (text mining).

Componenti aggiuntive possono essere installate una volta avviato Orange nei modi che vedremo nel prossimo capitolo.

2 Funzionamento

Una volta installato, Orange:

. può essere importato in uno script Pyhton con l'istruzione `import Orange` per utilizzarne le funzioni nello script stesso, ma questo non interessa in questa sede;

. può esserne avviata l'interfaccia grafica con il comando a terminale

```
orange-canvas
```

oppure

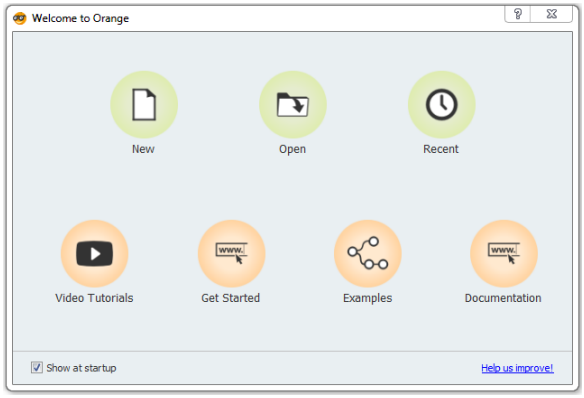
`python -m Orange.canvas` (o `python3 -m Orange.canvas` se sul sistema è installato anche Python 2)

Se utilizziamo la versione portable su Windows possiamo lanciare l'interfaccia grafica con il lanciatore `Orange` che si trova nella directory che ospita i file del programma.

La prima finestra che si presenta è una finestra di benvenuto

[3]Chi non conosca Anaconda può consultare l'allegato in formato PDF «python_anaconda» al mio articolo «Software libero per data scientists» dell'aprile 2019 sul mio blog *www.vittal.it*.

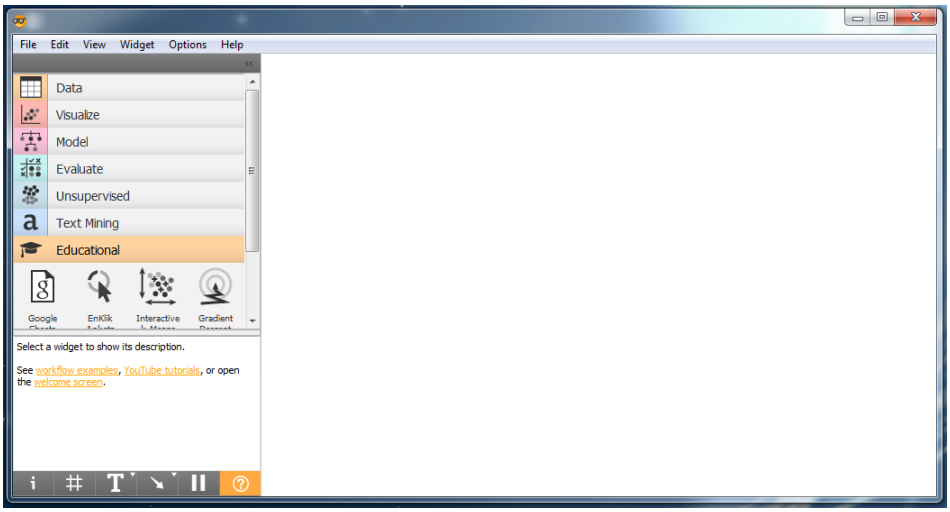
[4]Per il funzionamento del repository di Python rimando all'allegato in formato PDF «mondo_python» al mio articolo «Python per tutti» del febbraio 2017 sul mio blog *www.vittal.it*.



Questa finestra, che possiamo evitare si ripresenti deselezionando l'opzione SHOW AT STARTUP in fondo a sinistra, è utile per il neofita in quanto, con le icone in basso, dà accesso a tutta una serie di documentazioni di utile consultazione.

Le tre icone in alto aprono il canvas per lavorare su un nuovo progetto, per aprire un progetto già salvato o per aprire un progetto operato di recente.

Il canvas per un nuovo progetto si presenta così



Sulla sinistra abbiamo la finestra dove si trovano i raggruppamenti dei widget disponibili.

I primi cinque raggruppamenti (Data, Visualize, Model, Evaluate e Unsupervised) sono quelli che si rendono disponibili alla prima installazione.

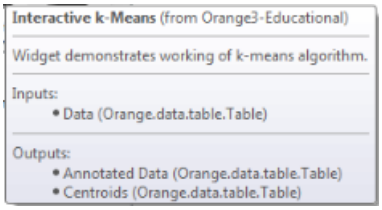
Gli altri due (Text mining e Educational) ritengo opportuno siano installati in aggiunta.

Per farlo, con collegamento internet attivo, si sceglie da menu OPTIONS ▷ ADD-ONS... e, nella finestra successiva, si selezionano le voci ORANGE3-EDUCATIONAL e ORANGE3-TEXT e si preme OK in fondo alla finestra.

Cliccando sulla voce che indica il raggruppamento si apre l'elenco dei widget che contiene. Nella figura è aperto l'elenco dei widget del raggruppamento EDUCATIONAL e se ne vedono i primi quattro; i successivi si possono vedere agendo sulla barra di scorrimento sulla destra della finestrella.

Fermando il puntatore del mouse sull'icona di un widget apriamo una finestrella che illustra il widget stesso.

Se, per esempio, fermiamo il puntatore del mouse sull'icona del widget INTERACTIVE K-MEANS, il terzo che vediamo tra quelli del raggruppamento EDUCATION nella figura, apriamo questa finestrella



nella quale ci viene detto che il widget dimostra il modo di lavorare dell'algoritmo K-Means, uno dei più utilizzati per il clustering, accetta in input dati da una tabella e fornisce due output leggibili su altrettante tabelle.

Per avere spiegazioni complete sul funzionamento dei widget è disponibile, con collegamento internet attivo, il Widget catalog. Lo possiamo aprire cliccando sull'icona DOCUMENTATION nella pagina di benvenuto prima illustrata oppure aprendo la pagina DOCS sul sito di Orange, scegliendo poi VISUAL PROGRAMMING ▷ WIDGET CATALOG. In questo catalogo sono illustrati tutti i widget di Orange, anche quelli non installati: in tal modo possiamo valutare l'utilità di ciò che non abbiamo installato ed eventualmente provvedere all'installazione nel modo illustrato prima.

La zona bianca a destra della finestra dei widget è l'area di lavoro, dove andremo a costruire il nostro workflow, fatto di widget collegati tra loro.

Inseriamo gli widgets che ci servono nell'area di lavoro scegliendo ciascuno di essi nella finestra di sinistra, dopo aver aperto il raggruppamento dove si trova, trascinando la relativa icona nell'area di lavoro o semplicemente cliccando sulla relativa icona.

L'icona inserita appare con un arco tratteggiato sulla destra e, se è richiesto un input, con un arco tratteggiato sulla sinistra.

L'workflow si costruisce collegando con il mouse archi di output e archi di input dei vari widgets.

Con doppio click su un widget nell'area di lavoro, se si tratta di un widget destinato a compiere una elaborazione apriamo una finestra in cui possiamo inserire opzioni per l'elaborazione stessa, se si tratta di un widget destinato a mostrare risultati apriamo una finestra in cui vediamo questi risultati.

Collegando ad uno di questi widget contenenti risultati il widget per la memorizzazione, che troviamo nel raggruppamento DATA, possiamo salvare i risultati stessi in un file.

3 Alcuni esempi

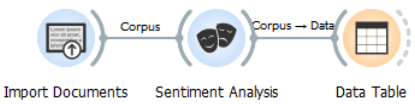
Senza altra pretesa se non quella di mostrare come concretamente funziona Orange propongo qualche semplice esempio.

3.1 Sentiment analysis

Il widget di Orange attribuisce valori positivi a testi di sentimento positivo (allegri) e valori negativi a testi di sentimento negativo (tristi), valore zero a testi normali.

Proponiamoci di analizzare le seguenti tre frasi:
. Quando passo per questa strada mi sento allegro.
. Quando incrocio i tuoi occhi tristi mi viene da piangere.
. Non riesco a scegliere un programma televisivo che mi piaccia.
che si trovano in altrettanti file di testo, denominati testo_1, testo_2 e testo_3, memorizzati in una directory chiamata testi.

Possiamo attribuire un indicatore di sentiment a ciascuna di esse utilizzando questo semplice workflow



Il primo widget (IMPORT DOCUMENTS) serve per importare i testi da esaminare. Lo troviamo nel raggruppamento TEXT MINING. Una volta trascinato nell'area di lavoro, con doppio click si di esso apriamo una finestrella in cui indichiamo la directory da cui importare i file di testo da esaminare.

Il secondo widget (SENTIMENT ANALYSIS) analizza il testo ed attribuisce l'indicatore di sentiment. Lo troviamo nel raggruppamento TEXT MINING. Una volta trascinato nell'area di lavoro lo collochiamo sulla destra del precedente e lo colleghiamo ad esso trascinando il mouse, con premuto il tasto sinistro, tra l'arco sulla destra del primo widget e l'arco sulla sinistra del secondo. A collegamento avvenuto apriamo la finestra di configurazione del widget con doppio click e scegliamo l'opzione MULTILINGUAL SENTIMENT: nella finestrella sulla destra della dicitura scorriamo e scegliamo ITALIAN (che è la lingua in cui sono scritti i testi da esaminare).

Il terzo widget (DATA TABLE) serve per leggere i risultati dell'elaborazione. Lo troviamo nel raggruppamento DATA. Una volta trascinato nell'area di lavoro lo collochiamo sulla destra del widget precedente e lo colleghiamo ad esso nel modo visto prima. Con doppio click su questo widget vediamo che al primo testo è stato attribuito sentiment positivo (11.1111), al secondo è stato attribuito sentiment negativo (-9,09091) ed il terzo testo è stato ritenuto neutrale, con sentiment zero.

Se vogliamo salvare questi risultati su file possiamo farlo cliccando sulla piccola icona (REPORT), la seconda che si trova nella barra di icone in fondo a sinistra della tabella che mostra i risultati, oppure ricorrendo al widget SAVE DATA che troviamo nel raggruppamento DATA (inserendo il widget nell'area di lavoro, collegandolo al terzo widget e configurandolo nella finestra che si apre con doppio click indicando dove e in che formato effettuare il salvataggio).

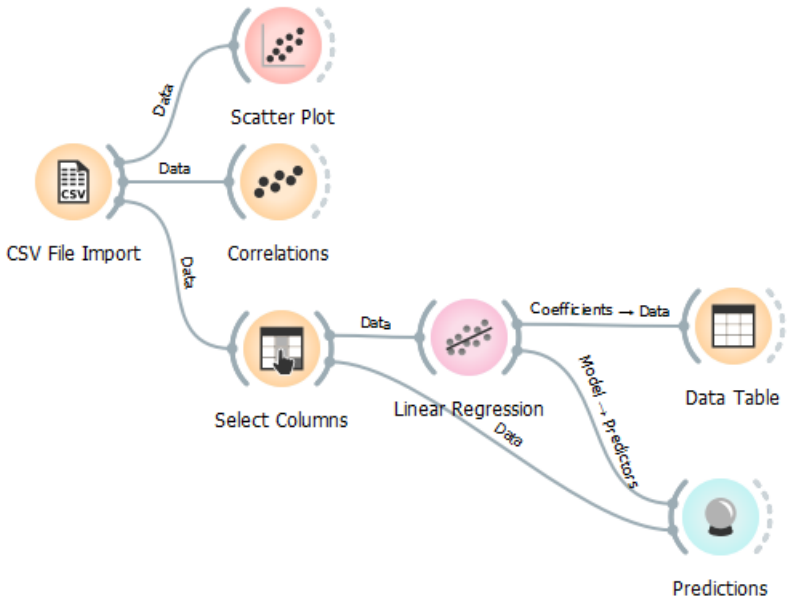
3.2 Relazioni tra dati

Poniamo di essere di fronte a questa serie di dati accostati

72	124
83	138
45	87
65	105
13	21
54	96

memorizzata in un file .csv (file di testo con dati separati da virgola) e di voler indagare sulle relazioni esistenti tra loro.

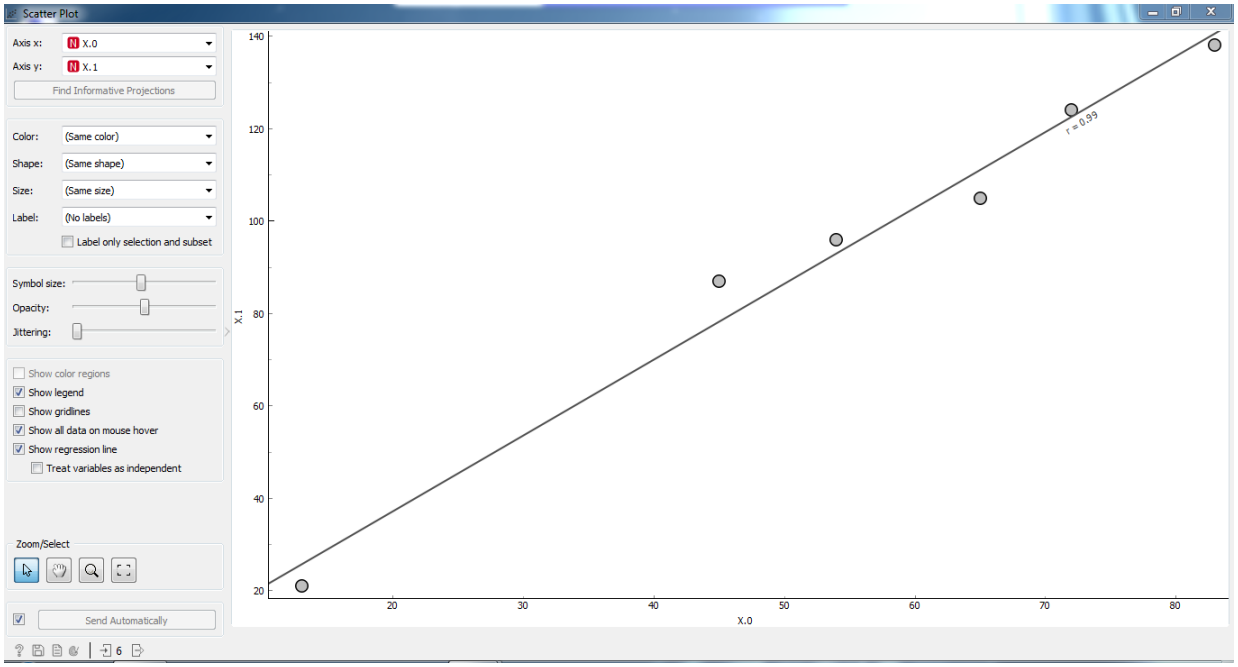
Con questo workflow



ci possiamo rendere conto di molte cose.

Innanzi tutto importiamo i dati e, visto che sono memorizzati in un file .csv, lo facciamo avvalendoci del widget CSV FILE IMPORT, che troviamo nel raggruppamento DATA. Una volta inserito il widget sul workflow ne apriamo la finestra di configurazione con doppio click e inseriamo percorso e nome del file da importare agendo nelle due finestrelle FILE:.

La prima cosa che possiamo fare è visualizzare come si relazionano i nostri dati attraverso un grafico a dispersione (scatter plot) allocando i dati su assi cartesiane ortogonali. Lo facciamo ricorrendo al widget SCATTER PLOT che troviamo nel raggruppamento VISUALIZE. Collegati i due widget nel modo che abbiamo visto nel precedente esempio, con doppio click sul widget SCATTER PLOT apriamo questa finestra



nella quale vediamo il grafico a dispersione e, avendo attivato l'opzione SHOW REGRESSION LINE, la retta interpolante accompagnata dall'indice di correlazione di Pearson `r` (del valore di 0,99, ad indicare elevatissima correlazione tra i dati).

Possiamo salvare il grafico agendo su una di queste icone in basso a sinistra . Con quella di sinistra salviamo il grafico in un formato grafico a scelta e con quella di destra lo salviamo su una pagina html.

Possiamo misurare la correlazione esistente tra i dati anche ricorrendo al widget dedicato CORRELATIONS, che troviamo nel raggruppamento DATA. Collegato questo widget a quello dell'importazione dei dati, con doppio click su di esso possiamo scegliere tra il calcolo dell'indice di correlazione di Pearson e l'indice di correlazione di Spearman.

Se poi riteniamo che tra le due serie di dati possa esistere un legame funzionale che faccia dipendere le variazioni dell'una dalle variazioni dell'altra possiamo scoprirlo e descriverlo con un modello matematico applicabile in sede previsiva.

Per fare questo dobbiamo innanzi tutto dichiarare quale, nella nostra coppia di dati, sia da considerarsi la variabile dipendente che varia in funzione dell'altra. Lo facciamo utilizzando il widget SELECT COLUMNS che troviamo nel raggruppamento DATA. Collegato questo widget a quello dell'importazione dei dati, con doppio click su di esso apriamo la finestra di configurazione e vediamo che le colonne dei dati letti sono chiamate `X0` e `X1` e sono inserite entrambe nella sotto-finestra FEATURES; se riteniamo, contrariamente a quanto fatto per default nel disegno del grafico, che la variabile dipendente sia quella della colonna `X0`, la colonna di sinistra della nostra tabella, con il mouse ne trasciniamo il simbolo nella sotto-finestra TARGET VARIABLE.

Ora ricorriamo al widget LINEAR REGRESSION, che troviamo nel raggruppamento MODEL, per descrivere in termini matematici la relazione funzionale tra le due variabili. Tipicamente la regressione lineare descrive la relazione funzionale con l'equazione di una retta del tipo $y = a + bx$ dove a è l'intercetta sull'asse verticale e b è il coefficiente angolare. Collegato questo widget al precedente che abbiamo inserito prima, con doppio click ne apriamo la finestra di configurazione e, per i nostri usi dilettanteschi, lasciamo tutto come si presenta di defalt; solo accertiamoci che sia selezionata l'opzione FIT INTERCEPT: in questo modo verrà calcolato anche il valore a dell'equazione della retta e non solo il coefficiente b.
Attraverso il widget DATA TABLE che troviamo nel raggruppamento DATA, una volta inserito nel workflow e collegato al LINEAR REGRESSION, con doppio click su di esso possiamo vedere gli elementi del nostro modello matematico:
. intercetta: -1,59722
. coefficiente della variabile indipendente `x1`: 0,598219

Visto che siamo nel machine learning la macchina ha imparato e, in previsione del manifestarsi futuro del valore, per esempio, 152 della variabile indipendente, ci dirà che, con ogni probabilità, il corrispondente valore della variabile dipendente sarà 89 (`-1,59722+0,598219x152`).

Se vogliamo avere un'idea dell'affidabilità di questa stima possiamo ricorrere al widget PREDICTIONS, che troviamo nel raggruppamento EVALUATE. Inserito questo widget nel workflow, lo colleghiamo al widget LINEAR REGRESSION e al widget SELECT COLUMNS (in tal modo applichiamo il modello regressivo ai dati della colonna di destra della nostra tabella di partenza e determiniamo i valori corrispondenti teorici della colonna di sinistra). Con doppio click sul widget apriamo una tabella in cui vediamo i nostri dati, effettivi e teorici e, soprattutto vediamo quattro indicatori di affidabilità del modello: lasciamo i primi tre agli statistici esperti e, da dilettanti, accontentiamoci del quarto, R2, detto anche R quadro o coefficiente di determinazione: nel nostro caso è pari a 0,982, valore molto elevato che ci tranquillizza sulla validità interpretativa del nostro modello.

3.3 Non tutto è lineare, però...

Supponiamo ora di trovarci di fronte a questa serie di dati accostati

12	24
45	125
56	223
84	512
92	798
116	985

Se misuriamo la correlazione sicuramente troviamo che esiste: anche a occhio vediamo che al crescere dei dati sulla sinistra crescono i dati sulla destra.

Sempre a occhio vediamo tuttavia che, al crescere dei dati sulla sinistra, i dati sulla destra crescono in maniera vistosamente più che proporzionale: da qui il sospetto che un modello lineare come quello visto nel paragrafo precedente non interpreti bene la relazione esistente tra i dati.

Orange ci offre uno strumento molto utile per ragionare su casi come questo, il widget POLYNOMIAL REGRESSION, che troviamo nel raggruppamento EDUCATION.

Lo troviamo in questo raggruppamento anziché, come il widget LINEAR REGRESSION, nel raggruppamento MODEL, probabilmente perché si tratta innanzi tutto di uno strumento di studio, che può diventare strumento di generazione di un modello dopo averci meditato molto bene.

In statistica abbiamo due tipi di interpolazione dei dati: l'interpolazione per punti noti e l'interpolazione tra punti noti.

Nel primo caso si tratta di trovare un modello matematico che disegni una linea interpolante che passi per tutti i punti dati. L'equazione di questa linea sarà un polinomio di grado pari

al numero di coppie di dati disponibili meno uno. Nel caso delle nostre sei coppie di dati un polinomio di quinto grado, del tipo

$$y = a + bx + cx^2 + dx^3 + ex^4 + fx^5$$

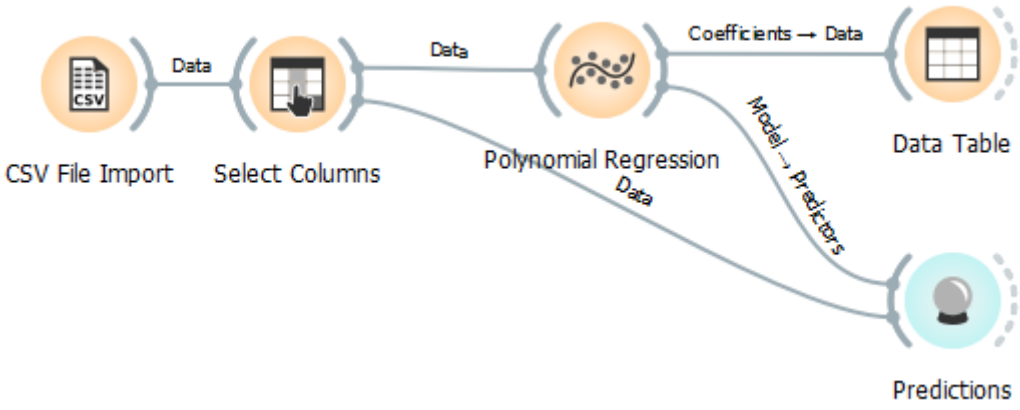
dove y (la variabile dipendente) è il dato nella colonna di destra e x (la variabile indipendente) è il dato nella colonna di sinistra.

Nel secondo caso si tratta di trovare una linea interpolante che passi attraverso i punti noti e che si avvicini il più possibile ad essi: il migliore avvicinamento si ha quando la somma dei quadrati degli scarti tra i valori effettivi e i valori stimati dalla linea è minima.

Il primo tipo di interpolazione è sicuramente valido per trovare i più probabili valori della variabile dipendente corrispondenti a valori ignoti della variabile indipendente all'interno della serie di dati noti di questa. Nel nostro caso, per esempio, per trovare il più probabile valore del dato di destra corrispondente ad un valore del dato di sinistra di 50.

Se il nostro obiettivo è quello di trovare il più probabile valore del dato di destra corrispondente ad un valore del dato di sinistra di 150, cioè quando al termine interpolazione è bene sostituire il termine estrapolazione, come a dire previsione per uno sviluppo al di fuori dai dati noti, è opinione comune che sia preferibile ragionare in termini di interpolazione tra punti noti, con preferenza per il metodo della regressione lineare, cioè con il polinomio di primo grado, la retta.

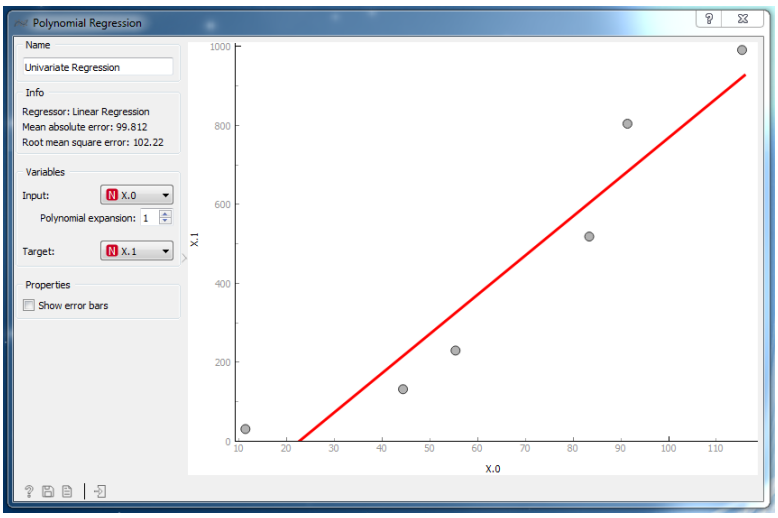
Vediamo come il widget POLYNOMIAL REGRESSION ci aiuti a ragionare su questi casi, costruendo il seguente workshop.



Importiamo i nostri dati, che ipotizzo si trovino su un file .csv, con il widget CSV FILE IMPORT; con il widget SELECT COLUMNS stabiliamo che la variabile dipendente (target) sia X.1 (nell'importazione dei dati viene dato il nome X.0 alla prima colonna e il nome X.1 alla seconda) e colleghiamo a questo widget il widget POLYNOMIAL REGRESSION. Inseriamo poi e colleghiamo i widget DATA TABLE e PREDICTIONS come abbiamo visto fare nell'esempio del precedente paragrafo.

Con doppio click sul widget POLYNOMIAL REGRESSION ne apriamo la finestra di configurazione e, nella finestrella POLYNOMIAL EXPANSION, inseriamo il grado del polinomio che intendiamo utilizzare per l'interpolazione dei dati.

Inserendo il valore 1 utilizziamo un polinomio di primo grado, che corrisponde alla linea retta. Pertanto se inseriamo il valore 1 è come se avessimo utilizzato il widget LINEAR REGRESSION e ci troviamo di fronte la seguente finestra, in cui vediamo disegnata la retta interpolante



Agendo sui widget DATA TABLE e PREDICTIONS individuiamo la seguente equazione della retta interpolante

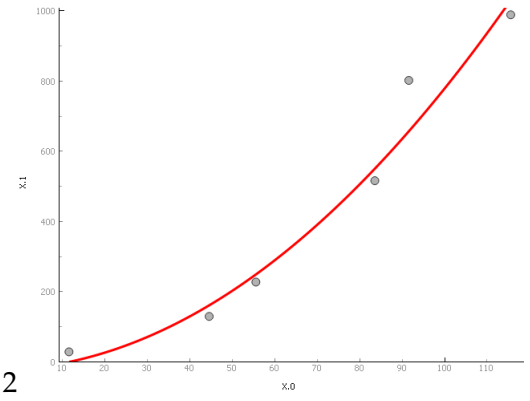
$$y = -226,879 + 9,94635x$$

e coefficiente di determinazione 0,916, elevato ma non esaltante.
Possiamo pertanto prevedere che, con buona probabilità, al valore di 150 della variabile indipendente corrisponda un valore di 1265 della variabile dipendente.

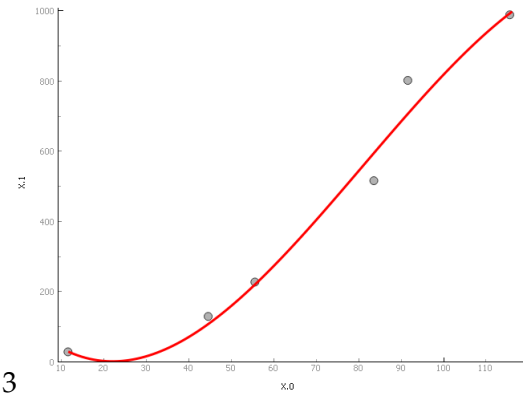
Rimane il fatto che i dati rivelano un andamento esponenziale della relazione tra di loro: lo si vede anche dal grafico se teniamo conto della differenza di scala tra gli assi delle ascisse e delle ordinate.

Agendo sulla finestrella POLYNOMIAL EXPANSION della finestra di configurazione del widget POLYNOMIAL REGRESSION possiamo vedere in tempo reale cosa succede inserendo il valore 2, alla ricerca di una parabola interpolante, il valore 3, alla ricerca di una curva flessuosa, del valore 4, fino al valore 5, corrispondente all'espressione analitica di una curva che tocchi tutti i punti del grafico (interpolazione per punti noti).

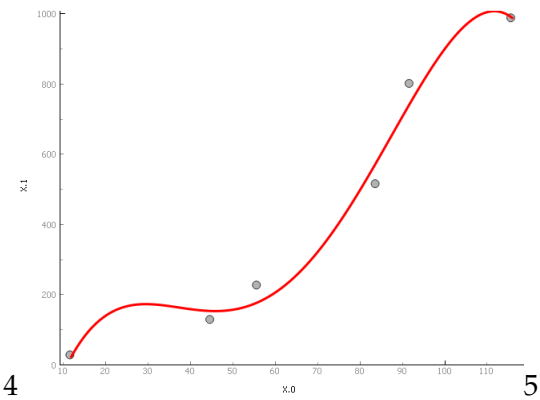
2



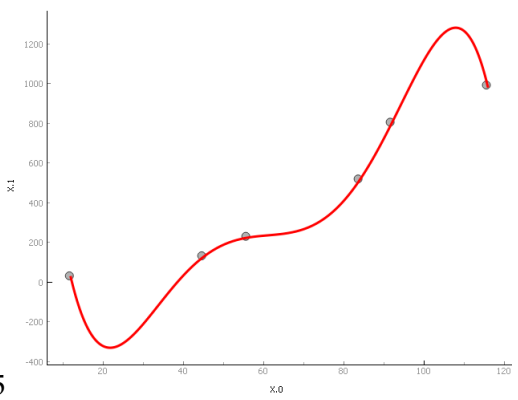
3



4



5



L'equazione della parabola di secondo grado risulta essere

$$y = -20,9841 + 0,901479x + 0,0708207x^2$$

con coefficiente di determinazione 0,965.
Possiamo pertanto prevedere che, con probabilità superiore a quella di prima, al valore di 150 della variabile indipendente corrisponda un valore di 1707 della variabile dipendente. Il valore previsto è più elevato di quello previsto con la retta in quanto qui si tiene maggiormente conto dell'andamento esponenziale.

La curva che rappresenta l'equazione di terzo grado rileva un andamento a flesso e, con un coefficiente di determinazione 0,979, lascia prevedere un valore di 1093 in corrispondenza al solito valore ipotetico di 150 della variabile indipendente. Il valore previsto è inferiore sia a quello previsto con la retta sia a quello previsto con la parabola in quanto si tiene conto della flessione del ritmo di crescita che si verifica a metà del grafico.

Con l'equazione di quarto grado la curva si avvicina sempre più ai punti noti, tanto che il coefficiente di determinazione sale a 0,987 e rileva una inversione di tendenza verso la fine del grafico.

Con l'equazione di quinto grado abbiamo una curva che tocca tutti i punti noti, pertanto il coefficiente di determinazione sale a 1, ma accentua la rilevazione dell'inversione di tendenza a fine grafico.

Le previsioni fatte applicando queste due equazioni fanno addirittura corrispondere valori negativi alla variabile dipendente in corrispondenza al valore di 150 della variabile indipendente.

Con questo esempio si dimostra come sia difficile fare previsioni, anche se si è aiutati dal rigore della matematica.

Soprattutto si dimostra che la ricerca di coefficienti di determinazione sempre più elevati può servirci per descrivere sempre meglio ciò che è successo ma può portarci parecchio fuori strada quando vogliamo capire ciò che può succedere.

Nel nostro caso specifico, comunque, ritengo valide le prime tre previsioni: quella ottimistica della parabola, quella media della retta e quella pessimistica del modello di terzo grado. Aiutino altre conoscenze a fare la scelta definitiva.

Ecco anche spiegato perché si tende comunque a preferire previsioni fatte attraverso la retta e la regressione lineare, ovviamente a patto che il coefficiente di determinazione sia abbastanza prossimo all'unità.

3.4 A proposito di coefficiente di determinazione

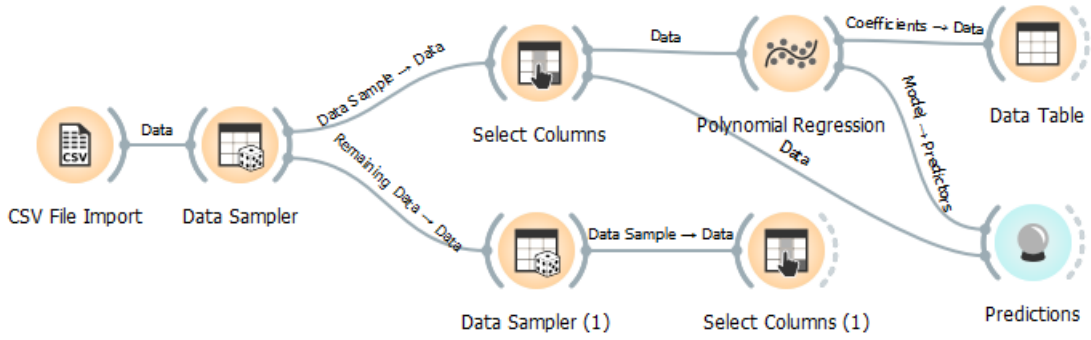
Il problema che abbiamo esaminato nel precedente paragrafo è esasperato dal fatto che abbiamo lavorato su una serie di dati molto limitata, come spesso avviene nella statistica tradizionalmente applicata a piccoli gruppi campionari o a dati di difficoltosa produzione.

Se abbiamo a che fare con dati disponibili in quantità superiore, addirittura con le grandi masse di dati che oggi sono facilmente prodotti dalla digitalizzazione di tutto ciò che facciamo, quelli che si usano chiamare big data, diventa più facile e sicuro per gli strumenti che stiamo vedendo individuare trend utili a scopo previsionale e possiamo anche meglio misurare l'affidabilità di ciò che scopriamo.

Nel caso visto nel precedente paragrafo, per esempio, data la scarsa numerosità dei dati da cui siamo partiti, il coefficiente di determinazione è stato calcolato, in modo autoreferenziale, utilizzando gli stessi dati che sono serviti per scoprire il modello.

Quando si lavora sui big data, per essere più sicuri, si scopre il modello avvalendosi di una parte dei dati a disposizione e poi lo si verifica anche su un'altra parte dei dati stessi e solo se entrambi i coefficienti di determinazione (quello calcolato sui dati utilizzati per scoprire il modello e quello calcolato sugli altri dati) sono accettabili si può essere tranquilli sulla validità del modello trovato.

Orange ci mette a disposizione un widget per estrarre da grandi masse di dati campioni più piccoli e lo possiamo utilizzare con profitto nel modo qui descritto.



Si tratta del widget DATA SAMPLER.

Importati i dati, con il widget CSV FILE IMPORT, sempre nell'ipotesi che i nostri dati siano su file .csv, inseriamo nel workflow il widget DATA SAMPLER e lo colleghiamo al precedente. Con doppio click apriamo la finestra di configurazione nella quale dobbiamo scegliere in che modo estrarre il campione di dati, avendo a disposizione due opzioni:

. FIXED PROPORTION OF DATA, scegliendo la quale indichiamo la percentuale di dati da estrarre per costruire il campione,

. FIXED SAMPLE SIZE, scegliendo la quale indichiamo il numero di dati da estrarre per costruire il campione.

A scelta effettuata premiamo il pulsante SAMPLE DATA.

A questo punto il widget ha pronto un doppio output: il DATA SAMPLE, cioè il campione di dati estratto, e i REMAINING DATA, cioè i dati che non sono entrati a far parte del campione.

Utilizziamo il DATA SAMPLE per applicare un modello regressivo (nell'illustrazione abbiamo il POLYNOMIAL REGRESSION ma potremmo avere il LINEAR REGRESSION). Preventivamente dobbiamo passare attraverso il widget SELECT COLUMNS per indicare la colonna che contiene il target.

Utilizziamo i REMAINING DATA per testare il modello regressivo. Se abbiamo a che fare con moltissimi dati conviene forse che estraiamo da questi dati un campione inserendo allo scopo un altro DATA SAMPLER, come si vede nell'illustrazione. A seguire il solito SELECT COLUMNS per indicare la colonna che contiene il target.

Per scegliere l'output giusto, visto che, per default, quando tracciamo con il mouse il collegamento, viene scelto l'output DATA SAMPLE, dobbiamo fare doppio click sulla linea del collegamento ed agire sulla finestrella del link editor che si apre.

Inserito il widget PREDICTIONS e collegatolo al widget del modello regressivo, attraverso il collegamento tra il widget PREDICTIONS e il widget SELECT COLUMNS del DATA SAMPLE, con doppio click sul widget PREDICTIONS vedremo il coefficiente di determinazione autoreferenziale calcolato sui dati da cui abbiamo derivato il modello. Se questo secondo collegamento lo effettuiamo tra il widget PREDICTIONS e il widget SELECT COLUMNS del REMANINIG DATA vedremo il coefficiente di determinazione calcolato su dati che non hanno contribuito all'individuazione del modello. Se entrambi i coefficienti di determinazione ci confortano per la loro vicinanza al numero 1 possiamo con maggiore tranquillità accettare per buono il nostro modello.

3.5 Regressione multipla[5]

In questo piccolo dataset, sempre di dimensioni sufficienti per capirci,

[5]Questo esercizio è stato affrontato nel Capitolo 3 dell'allegato in formato PDF «knime» al mio articolo «KNIME: l'alternativa a Python per i data scientists». In quella sede ho mostrato come si possa risolvere il problema con un foglio di calcolo, con uno script Python e con il software KNIME. KNIME, scritto in linguaggio Java, è un'ottima alternativa a Orange per fare le cose che stiamo vedendo senza scrivere codice.

y	x1	x2	x3
34	12	11	21
72	26	20	45
87	44	12	58
28	32	14	37
16	21	11	24
97	76	7	75

abbiamo una variabile dipendente y accostata al manifestarsi di sei diverse terne di variabili x.

Ci proponiamo di trovare il più probabile valore della y in corrispondenza delle terne

11	22	33
32	45	64
62	28	45

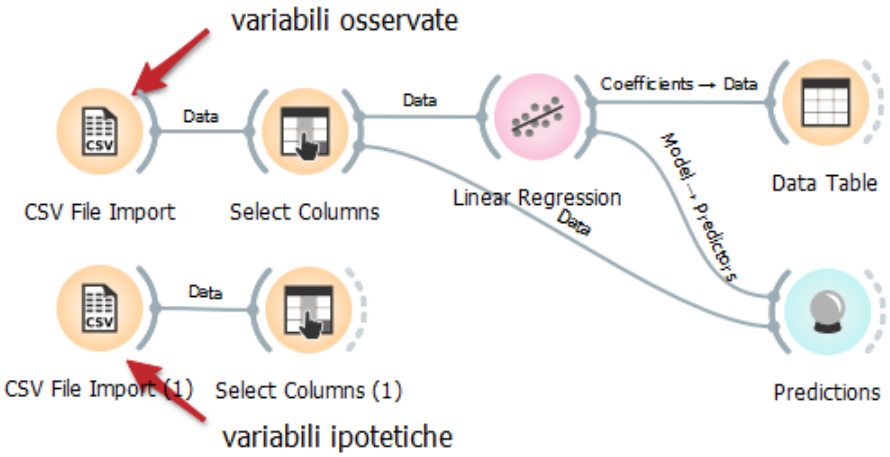
valori che dobbiamo introdurre in un file .csv in questo modo

```
,11,22,33
,32,45,64
,62,28,45
```

(iniziamo con una virgola in quanto il file deve avere la stessa formattazione di quello contenente i dati e il primo campo, quello della y, deve essere vuoto).

In questo modo abbiamo due file .csv: quello contenente le variabili osservate e quello contenente le variabili ipotetiche.

Risolviamo il nostro problema in questo modo.



Carichiamo il file .csv delle variabili osservate, nel relativo widget SELECT COLUMNS spostiamo la colonna `X0` nella finestrella TARGET e poi proseguiamo come abbiamo fatto nel Paragrafo 3.2 con i widget LINEAR REGRESSION, DATA TABLE e PREDICTIONS.
Con doppio click sul widget DATA TABLE troviamo gli elementi per scrivere il modello regressivo trovato:

$$y = 2,87997 - 2,64835x_1 - 2,61655x_2 + 4,12217x_3$$

Collegati i widget SELECT COLUMNS e PREDICTIONS, con doppio click sul widget PREDICTIONS constatiamo il coefficiente di determinazione di ottimo livello 0,973.

Ora importiamo il file .csv delle variabili ipotetiche, lo colleghiamo a un nuovo widget SELECT COLUMNS e lasciamo tutte le colonne importate nelle FEATURES senza indicare alcun TARGET.
Collegato questo nuovo SELECT COLUMNS a PREDICTIONS, con doppio click sul widget PREDICTIONS vediamo che il valore calcolato in corrispondenza alla terna 11 22 33 è 52, quello calcolato in corrispondenza della terna 32 45 64 è 64 e quello calcolato in corrispondenza alla terna 62 28 45 è -49.

3.6 Clustering[6]

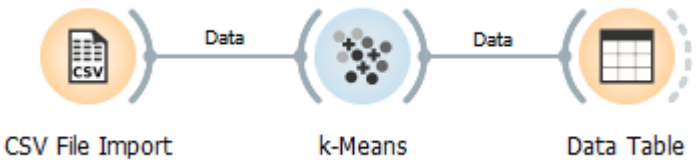
Abbiamo fatto un'indagine sui nostri clienti - nel caso sono solo 12 ma potrebbero anche essere 120.000 - e siamo venuti a conoscere ciò che vediamo in questa tabella, memorizzata in un file .csv:

cliente	cambio	prezzo	marca	frequenza
Luigi	5	5	5	3
Maria	7	7	4	2
Vittorio	6	5	5	3
Giovanni	6	6	4	2
Beatrice	5	5	4	3
Giuseppe	9	8	2	4
Elena	8	8	2	5
Antonio	9	7	3	4
Paola	10	7	3	4
Mario	9	8	2	4
Carlo	9	10	2	5
Cecilia	10	8	3	3

Le grandezze numeriche variano da 1 (per niente) a 10 (moltissimo).
La colonna «cambio» indica la propensione del cliente a cambiare negozio.
La colonna «prezzo» indica la sensibilità del cliente al prezzo della merce.
La colonna «marca» indica la preferenza del cliente per articoli di marca.
La colonna «frequenza» indica il numero medio di acquisti in un anno.

Aiutati dalla piccola dimensione del data set e dall'ordine con cui ho esposto i dati notiamo subito a occhio che abbiamo un gruppo di clienti, i primi cinque, tendenzialmente quelli che fanno meno acquisti, che hanno una moderata propensione a cambiare negozio, e sono moderatamente sensibili al prezzo e alla marca; questo gruppo si contrappone al gruppo degli altri sette, con molto più elevata propensione a cambiare negozio, molto più sensibili al prezzo, meno sensibili alla marca e mediamente migliori clienti sul piano della frequenza di acquisto. Tutte cose utili per le nostre strategie di marketing o semplicemente per qualche intervento tendente a fidelizzare meglio i sette clienti più «volatili».

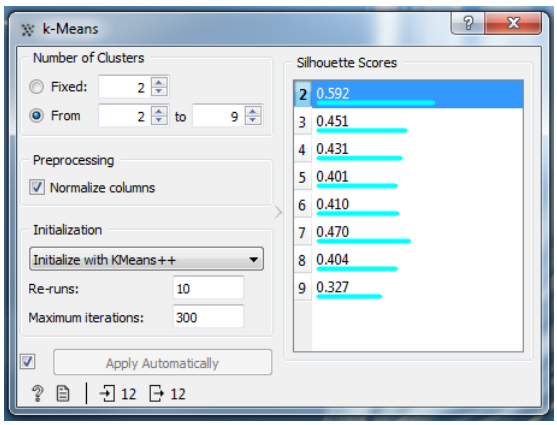
Per compiere questa analisi non più a occhio ma in maniera rigorosa, anche quando i dati sono migliaia e non solo 12 come in questo esempio, Orange ci mette a disposizione il widget K-MEANS.



Al solito, importiamo innanzitutto i dati con il widget CSV FILE IMPORT e poi colleghiamo a questo widget il widget K-MEANS, che troviamo nel raggruppamento UNSUPERVISED.

Con doppio click su questo widget ne apriamo la finestra di configurazione

[6]Questo esercizio è stato affrontato nel Capitolo 5 dell'allegato in formato PDF «knime» al mio articolo «KNIME: l'alternativa a Python per i data scientists». In quella sede ho mostrato come si possa risolvere il problema con uno script Python e con il software KNIME. KNIME, scritto in linguaggio Java, è un'ottima alternativa a Orange per fare le cose che stiamo vedendo senza scrivere codice.



Quella che si apre di default è solo la parte di sinistra della finestra, senza la parte di destra intitolata SILHOUETTE SCORES.

Se abbiamo già le idee chiare sul numero di cluster che intendiamo individuare, scegliamo l'opzione FIXED, nella finestrella indichiamo il numero dei cluster e chiudiamo.

Se vogliamo essere preventivamente orientati sul numero dei cluster più attendibili che possiamo individuare nel nostro dataset scegliamo l'opzione FROM... TO... indicando le relative quantità ipotetiche. Per il nostro piccolo dataset la figura mostra la scelta di mettere a prova un numero di cluster compreso tra 2 e 9.

Un cluster è attendibile quando gli oggetti che contiene sono sufficientemente simili tra loro e sufficientemente diversi dagli oggetti appartenenti ad altri gruppi.

Tra i vari modi per rappresentare con un indicatore l'attendibilità, Orange, per il suo k-Means, ha scelto l'indice di silhouette. Esso varia tra -1 e 1 e l'esperienza insegna che

. se è compreso tra 0,70 e 1 il partizionamento è estremamente attendibile,

. se è compreso tra 0,50 e 0,70 il partizionamento è attendibile,

. se è compreso tra 0,25 e 0,50 il partizionamento è poco attendibile,

. se è compreso tra -1 e 0,25 il partizionamento non è per niente attendibile.

Scelta e configurata l'opzione con l'indicazione dei cluster da sottoporre al test della silhouette, nella parte destra della finestra compare la lista degli indici di silhouette corrispondenti al range di possibili quantità di cluster.

Nel nostro caso vediamo che il valore massimo dell'indicatore corrisponde ad un partizionamento su due cluster con indice di silhouette 0,592, ad indicare che anche il partizionamento su due cluster è appena attendibile. Per tutti gli altri possibili partizionamenti non si raggiunge la soglia di attendibilità.

Selezioniamo pertanto, come fatto in figura, la riga corrispondente a 2 cluster e chiudiamo.

Colleghiamo un widget DATA TABLE in cui far confluire il risultato del partizionamento, che vediamo corrispondere esattamente all'impressione che avevamo percepito a occhio.

Se vogliamo salvare i risultati in formato html possiamo utilizzare il pulsante REPORT del widget DATA TABLE. Per un salvataggio in altri formati dobbiamo ricorrere al widget SAVE DATA, preventivamente selezionando, nella DATA TABLE, quanto vogliamo salvare.