Dispensa di Machine Learning

Federico Luzzi
Christian Uccheddu

1 Introduzione

Gli ambiti più importanti nei quali vengono applicate tecniche di machine learning nella vita reale sono diversi, in particolare si usano in: finanza, sanità, agricoltura, e-commerce, social, chatbot, sensoristica (come i veicoli a guida autonoma).

Vista la grande mole di dati la necessità è capire come trattare i dati.

L'obiettivo del Machine Learning è sviluppare metodologia per dare valore ai dati in funzione di una particolare domanda che ci stiamo facendo.

Tipicamente le tecninche di machine learning si dividono nelle seguenti tre macro categorie:

- Apprendimento supervisionato o predittivo: qualcuno ha gia catalogato ad esempio delle immagini o dei dati e noi prendendo questi modelli dovremmo essere in grado di predire.
- Apprendimento descrittivo: ci sono delle funzioni obiettivo che vanno ottimizzate. Non usiamo etichette della singola istanza ma in qualche modo sappiamo dove arrivare.
- Apprendimento rinforzato: E' quello più utile in questa epoca: funziona sui premi.

In questo corso ci concentreremo sui primi due tipi di apprendimento.

L' *apprendimento supervisionato* si divide a sua volta nelle seguenti due categorie:

- Classificazione: quando il problema consiste nel dividere in classi delle quantità discrete.
- Regressione: quando il problema consiste nel ricostruire una certa variabile date delle condizioni pregresse.

L' *apprendimento non supervisionato* si divide a sua volta in:

- Clustering: Quando il problema consistere nel ricostruire delle classi delle istanze che ci vengono consegnate, senza sapere nulla sulla storia pregressa.
- Associativa: Quando il problema consiste nel scoprire pattern che descrivono bene caratteristiche associate ad un certo fenomeno.

Per alcuni compiti la correlazione statistica va benissimo, in alcuni casi però essa è addirittura deleteria. Se infatti provassi a vedere la correlazione tra il numero di omicidi in america e il numero di fondi investiti sulla ricerca scientifica vedrei che statisticamente sono strettamente correlate. Questo è un no-sense ed è il classico esempio di *correlazione spuria.*

A confermare questa visione in cui il modello per una certa trattazione dati è fondamentale è il cosidetto paradosso di Simpson.

Paradosso di Simpson: Se uso solo i dati senza modello no c'è alcun modo di scoprire la verità.

1.1 Data Types

Il primo passo fondamentale è sicuramente quello di prendere confidenza coi dati. E' fondamentale capire la natura intrinseca dei dati che abbiamo a disposizione, di solito essi sono organizzati in strutture che chiamiamo **dataset**. Per le analisi con le tecniche di Machine Learning esse non sono altro che tabelle fatte da righe e da colonne.

Ci possono essere valori missing in un dataset e possono mancare per diversi motivi.

Le colonne sono chiamate **Attributi.**

Le righe sono chiamate **Istanze.**

A volte ci sono anche attributi duplicati che possiamo tranquillamente buttare fuori. Ogni attributo è caratterizzato dal fatto di avere un tipo. Conoscerlo è fondamentale per trattare i dati. Gli attributi si dividono in due grandi gruppi:

- Categorici:
  - Nominali: ad esempio il colore degli occhi.
  - Ordinali: ad esempio possono essere i giudizi.
- Numerici:
  - Intervallo: ammettono operazioni di somma e sottrazione.
  - Ratio: possiamo applicare tutte le operazioni logico/matematiche.

Dall'alto al basso il livello gerarchico sale e le proprietà aumentano.
Si possono anche dividere in attributi *discreti* che possono essere:

- Categorici
- Numerici

- Binari: sono i più particolari da trattare, e hanno una serie di proprietà strane.

Oppure possono essere *Continui.*

1.2 Data exploration

Dobbiamo però anche sapere come esplorare i dati. Per farlo facciamo cose molto elementari. Per farlo si usano tutti gli strumenti a nostra disposizione.

Il concetto di **quantile** è trovare il numero di osservazione che ci indica quanti attributi sono più piccoli di un dato valore. Un quantile molto importante è il quantile di ordine $\frac{1}{2}$ e si chiama **Mediana** che è quello che ha esattamente minori di lui la metà dei dati.

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media non è un buon modo di visualizzare i dati perché dice poco ma quanto meno dice qualcosa. Siccome la media è basata su singole osservazioni si possono vedere le presenze di outliar, ossia di elementi troppo discordanti dalla media e che si presenta poche volte. Sono quindi elementi che è necessario trattare per vedere la provenienza.

Per prevenire questo si usa la **media trimmed** in cui si buttano via il valore più piccolo e il valore più grande. Se si trova un grosso scostamento probabilmente è presente un outlier.

Si può definire anche il range anche se di solito si usa la varianza:

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Di solito si usa la deviazione standard che ha lo stesso ordine di grandezza dei dati:

$$std = \sqrt{var}$$

Si usa anche il range interquartile (IQR) sempre per ovviare alla presenza di outliar. Se ho a che fare con coppie di attributi allora è naturale parlare di **covarianza** ossia la varianza calcolata su due attributi diversi:

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Di solito si fanno scalature perché sennò le covarianze vengono troppo sbagliate. Per ovviare uso la **correlazione di Pearson** che può prendere valori $[-1, 1]$

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

Possono organizzare i dati in istogrammi in cui posso usare una ampiezza fissa o variabile (bin). A seconda dell'ampiezza che uso posso ottenere due disegni molto diversi.

Un altro modo di rappresentare i dati particolarmente utile è il **grafico Box and Whiskers** applicato solo ad attributi quantitativi.

1.3 Missing replacement

E' un problema enorme di per sè. Le operazioni più elementari sono le seguenti. In alcuni valori degli attributi un valore non è registrato. Ci possono esser tante ragioni: ad esempio un attributo non è sempre stato osservabile (penso all'ambito clinico), o ad esempio un attributo prima non veniva considerato rilevante.

Il primo nmetodo è il **Record removal** che è molto drastico come metodo perché comunque sia vengono eliminati dei valori che sarebbero potuti essere molto importanti.

Il secondo metodo è quello di **imputazione manuale**: è fatta da umani e tramite osservazioni ci si chiede se sia possibile inserirlo, è tremendamente difficile dal punto di vista computazionale.

Il terzo metodo è quello della **global constant**: ossia metto un numero là chiamato place holder con un valore costante, non troppo efficiente.

Il quarto metodo è quello di **rimpiazzarlo con la moda**, anche questo però è fortemente criticabile. Se gli attributi sono continui si fa la stessa cosa ma con la media.

Il quinto metodo è **Conditional mean replacement** ossia bisogna rimpiazzare con la media solo se è presente un altro determinato attributo.

Il sesto metodo è quello del **most probable** ossia di prendere un modello e sostituire il valore.

E' molto difficile dare il confine tra l'esplorazione dei dati e la modellizzazione dei dati.

1.4 Data Preprocessing

Tutti questi processi impattano fortemente tutta l'analisi che farò dopo.

2 Introduzione

Gli ambiti più importanti nei quali vengono applicate tecniche di machine learning nella vita reale sono:

- Finanza
- Sanità
- Agricoltura
- e-commerce
- Social
- Chatbot
- Sensoristica (come i veicoli a guida autonoma)

Vista la grande mole di dati la necessità è capire come trattare i dati.

L'obiettivo del Machine Learning è sviluppare metodologia per dare valore ai dati in funzione di una particolare domanda che ci stiamo facendo.

Tipicamente le tecninche di machine learning si dividono nelle seguenti tre macro categorie:

- Apprendimento supervisionato o predittivo: qualcuno ha gia catalogato ad esempio delle immagini o dei dati e noi prendendo questi modelli dovremmo essere in grado di predire.
- Apprendimento descrittivo: ci sono delle funzioni obiettivo che vanno ottimizzate. Non usiamo etichette della singola istanza ma in qualche modo sappiamo dove arrivare.
- Apprendimento rinforzato: E' quello più utile in questa epoca: funziona sui premi.

In questo corso ci concentreremo sui primi due tipi di apprendimento.

L' *apprendimento supervisionato* si divide a sua volta nelle seguenti due categorie:

- Classificazione: quando il problema consiste nel dividere in classi delle quantità discrete.
- Regressione: quando il problema consiste nel ricostruire una certa variabile date delle condizioni pregresse.

L' *apprendimento non supervisionato* si divide a sua volta in:

- Clustering: Quando il problema consistere nel ricostruire delle classi delle istanze che ci vengono consegnate, senza sapere nulla sulla storia pregressa.
- Associativa: Quando il problema consiste nel scoprire pattern che descrivono bene caratteristiche associate ad un certo fenomeno.

Per alcuni compiti la correlazione statistica va benissimo, in alcuni casi però essa è addirittura deleteria. Se infatti provassi a vedere la correlazione tra il numero di omicidi in america e il numero di fondi investiti sulla ricerca scientifica vedrei che statisticamente sono strettamente correlate. Questo è un no-sense ed è il classico esempio di *correlazione spuria.*

A confermare questa visione in cui il modello per una certa trattazione dati è fondamentale è il cosidetto paradosso di Simpson.

Paradosso di Simpson: Se uso solo i dati senza modello no c'è alcun modo di scoprire la verità.

2.1 Data Types

Il primo passo fondamentale è sicuramente quello di prendere confidenza coi dati. E' fondamentale capire la natura dei dati che abbiamo a disposizione (**dataset**), c'è un fenomeno che si chiama **churn**, quando non siamo soddisfatti di un servizio ci affidiamo al competitor.

Ci possono essere valori missing in un dataset e possono mancare per diversi motivi.

Le colonne sono chiamate **Attributi.**

Le righe sono chiamate **Istanze.**

A volte ci sono anche attributi duplicati che possiamo tranquillamente buttare fuori. Ogni attributo è caratterizzato dal fatto di avere un tipo. Conoscerlo è fondamentale per trattare i dati. Gli attributi si dividono in due grandi gruppi:

- Categorici:
  - Nominali: ad esempio il colore degli occhi.
  - Ordinali: ad esempio possono essere i giudizi.
- Numerici:
  - Intervallo: ammettono operazioni di somma e sottrazione.
  - Ratio: possiamo applicare tutte le operazioni logico/matematiche.

Dall'alto al basso il livello gerarchico sale e le proprietà aumentano.
Si possono anche dividere in attributi *discreti* che possono essere:

- Categorici
- Numerici
- Binari: sono i più particolari da trattare, e hanno una serie di proprietà strane.

Oppure possono essere *Continui.*

2.2 Data exploration

Dobbiamo però anche sapere come esplorare i dati. Per farlo facciamo cose molto elementari. Per farlo si usano tutti gli strumenti a nostra disposizione.

Il concetto di **quantile** è trovare il numero di osservazione che ci indica quanti attributi sono più piccoli di un dato valore. Un quantile molto importante è il quantile di ordine $\frac{1}{2}$ e si chiama **Mediana** che è quello che ha esattamente minori di lui la metà dei dati.

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media non è un buon modo di visualizzare i dati perché dice poco ma quanto meno dice qualcosa. Siccome la media è basata su singole osservazioni si possono vedere le presenze di outliar, ossia di elementi troppo discordanti dalla media e che si presenta poche volte. Sono quindi elementi che è necessario trattare per vedere la provenienza.

Per prevenire questo si usa la **media trimmed** in cui si buttano via il valore più piccolo e il valore più grande. Se si trova un grosso scostamento probabilmente è presente un outlier.

Si puù definire anche il range anche se di solito si usa la varianza:

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Di solito si usa la deviazione standard che ha lo stesso ordine di grandezza dei dati:

$$std = \sqrt{var}$$

Si usa anche il range interquartile (IQR) sempre per ovviare alla presenza di outliar. Se ho a che fare con coppie di attributi allora è naturale parlare di **covarianza** ossia la varianza calcolata su due attributi diversi:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Di solito si fanno scalature perché sennò le covarianze vengono troppo sbagliate. Per ovviare uso la **correlazione di Pearson** che può prendere valori $[-1,1]$

$$corr(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

Possono organizzare i dati in istogrammi in cui posso usare una ampiezza fissa o variabile (bin). A seconda dell'ampiezza che uso posso ottenere due disegni molto diversi.

Un altro modo di rappresentare i dati particolarmente utile è il **grafico Box and Whiskers** applicato solo ad attributi quantitativi.

2.3 Missing replacement

E' un problema enorme di per sè. Le operazioni più elementari sono le seguenti. In alcuni valori degli attributi un valore non è registrato. Ci possono esser tante ragioni: ad esempio un attributo non è sempre stato osservabile (penso all'ambito clinico), o ad esempio un attributo prima non veniva considerato rilevante.

Il primo nmetodo è il **Record removal** che è molto drastico come metodo perché comunque sia vengono eliminati dei valori che sarebbero potuti essere molto importanti.

Il secondo metodo è quello di **imputazione manuale**: è fatta da umani e tramite osservazioni ci si chiede se sia possibile inserirlo, è tremendamente difficile dal punto di vista computazionale.

Il terzo metodo è quello della **global constant**: ossia metto un numero là chiamato place holder con un valore costante, non troppo efficiente.

Il quarto metodo è quello di **rimpiazzarlo con la moda**, anche questo però è fortemente criticabile. Se gli attributi sono continui si fa la stessa cosa ma con la media.

Il quinto metodo è **Conditional mean replacement** ossia bisogna rimpiazzare con la media solo se è presente un altro determinato attributo.

Il sesto metodo è quello del **most probable** ossia di prendere un modello e sostituire il valore.

E' molto difficile dare il confine tra l'esplorazione dei dati e la modellizzazione dei dati.

2.4 Data Preprocessing

Tutti questi processi impattano fortemente tutta l'analisi che farò dopo.

3 Introduzione

Gli ambiti più importanti nei quali vengono applicate tecniche di machine learning nella vita reale sono:

- Finanza
- Sanità
- Agricoltura
- e-commerce
- Social
- Chatbot
- Sensoristica (come i veicoli a guida autonoma)

Vista la grande mole di dati la necessità è capire come trattare i dati.
L'obiettivo del Machine Learning è sviluppare metodologia per dare valore ai dati in funzione di una particolare domanda che ci stiamo facendo.
Tipicamente le tecninche di machine learning si dividono nelle seguenti tre macro categorie:

- Apprendimento supervisionato o predittivo: qualcuno ha gia catalogato ad esempio delle immagini o dei dati e noi prendendo questi modelli dovremmo essere in grado di predire.
- Apprendimento descrittivo: ci sono delle funzioni obiettivo che vanno ottimizzate. Non usiamo etichette della singola istanza ma in qualche modo sappiamo dove arrivare.
- Apprendimento rinforzato: E' quello più utile in questa epoca: funziona sui premi.

In questo corso ci concentreremo sui primi due tipi di apprendimento.

L' *apprendimento supervisionato* si divide a sua volta nelle seguenti due categorie:

- Classificazione: quando il problema consiste nel dividere in classi delle quantità discrete.
- Regressione: quando il problema consiste nel ricostruire una certa variabile date delle condizioni pregresse.

L' *apprendimento non supervisionato* si divide a sua volta in:

- Clustering: Quando il problema consistere nel ricostruire delle classi delle istanze che ci vengono consegnate, senza sapere nulla sulla storia pregressa.
- Associativa: Quando il problema consiste nel scoprire pattern che descrivono bene caratteristiche associate ad un certo fenomeno.

Per alcuni compiti la correlazione statistica va benissimo, in alcuni casi però essa è addirittura deleteria. Se infatti provassi a vedere la correlazione tra il numero di omicidi in america e il numero di fondi investiti sulla ricerca scientifica vedrei che statisticamente sono strettamente correlate. Questo è un no-sense ed è il classico esempio di *correlazione spuria.*

A confermare questa visione in cui il modello per una certa trattazione dati è fondamentale è il cosidetto paradosso di Simpson.

Paradosso di Simpson: Se uso solo i dati senza modello no c'è alcun modo di scoprire la verità.

3.1 Data Types

Il primo passo fondamentale è sicuramente quello di prendere confidenza coi dati. E' fondamentale capire la natura dei dati che abbiamo a disposizione (**dataset**), c'è un fenomeno che si chiama **churn**, quando non siamo soddisfatti di un servizio ci affidiamo al competitor.

Ci possono essere valori missing in un dataset e possono mancare per diversi motivi.

Le colonne sono chiamate **Attributi.**

Le righe sono chiamate **Istanze.**

A volte ci sono anche attributi duplicati che possiamo tranquillamente buttare fuori. Ogni attributo è caratterizzato dal fatto di avere un tipo. Conoscerlo è fondamentale per trattare i dati. Gli attributi si dividono in due grandi gruppi:

- Categorici:
  - Nominali: ad esempio il colore degli occhi.
  - Ordinali: ad esempio possono essere i giudizi.
- Numerici:
  - Intervallo: ammettono operazioni di somma e sottrazione.
  - Ratio: possiamo applicare tutte le operazioni logico/matematiche.

Dall'alto al basso il livello gerarchico sale e le proprietà aumentano.
Si possono anche dividere in attributi *discreti* che possono essere:

- Categorici
- Numerici
- Binari: sono i più particolari da trattare, e hanno una serie di proprietà strane.

Oppure possono essere *Continui.*

3.2 Data exploration

Dobbiamo però anche sapere come esplorare i dati. Per farlo facciamo cose molto elementari. Per farlo si usano tutti gli strumenti a nostra disposizione.

Il concetto di **quantile** è trovare il numero di osservazione che ci indica quanti attributi sono più piccoli di un dato valore. Un quantile molto importante è il quantile di ordine $\frac{1}{2}$ e si chiama **Mediana** che è quello che ha esattamente minori di lui la metà dei dati.

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media non è un buon modo di visualizzare i dati perché dice poco ma quanto meno dice qualcosa. Siccome la media è basata su singole osservazioni si possono vedere le presenze di outliar, ossia di elementi troppo discordanti dalla media e che si presenta poche volte. Sono quindi elementi che è necessario trattare per vedere la provenienza.

Per prevenire questo si usa la **media trimmed** in cui si buttano via il valore più piccolo e il valore più grande. Se si trova un grosso scostamento probabilmente è presente un outlier.

Si può definire anche il range anche se di solito si usa la varianza:

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Di solito si usa la deviazione standard che ha lo stesso ordine di grandezza dei dati:

$$std = \sqrt{var}$$

Si usa anche il range interquartile (IQR) sempre per ovviare alla presenza di outliar. Se ho a che fare con coppie di attributi allora è naturale parlare di **covarianza** ossia la varianza calcolata su due attributi diversi:

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Di solito si fanno scalature perché sennò le covarianze vengono troppo sbagliate. Per ovviare uso la **correlazione di Pearson** che può prendere valori $[-1, 1]$

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

Possono organizzare i dati in istogrammi in cui posso usare una ampiezza fissa o variabile (bin). A seconda dell'ampiezza che uso posso ottenere due disegni molto diversi.

Un altro modo di rappresentare i dati particolarmente utile è il **grafico Box and Whiskers** applicato solo ad attributi quantitativi.

3.3 Missing replacement

E' un problema enorme di per sè. Le operazioni più elementari sono le seguenti. In alcuni valori degli attributi un valore non è registrato. Ci possono esser tante ragioni: ad esempio un attributo non è sempre stato osservabile (penso all'ambito clinico), o ad esempio un attributo prima non veniva considerato rilevante.

Il primo nmetodo è il **Record removal** che è molto drastico come metodo perché comunque sia vengono eliminati dei valori che sarebbero potuti essere molto importanti.

Il secondo metodo è quello di **imputazione manuale**: è fatta da umani e tramite osservazioni ci si chiede se sia possibile inserirlo, è tremendamente difficile dal punto di vista computazionale.

Il terzo metodo è quello della **global constant**: ossia metto un numero là chiamato place holder con un valore costante, non troppo efficiente.

Il quarto metodo è quello di **rimpiazzarlo con la moda**, anche questo però è fortemente criticabile. Se gli attributi sono continui si fa la stessa cosa ma con la media.

Il quinto metodo è **Conditional mean replacement** ossia bisogna rimpiazzare con la media solo se è presente un altro determinato attributo.

Il sesto metodo è quello del **most probable** ossia di prendere un modello e sostituire il valore.

E' molto difficile dare il confine tra l'esplorazione dei dati e la modellizzazione dei dati.

3.4 Data Preprocessing

Tutti questi processi impattano fortemente tutta l'analisi che farò dopo.

4 Introduzione

Gli ambiti più importanti nei quali vengono applicate tecniche di machine learning nella vita reale sono:

- Finanza
- Sanità
- Agricoltura
- e-commerce
- Social
- Chatbot
- Sensoristica (come i veicoli a guida autonoma)

Vista la grande mole di dati la necessità è capire come trattare i dati.

L'obiettivo del Machine Learning è sviluppare metodologia per dare valore ai dati in funzione di una particolare domanda che ci stiamo facendo.

Tipicamente le tecninche di machine learning si dividono nelle seguenti tre macro categorie:

- Apprendimento supervisionato o predittivo: qualcuno ha gia catalogato ad esempio delle immagini o dei dati e noi prendendo questi modelli dovremmo essere in grado di predire.

- Apprendimento descrittivo: ci sono delle funzioni obiettivo che vanno ottimizzate. Non usiamo etichette della singola istanza ma in qualche modo sappiamo dove arrivare.

- Apprendimento rinforzato: E' quello più utile in questa epoca: funziona sui premi.

In questo corso ci concentreremo sui primi due tipi di apprendimento.
L' *apprendimento supervisionato* si divide a sua volta nelle seguenti due categorie:

- Classificazione: quando il problema consiste nel dividere in classi delle quantità discrete.

- Regressione: quando il problema consiste nel ricostruire una certa variabile date delle condizioni pregresse.

L' *apprendimento non supervisionato* si divide a sua volta in:

- Clustering: Quando il problema consistere nel ricostruire delle classi delle istanze che ci vengono consegnate, senza sapere nulla sulla storia pregressa.

- Associativa: Quando il problema consiste nel scoprire pattern che descrivono bene caratteristiche associate ad un certo fenomeno.

Per alcuni compiti la correlazione statistica va benissimo, in alcuni casi però essa è addirittura deleteria. Se infatti provassi a vedere la correlazione tra il numero di omicidi in america e il numero di fondi investiti sulla ricerca scientifica vedrei che statisticamente sono strettamente correlate. Questo è un no-sense ed è il classico esempio di *correlazione spuria.*
A confermare questa visione in cui il modello per una certa trattazione dati è fondamentale è il cosidetto paradosso di Simpson.
Paradosso di Simpson: Se uso solo i dati senza modello no c'è alcun modo di scoprire la verità.

4.1 Data Types

Il primo passo fondamentale è sicuramente quello di prendere confidenza coi dati. E' fondamentale capire la natura dei dati che abbiamo a disposizione (**dataset**), c'è un fenomeno che si chiama **churn**, quando non siamo soddisfatti di un servizio ci affidiamo al competitor.

Ci possono essere valori missing in un dataset e possono mancare per diversi motivi.

Le colonne sono chiamate **Attributi.**

Le righe sono chiamate **Istanze.**

A volte ci sono anche attributi duplicati che possiamo tranquillamente buttare fuori. Ogni attributo è caratterizzato dal fatto di avere un tipo. Conoscerlo è fondamentale per trattare i dati. Gli attributi si dividono in due grandi gruppi:

- Categorici:
  - Nominali: ad esempio il colore degli occhi.
  - Ordinali: ad esempio possono essere i giudizi.
- Numerici:
  - Intervallo: ammettono operazioni di somma e sottrazione.
  - Ratio: possiamo applicare tutte le operazioni logico/matematiche.

Dall'alto al basso il livello gerarchico sale e le proprietà aumentano.

Si possono anche dividere in attributi *discreti* che possono essere:

- Categorici
- Numerici
- Binari: sono i più particolari da trattare, e hanno una serie di proprietà strane.

Oppure possono essere *Continui.*

4.2 Data exploration

Dobbiamo però anche sapere come esplorare i dati. Per farlo facciamo cose molto elementari. Per farlo si usano tutti gli strumenti a nostra disposizione.

Il concetto di **quantile** è trovare il numero di osservazione che ci indica quanti attributi sono più piccoli di un dato valore. Un quantile molto importante è il quantile di ordine $\frac{1}{2}$ e si chiama **Mediana** che è quello che ha esattamente minori di lui la metà dei dati.

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media non è un buon modo di visualizzare i dati perché dice poco ma quanto meno dice qualcosa. Siccome la media è basata su singole osservazioni si possono vedere le presenze di outliar, ossia di elementi troppo discordanti dalla media e che si presenta poche volte. Sono quindi elementi che è necessario trattare per vedere la provenienza.

Per prevenire questo si usa la **media trimmed** in cui si buttano via il valore più piccolo e il valore più grande. Se si trova un grosso scostamento probabilmente è presente un outlier.

Si può definire anche il range anche se di solito si usa la varianza:

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Di solito si usa la deviazione standard che ha lo stesso ordine di grandezza dei dati:

$$std = \sqrt{var}$$

Si usa anche il range interquartile (IQR) sempre per ovviare alla presenza di outliar. Se ho a che fare con coppie di attributi allora è naturale parlare di **covarianza** ossia la varianza calcolata su due attributi diversi:

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Di solito si fanno scalature perché sennò le covarianze vengono troppo sbagliate. Per ovviare uso la **correlazione di Pearson** che può prendere valori $[-1, 1]$

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

Possono organizzare i dati in istogrammi in cui posso usare una ampiezza fissa o variabile (bin). A seconda dell'ampiezza che uso posso ottenere due disegni molto diversi.

Un altro modo di rappresentare i dati particolarmente utile è il **grafico Box and Whiskers** applicato solo ad attributi quantitativi.

4.3 Missing replacement

E' un problema enorme di per sè. Le operazioni più elementari sono le seguenti. In alcuni valori degli attributi un valore non è registrato. Ci possono esser tante ragioni: ad esempio un attributo non è sempre stato osservabile (penso all'ambito clinico), o ad esempio un attributo prima non veniva considerato rilevante.

Il primo nmetodo è il **Record removal** che è molto drastico come metodo perché comunque sia vengono eliminati dei valori che sarebbero potuti essere molto importanti.

Il secondo metodo è quello di **imputazione manuale**: è fatta da umani e tramite osservazioni ci si chiede se sia possibile inserirlo, è tremendamente difficile dal punto di vista computazionale.

Il terzo metodo è quello della **global constant**: ossia metto un numero là chiamato place holder con un valore costante, non troppo efficiente.

Il quarto metodo è quello di **rimpiazzarlo con la moda**, anche questo però è fortemente criticabile. Se gli attributi sono continui si fa la stessa cosa ma con la media.

Il quinto metodo è **Conditional mean replacement** ossia bisogna rimpiazzare con la media solo se è presente un altro determinato attributo.

Il sesto metodo è quello del **most probable** ossia di prendere un modello e sostituire il valore.

E' molto difficile dare il confine tra l'esplorazione dei dati e la modellizzazione dei dati.

4.4 Data Preprocessing

Tutti questi processi impattano fortemente tutta l'analisi che farò dopo.