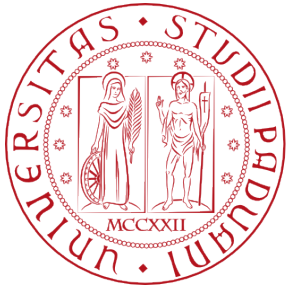
Statistical learning methods for classification and profiling

AGCom Workshop on the impact of online platforms on information freedom and media pluralism: Fake News and Other regulatory challenges

Antonio Canale

March 17, 2017



Università degli Studi di Padova

1. Machine learning e Statistical learning
2. Classificazione statistica
3. Dal modello di classificazione all'algoritmo di classificazione
4. Un esempio: Google AdWords
5. Spunti di discussione
   - Distorsioni e preconcetti (bias in classification)
   - Trasparenza
   - Influenzare il futuro

Machine learning / Statistical learning

- Il Machine learning costruisce algoritmi che possono imparare dai dati

Machine learning / Statistical learning

- Il Machine learning costruisce algoritmi che possono imparare dai dati
- Statistical learning è quella parte della statistica applicata che emerge come risposta al machine learning, enfatizzando i modelli statistici e la valutazione dell'incertezza

Machine learning / Statistical learning

- Il Machine learning costruisce algoritmi che possono imparare dai dati
- Statistical learning è quella parte della statistica applicata che emerge come risposta al machine learning, enfatizzando i modelli statistici e la valutazione dell'incertezza
- Data science è l'estrazione di conoscenza dai dati, usando idee dalla matematica, statistica, machine learning, informatica, ingegneria, fisica, ...

Scopi simili, focus diversi

- Gli scopi sono i medesimi: descrivere trend e andamenti, capire il presente, prevedere il futuro

Scopi simili, focus diversi

- Gli scopi sono i medesimi: descrivere trend e andamenti, capire il presente, prevedere il futuro
- Per il machine learning l'algoritmo e il suo output è centrale

Scopi simili, focus diversi

- Gli scopi sono i medesimi: descrivere trend e andamenti, capire il presente, prevedere il futuro
- Per il machine learning l'algoritmo e il suo output è centrale
- Per lo Statistical learning i modelli statistici e la quantificazione dell'incertezza sono centrali

Classificazione (in ottica previsionale)

Facciamo un esempio

- Una società di telecomunicazioni vuole prevedere i clienti che nel prossimo mese disdiranno il loro contratto per passare alla concorrenza (churn analysis)

Classificazione (in ottica previsionale)

Facciamo un esempio

- Una società di telecomunicazioni vuole prevedere i clienti che nel prossimo mese disdiranno il loro contratto per passare alla concorrenza (churn analysis)
- Con classificazione si intende l'attribuzione dell'etichetta `lascio/rimango` (o `churn/not churn`)

Classificazione (in ottica previsionale)

Facciamo un esempio

- Una società di telecomunicazioni vuole prevedere i clienti che nel prossimo mese disdiranno il loro contratto per passare alla concorrenza (churn analysis)
- Con classificazione si intende l'attribuzione dell'etichetta `lascio/rimango` (o `churn/not churn`)
- Lo scopo è prevedere il churn di domani oggi

Classificazione (in ottica previsionale)

Facciamo un esempio

- Una società di telecomunicazioni vuole prevedere i clienti che nel prossimo mese disdiranno il loro contratto per passare alla concorrenza (churn analysis)
- Con classificazione si intende l'attribuzione dell'etichetta `lascio/rimango` (o `churn/not churn`)
- Lo scopo è prevedere il churn di domani oggi
- Vogliamo "etichettare" il futuro con usando le informazioni note fino ad oggi

Un po' di terminologia...

- L'etichetta `lascio/rimango` è chiamata variabile risposta o target

Un po' di terminologia...

- L'etichetta `lascio/rimango` è chiamata variabile risposta o target
- Le informazioni note fino ad oggi possono essere rappresentate da variabili numeriche (n. connessioni alla rete, n. chiamate, durata chiamate, età del cliente ...) o variabili qualitative (genere del cliente, piano tariffario, handset...)

Un po' di terminologia...

- L'etichetta `lascio/rimango` è chiamata variabile risposta o target
- Le informazioni note fino ad oggi possono essere rappresentate da variabili numeriche (n. connessioni alla rete, n. chiamate, durata chiamate, età del cliente ...) o variabili qualitative (genere del cliente, piano tariffario, handset...)
- Queste vengono chiamate variabili esplicative o variabili di input

Un po' di terminologia...

- L'etichetta `lascio/rimango` è chiamata variabile risposta o target
- Le informazioni note fino ad oggi possono essere rappresentate da variabili numeriche (n. connessioni alla rete, n. chiamate, durata chiamate, età del cliente ...) o variabili qualitative (genere del cliente, piano tariffario, handset...)
- Queste vengono chiamate variabili esplicative o variabili di input
- E' evidente che il target non è definito perfettamente dagli input

Un po' di terminologia...

- L'etichetta `lascio/rimango` è chiamata variabile risposta o target
- Le informazioni note fino ad oggi possono essere rappresentate da variabili numeriche (n. connessioni alla rete, n. chiamate, durata chiamate, età del cliente ...) o variabili qualitative (genere del cliente, piano tariffario, handset...)
- Queste vengono chiamate variabili esplicative o variabili di input
- E' evidente che il target non è definito perfettamente dagli input
- C'è una componente casuale/aleatoria

Classificazione statistica

- Un modello statistico prevede

$$Pr(\texttt{target} = \texttt{churn} \mid \texttt{input}) = f(\texttt{input})$$

Classificazione statistica

- Un modello statistico prevede

$$Pr(\texttt{target} = \texttt{churn} \mid \texttt{input}) = f(\texttt{input})$$

- è evidente che anche se conoscessimo in maniera esatta la probabilità non è detto che sapremo con certezza l'etichetta

Classificazione statistica

- Un modello statistico prevede

$$Pr(\texttt{target} = \texttt{churn} \mid \texttt{input}) = f(\texttt{input})$$

- è evidente che anche se conoscessimo in maniera esatta la probabilità non è detto che sapremo con certezza l'etichetta
- esempio: so che la probabilità di testa è 50% ma non so se uscirà testa o croce!

Classificazione statistica

- Un modello statistico prevede

$$Pr(\texttt{target} = \texttt{churn} \mid \texttt{input}) = f(\texttt{input})$$

- è evidente che anche se conoscessimo in maniera esatta la probabilità non è detto che sapremo con certezza l'etichetta
- esempio: so che la probabilità di testa è 50% ma non so se uscirà testa o croce!
- è ragionevole pensare che etichetterò il tal cliente come futuro abbandono se la sua $Pr(\texttt{target} = \texttt{churn})$ è maggiore del 50%

Apprendimento statistico (statistical learning)

- in ogni caso non sappiamo come è fatta $f()$

Apprendimento statistico (statistical learning)

- in ogni caso non sappiamo come è fatta $f()$
- dobbiamo stimare f usando le informazioni a nostra disposizione ovvero imparare il meccanismo con cui dagli input f ricaviamo l'output

Apprendimento statistico (statistical learning)

- in ogni caso non sappiamo come è fatta $f()$
- dobbiamo stimare f usando le informazioni a nostra disposizione ovvero imparare il meccanismo con cui dagli input f ricaviamo l'output
- si tratta di arrivare a una stima di $f()$ che chiameremo $\hat{f}()$

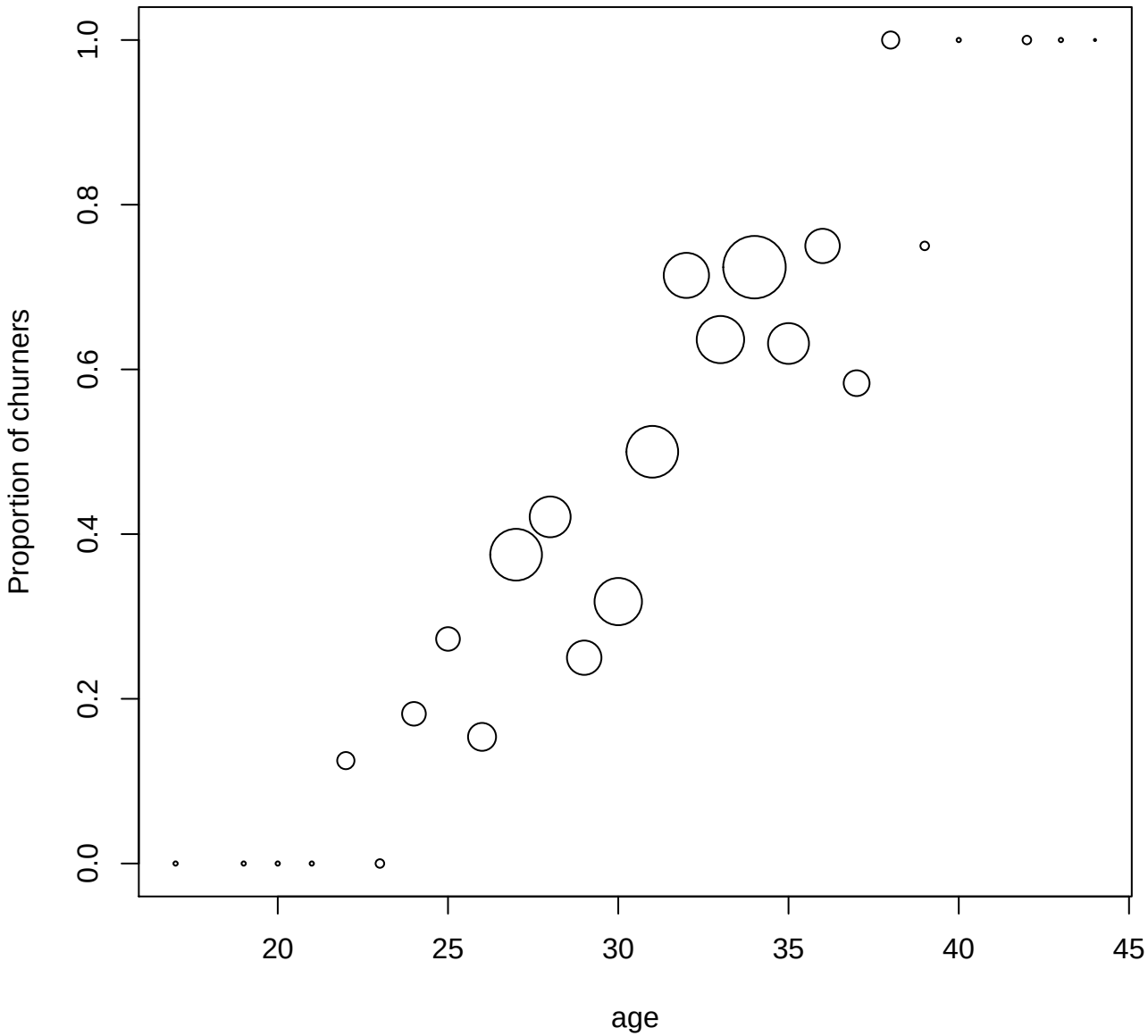
Due fonti di incertezza

- non conosciamo $f()$ ma solo una sua stima $\hat{f}()$

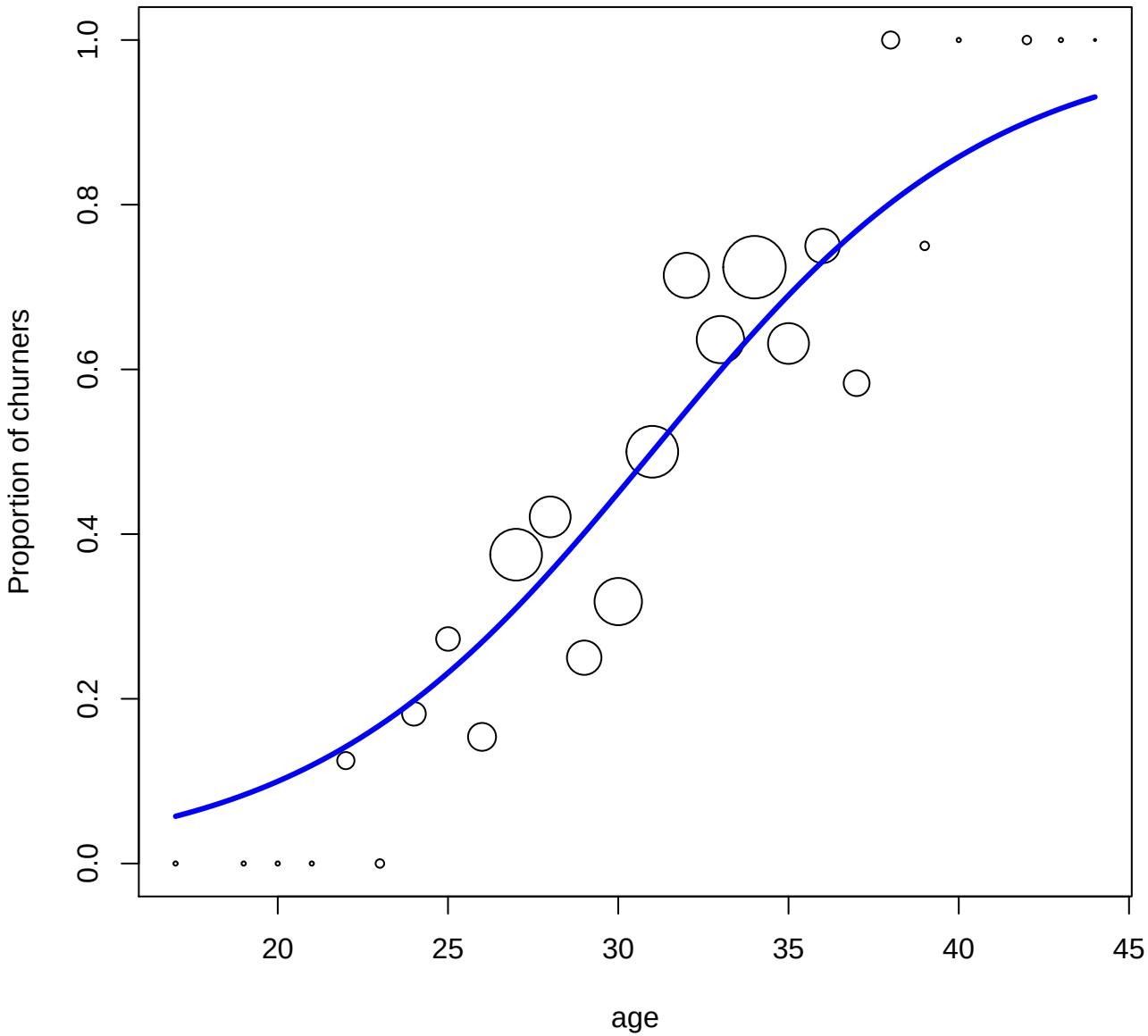
Due fonti di incertezza

- non conosciamo $f()$ ma solo una sua stima $\hat{f}()$
- in ogni caso $\hat{f}()$ ci fornisce una probabilità

Caso semplice



Caso semplice



Algoritmo di classificazione

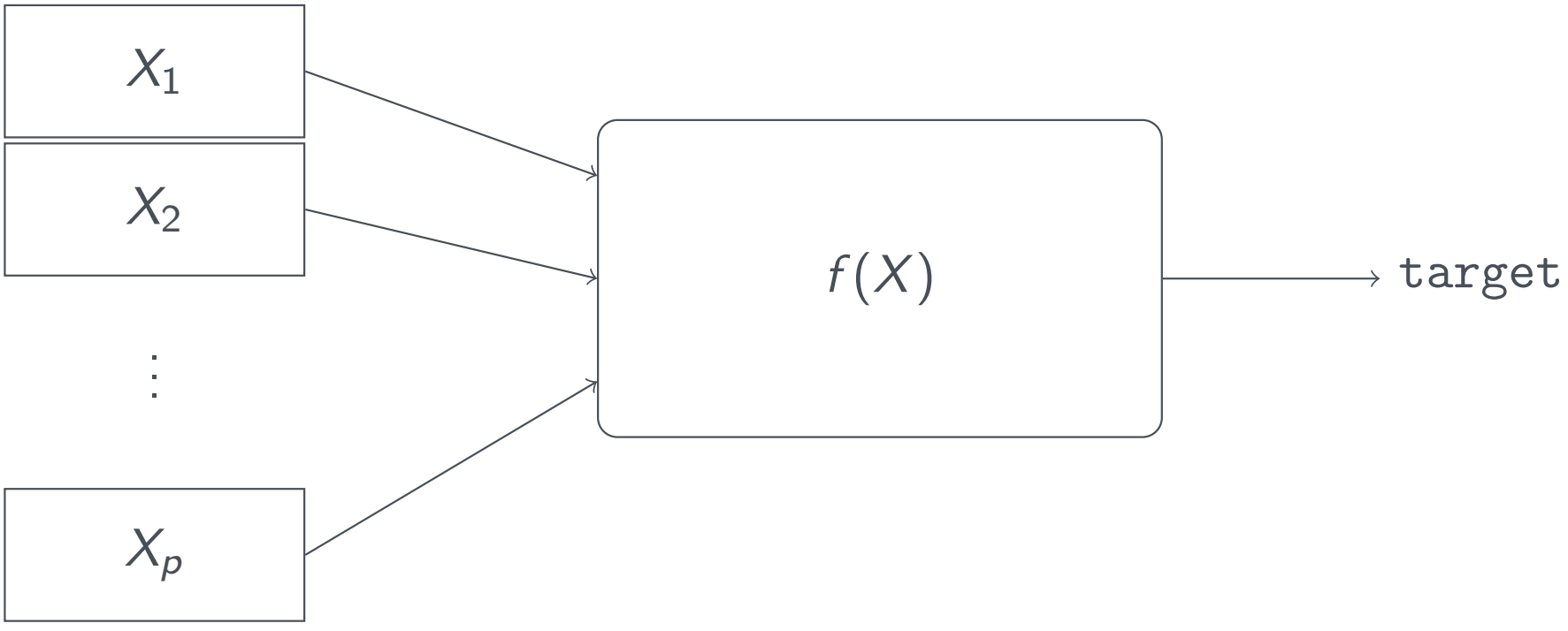
1. Decido che famiglia di f considerare
2. Uso lo storico per stimare $\hat{f}$ secondo alcuni criteri che minimizzano una qualche misura dell'errore di previsione
3. Applico $\hat{f}$ a nuovi dati di cui conosco solamente gli `input` per avere una probabilità
4. Classifico come `churn` se questa probabilità è maggiore del 50%

Algoritmo di classificazione

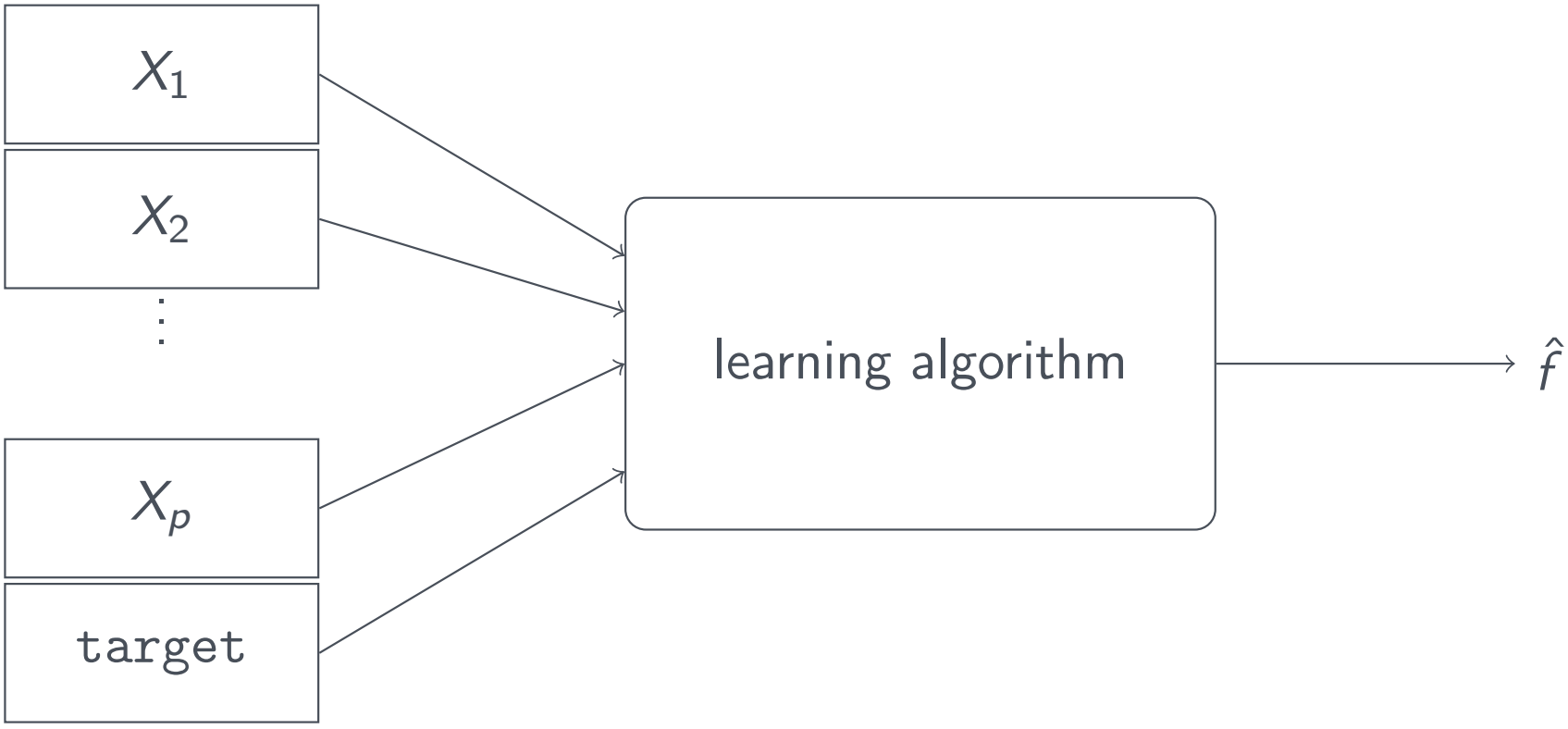
1. Decido che famiglia di f considerare
2. Uso lo storico per stimare $\hat{f}$ secondo alcuni criteri che minimizzano una qualche misura dell'errore di previsione
3. Applico $\hat{f}$ a nuovi dati di cui conosco solamente gli `input` per avere una probabilità
4. Classifico come `churn` se questa probabilità è maggiore del 50%

Operazione automatica - Decisione umana

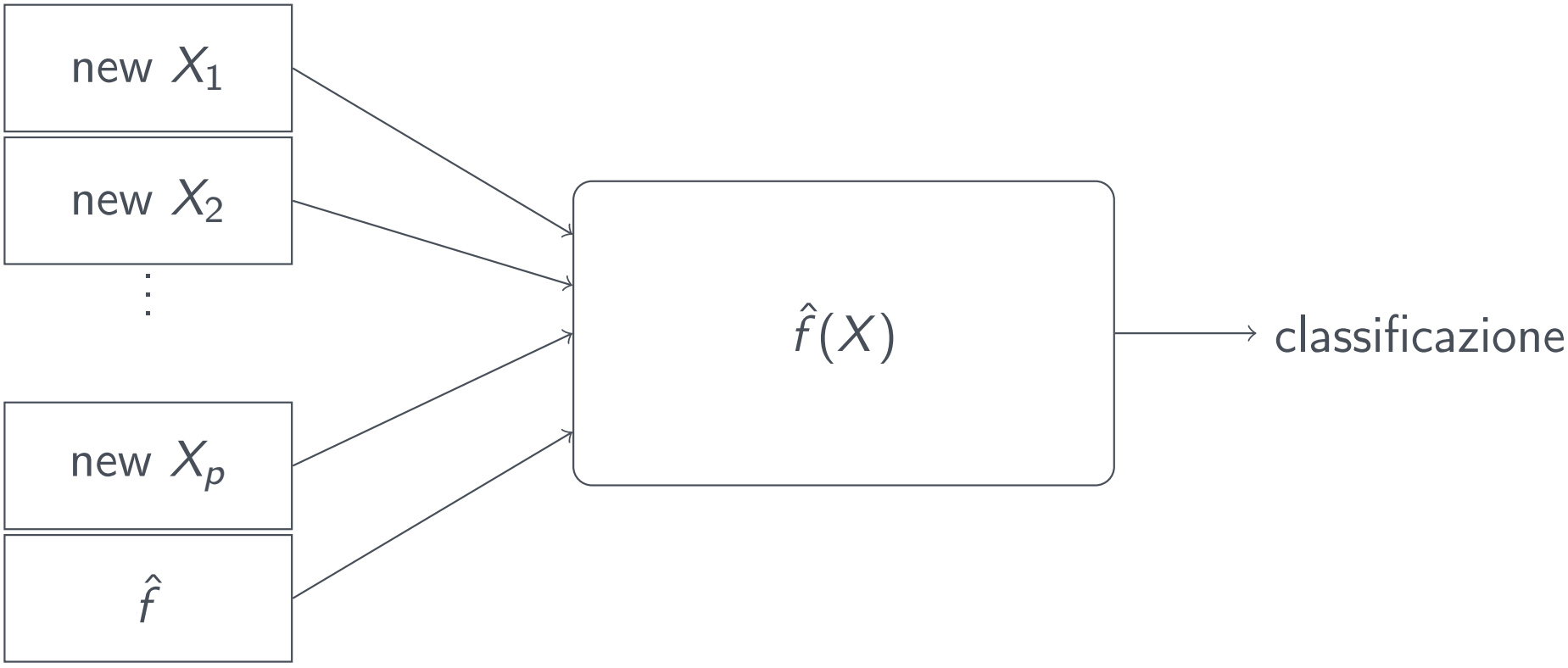
Riassumendo: modello semplificato "vero"



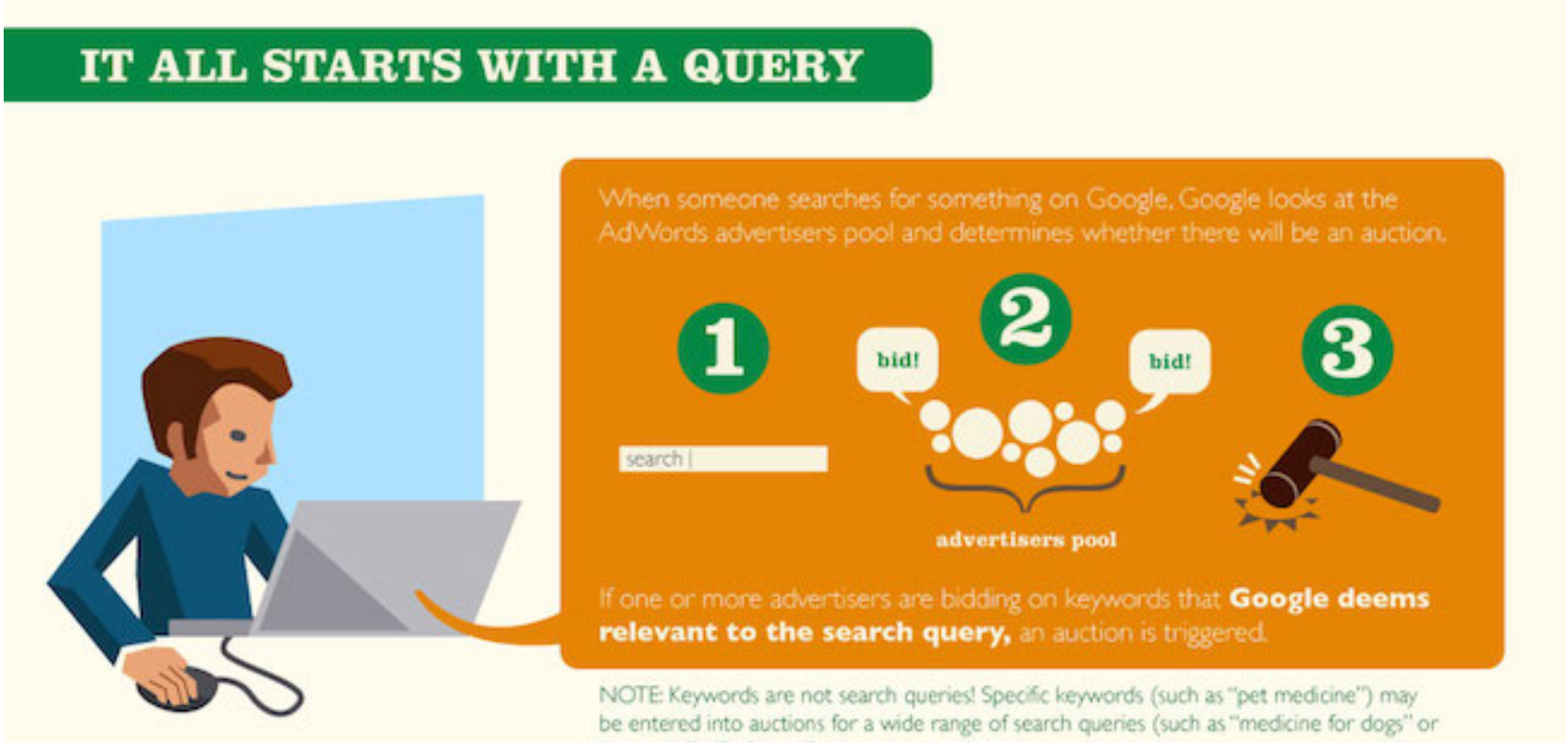
Riassumendo: stima



Riassumendo: previsione/classificazione



Google AdWords



Fonte: hubspot.com

Google AdWords

WHAT GETS ENTERED INTO THESE AUCTIONS?

Advertisers **identify keywords they want to bid on,** how much they want to spend, and create groupings of these keywords that are paired with ads.

Google then enters the keyword from your account it deems **most relevant** into the auction with the maximum bid you've specified as well as the associated ad.

NOTE: You can only have one entry into any query auction from your account.

Fonte: hubspot.com

Google AdWords

Fonte: hubspot.com

Google AdWords: ricapitolando

- L'utente fa una ricerca su Google

Google AdWords: ricapitolando

- L'utente fa una ricerca su Google
- Google associa alla ricerca delle keyword

Google AdWords: ricapitolando

- L'utente fa una ricerca su Google
- Google associa alla ricerca delle keyword
- Google, per ogni cliente che ha delle offerte su quella keyword associa un indice (quality score) sulla rilevanza dell'inserzione sulla query

Google AdWords: ricapitolando

- L'utente fa una ricerca su Google
- Google associa alla ricerca delle keyword
- Google, per ogni cliente che ha delle offerte su quella keyword associa un indice (quality score) sulla rilevanza dell'inserzione sulla query
- Google calcola l'AdRank

$$\texttt{max bid} \times \texttt{quality score} = \texttt{AdRank}$$

e restituisce i risultati

Google AdWords: keyword e classificazione

- Google associa alla ricerca delle keyword

Google AdWords: keyword e classificazione

- Google associa alla ricerca delle keyword
- La query costituisce l'input (e.g. "apple")

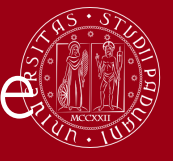
Google AdWords: keyword e classificazione

- Google associa alla ricerca delle keyword
- La query costituisce l'input (e.g. "apple")
- le keyword sono i possibili argomenti, ad esempio: `frutta fresca`, `computer Apple`, ...

Google AdWords: keyword e classificazione

- Google associa alla ricerca delle keyword
- La query costituisce l'input (e.g. "apple")
- le keyword sono i possibili argomenti, ad esempio: `frutta fresca`, `computer Apple`, ...
- nello scegliere le keyword Google cerca di prevedere la probabilità con cui l'utente intende un certo significato

Google AdWords: keyword e classificazione



- Google associa alla ricerca delle keyword
- La query costituisce l'input (e.g. "apple")
- le keyword sono i possibili argomenti, ad esempio: `frutta fresca`, `computer Apple`, ...
- nello scegliere le keyword Google cerca di prevedere la probabilità con cui l'utente intende un certo significato
- è un problema di classificazione

Bias in classification

- l'algoritmo di apprendimento dipende dall'uomo non solo per la scelta del tipo di modello ma anche per quanto riguarda i dati con cui è "allenato"

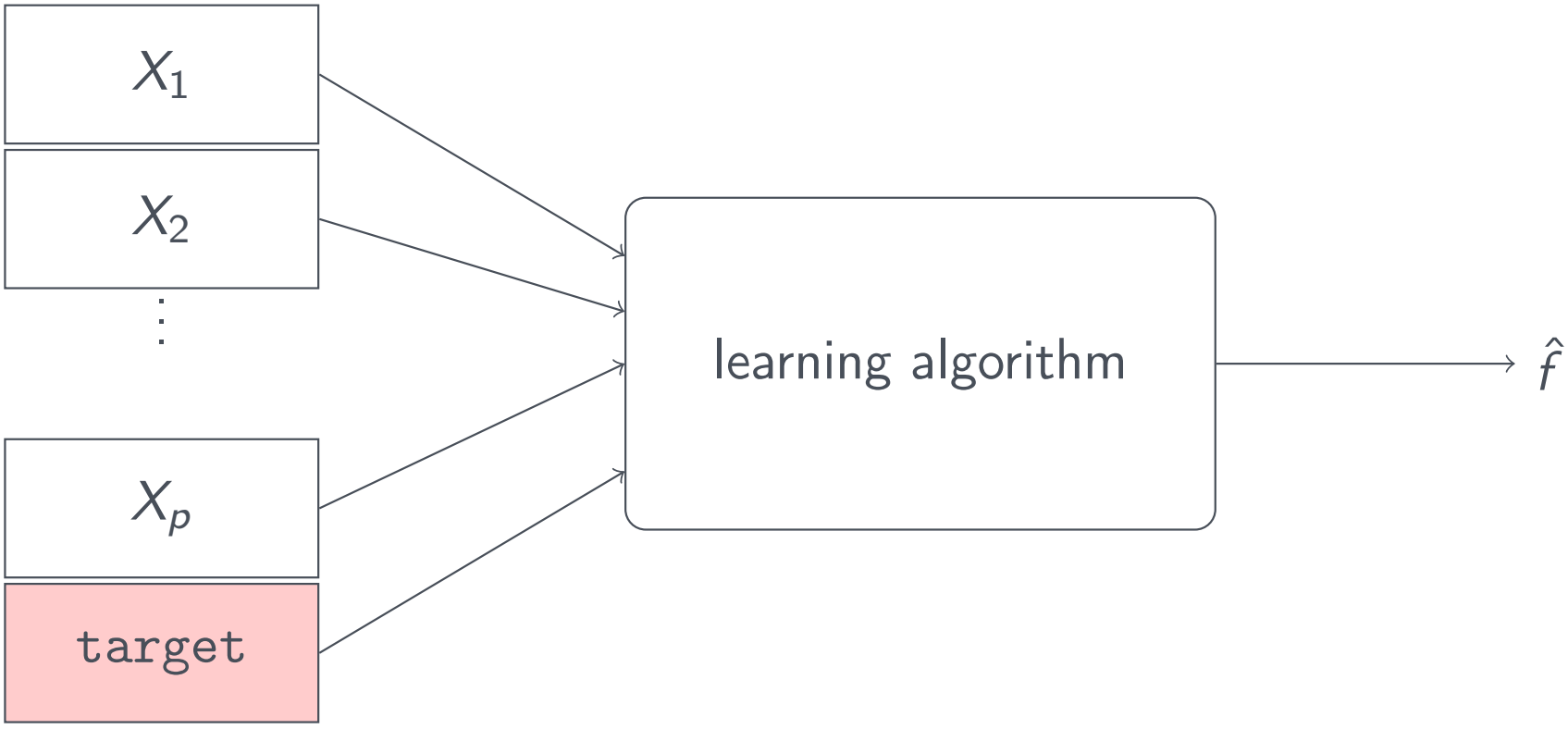
Bias in classification

- l'algoritmo di apprendimento dipende dall'uomo non solo per la scelta del tipo di modello ma anche per quanto riguarda i dati con cui è "allenato"
- la categorizzazione distorta delle etichette da parte dell'uomo porta inevitabilmente in una distorsione del modello di classificazione

Bias in classification

- l'algoritmo di apprendimento dipende dall'uomo non solo per la scelta del tipo di modello ma anche per quanto riguarda i dati con cui è "allenato"
- la categorizzazione distorta delle etichette da parte dell'uomo porta inevitabilmente in una distorsione del modello di classificazione
- e.g. nella classificazione delle immagini con contenuti espliciti, la classificazione dipende da fattori culturali.

Fase di stima



Trasparenza

- la scelta della famiglia di modelli, dei parametri di settaggio e di tutti i dettagli tecnici sono tipicamente non note negli algorimi di classificazione degli operatori digitali

Trasparenza

- la scelta della famiglia di modelli, dei parametri di settaggio e di tutti i dettagli tecnici sono tipicamente non note negli algorimi di classificazione degli operatori digitali
- Google ha una policy di trasparenza sui dati di input (ci dice cosa usa di noi per profilarci) ma non come ci profila

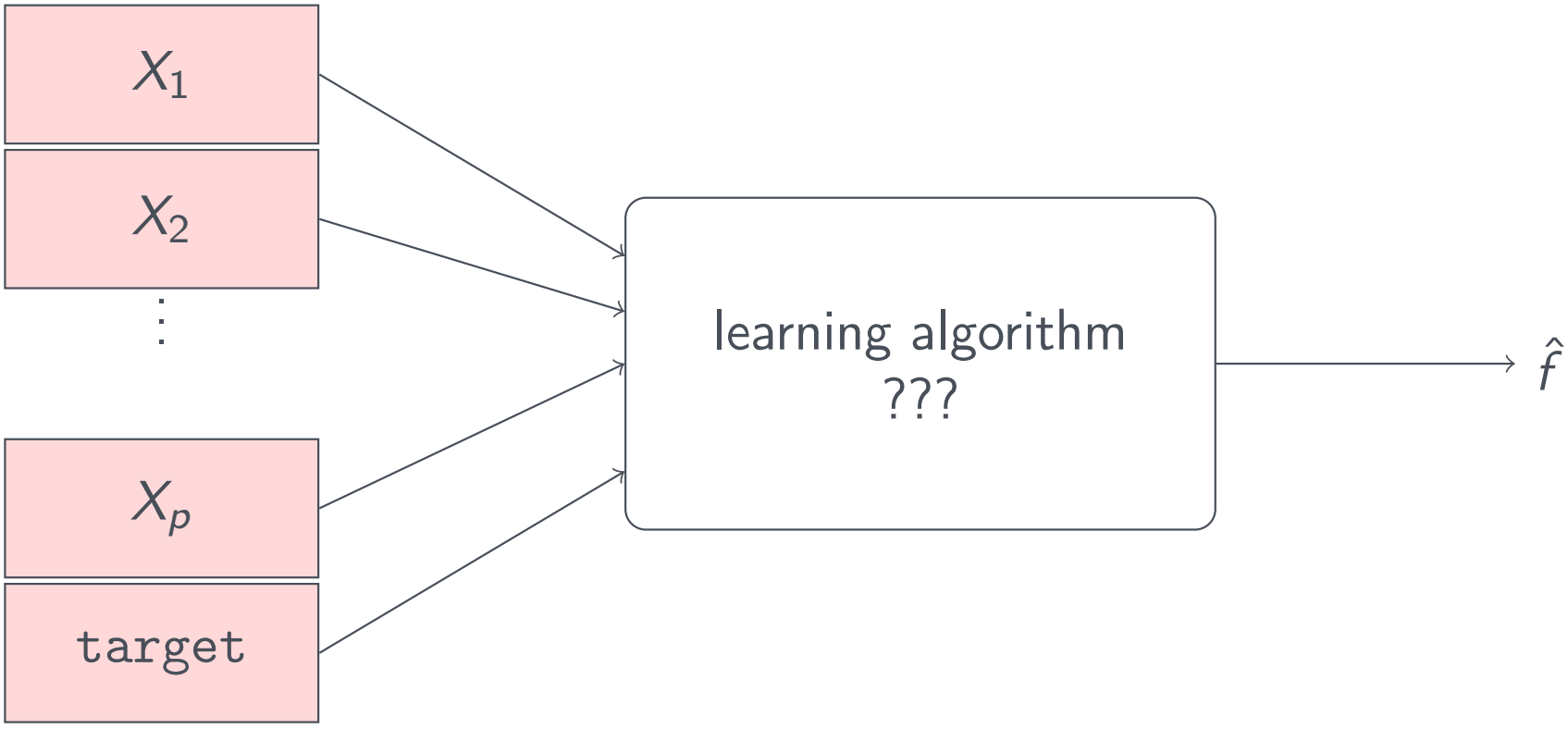
Trasparenza

- la scelta della famiglia di modelli, dei parametri di settaggio e di tutti i dettagli tecnici sono tipicamente non note negli algorimi di classificazione degli operatori digitali
- Google ha una policy di trasparenza sui dati di input (ci dice cosa usa di noi per profilarci) ma non come ci profila
- evidentemente migliore è il modello (e pertanto l'algoritmo di classificazione) migliore è la performance commerciale (si pensi di nuovo a Google AdWords)

Trasparenza

- la scelta della famiglia di modelli, dei parametri di settaggio e di tutti i dettagli tecnici sono tipicamente non note negli algorimi di classificazione degli operatori digitali
- Google ha una policy di trasparenza sui dati di input (ci dice cosa usa di noi per profilarci) ma non come ci profila
- evidentemente migliore è il modello (e pertanto l'algoritmo di classificazione) migliore è la performance commerciale (si pensi di nuovo a Google AdWords)
- tutto questo fa si che i dettegli rimangano segreto industriale

Trasparenza



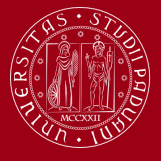
Classificazione di contenuti: le fake news

- A Gennaio Google ha rimosso da AdSense siti che riportavano fake news (fonte: recode.net, gennaio 2017)

Classificazione di contenuti: le fake news

- A Gennaio Google ha rimosso da AdSense siti che riportavano fake news (fonte: recode.net, gennaio 2017)
- Facebook sta aggiungendo filtri e algoritmi per mettere in risalto i contenuti più genuini (fonte: ilpost.it, febbraio 2017)

Classificazione di contenuti: le fake news



- A Gennaio Google ha rimosso da AdSense siti che riportavano fake news (fonte: recode.net, gennaio 2017)
- Facebook sta aggiungendo filtri e algoritmi per mettere in risalto i contenuti più genuini (fonte: ilpost.it, febbraio 2017)
- In ogni caso si tratta di implementazione di strumenti di classificazione

Influenzare il futuro

- Le previsioni, influenzano il futuro?

Influenzare il futuro

- Le previsioni, influenzano il futuro?
- La storia di Edipo, Laio e l'oracolo di Delfi

Influenzare il futuro

- Le previsioni, influenzano il futuro?
- La storia di Edipo, Laio e l'oracolo di Delfi
- Di fatto ogni previsione influenza il futuro!

Influenzare il futuro

Ipotizziamo di operare in un mercato azionario

- Un nostro modello (statistico?) prevede che il prezzo dell'azione A domani si alzerà di molto

Influenzare il futuro

Ipotizziamo di operare in un mercato azionario

- Un nostro modello (statistico?) prevede che il prezzo dell'azione A domani si alzerà di molto
- In virtù di questa previsione saremo portati a comprare un elevato numero di quelle azioni

Influenzare il futuro

Ipotizziamo di operare in un mercato azionario

- Un nostro modello (statistico?) prevede che il prezzo dell'azione A domani si alzerà di molto
- In virtù di questa previsione saremo portati a comprare un elevato numero di quelle azioni
- Se tutti fanno cosi, il prezzo si alzerà

Influenzare il futuro

Ipotizziamo di operare in un mercato azionario

- Un nostro modello (statistico?) prevede che il prezzo dell'azione A domani si alzerà di molto
- In virtù di questa previsione saremo portati a comprare un elevato numero di quelle azioni
- Se tutti fanno cosi, il prezzo si alzerà
- Il prezzo che si alza è pertanto causa o effetto?

Grazie per l'attenzione

Antonio Canale
Email: canale@stat.unipd.it
Twitter: tonycanale_