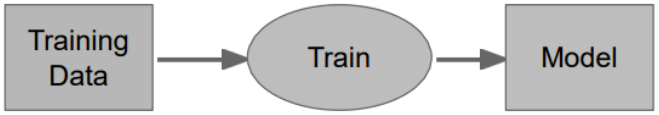
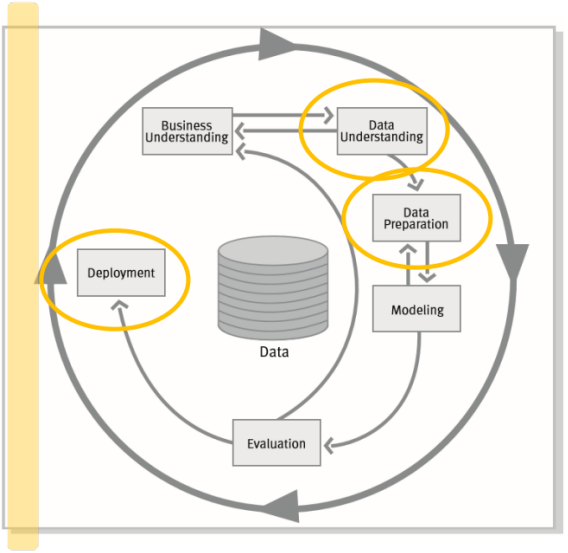
Lezione 8 10/04/2024

Use case: data management for machine learning

A machine learning avevamo i dati di training, facevamo il training e costruivamo il modello.



Quale potrebbe essere il ciclo di vita, il modello di riferimento, quando si utilizzano sistemi di machine learning?



Questo è il CRISP-DM (data mining) che parte dalla comprensione del problema, poi c'è la fase di data understanding, perché per risolvere il problema ho bisogno dei dati. Se prima i dati ci venivano già forniti, nel mondo reale i dati non ci sono ancora.

Poi c'è la data preparation, dove i dati vengono rielaborati, ai fini di considerare le features più identificative. Magari i dati vengono uniti, puliti, in generale preparati.

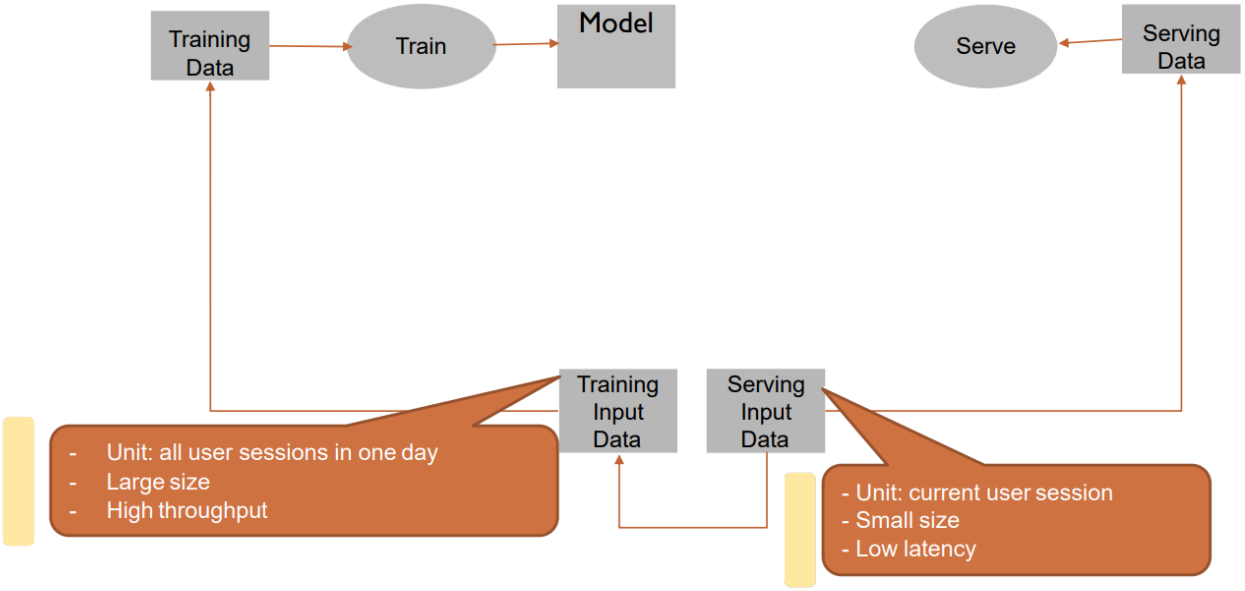
Nel modeling invece viene applicato il modello di machine learning. Infine c'è la fase di valutazione e deployment.

In questo use case vediamo le parti cerchiate in giallo.

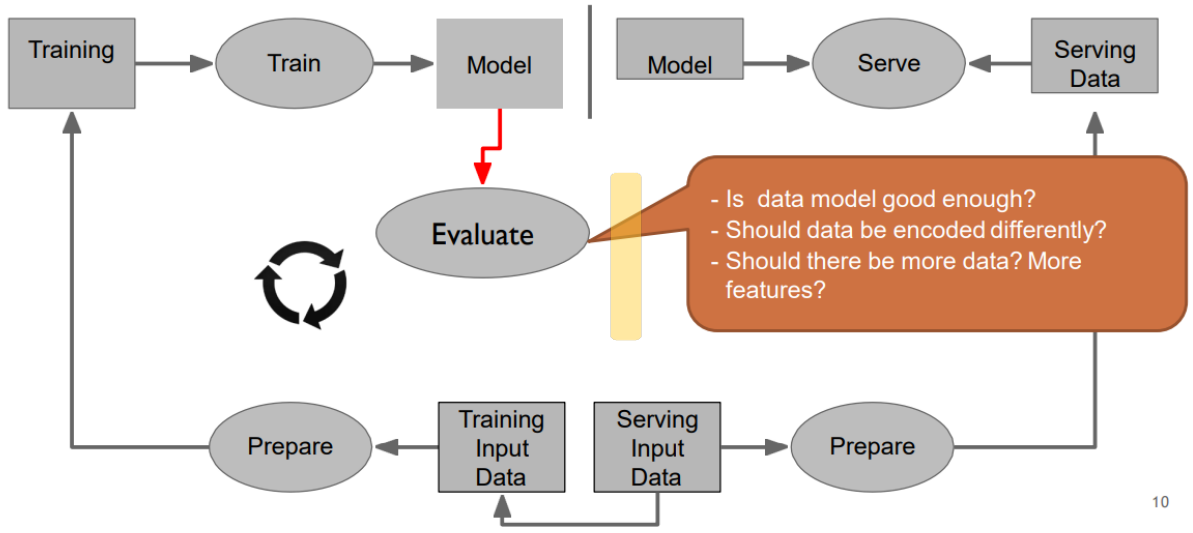
Dal punto di vista dei dati prima si definisce il modello, e poi il modello riceve dei dati reali, e applicando il modello andrà a classificarli.

Dal punto di vista dei dati, questi devono essere giusti. A volte i dati sono in mano all'azienda, e dati esterni non sono applicabili al problema (per esempio se netflix deve capire quali utenti potrebbero disiscriversi, se uso quelli di spotify il comportamento dell'utente è diverso e quindi non funziona). I dati vanno quindi preparati per renderli utilizzabili.

Sia i dati di training che quelli di serving nascono dallo stesso gruppo di dati, vengono divisi tra i due.



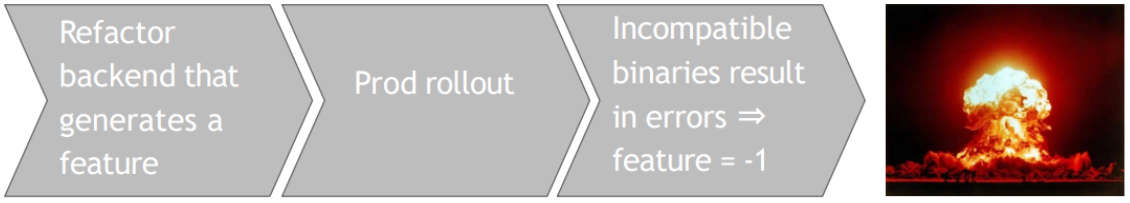
I dati di training sono più grandi, nel caso di netflix avrei quindi tutte le sessioni dell'utente. Mentre quelli di serving sono di meno, e contengono la sessione attuale dell'utente per valutarla con il modello. Qui è fondamentale la latenza, perchè la creazione del modello può anche richiedere giorni, ma l'esecuzione deve essere nell'ordine dei millisecondi.



I dati vanno quindi preparati in entrambe le direzioni, costruendo i dati in modo corretto e trovando le features più interessanti, questo poi dipende anche dall'algoritmo, alcuni preferiscono alcune features, altri preferiscono altre.

Infine il modello va validato, per vedere se è un buon modello, se è migliorabile cambiando la rappresentazione dei dati, se i dati vanno codificati in modo diverso, se vanno aggiunti più dati, più features.

Example of data failure



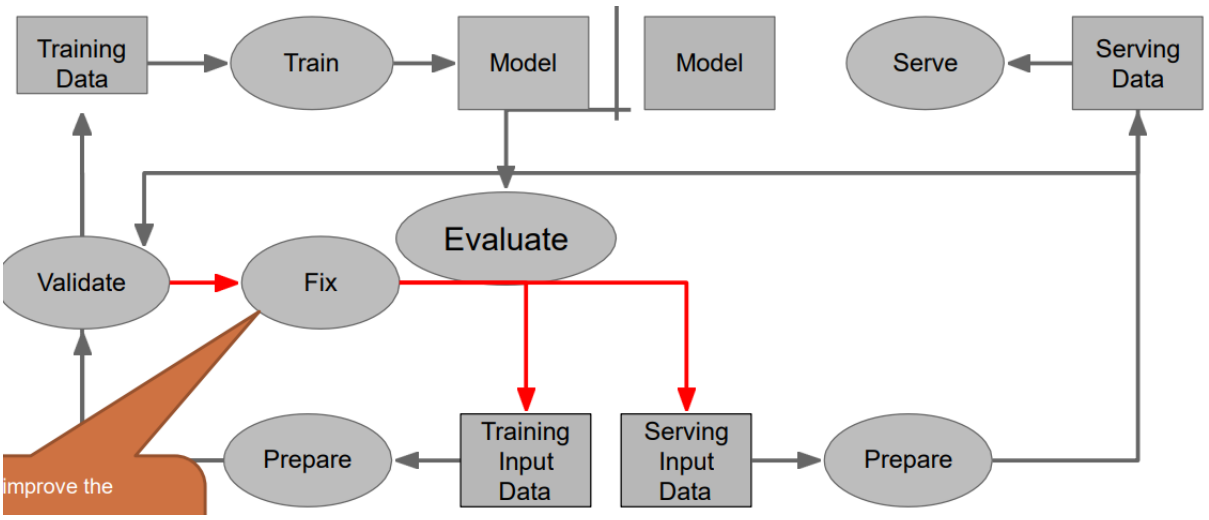
Per esempio, I dati potrebbero arrivare da un altro sistema che crea un csv. Può succedere che il sistema che crea i dati viene modificato e quindi crea un file csv diverso, che quindi non è più compatibile con il modello.

Per limitare e prevenire questi eventi, i dati in input vanno validati. Bisogna pensare a quali proprietà dei dati hanno un effetto significativo sulla qualità del modello. Fare questa cosa è complicato, perché i dati che cambiano sono quelli nuovi che vanno sul serving, bisogna vedere se questi dati nuovi possono ancora essere utilizzati con il sistema vecchio o no.

Quindi il problema è anche capire che i dati sono cambiati. Bisogna avere degli alert per capire che c'è qualcosa che non va, e poi bisogna decidere se non

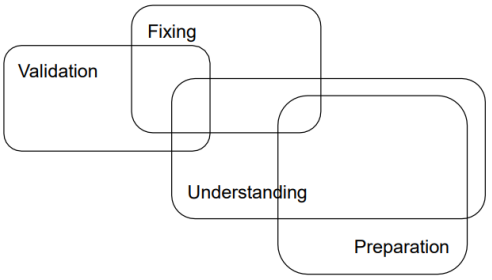
fare nulla, se fare qualcosa nell'immediato o se ad un certo punto ri-addestrare il modello.

I dati possono essere sistemati nel fix, per sistemare anche solo dei problemi dove semplicemente la formattazione viene cambiata.



Quindi questi sono i pezzi fondamentali delle attività di fare sui dati, senza parlare del modello. Questa parte è ancora più importante dell'addestramento stesso.

La realtà dei fatti è che la pulizia, l'ottimizzazione, l'organizzazione dei dati occupa il 80% del tempo.

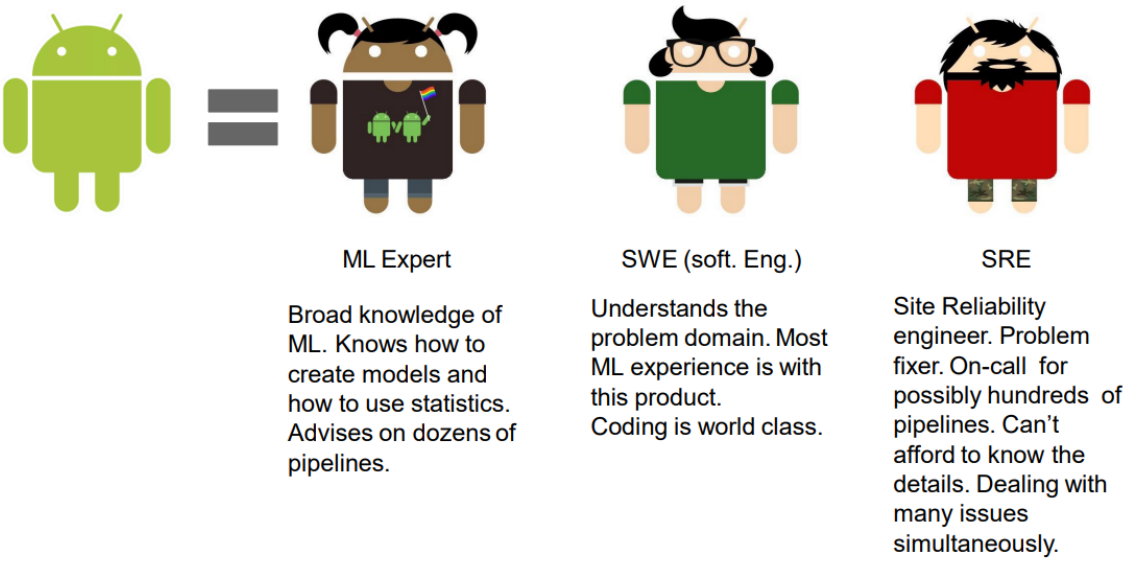


Data acquisition

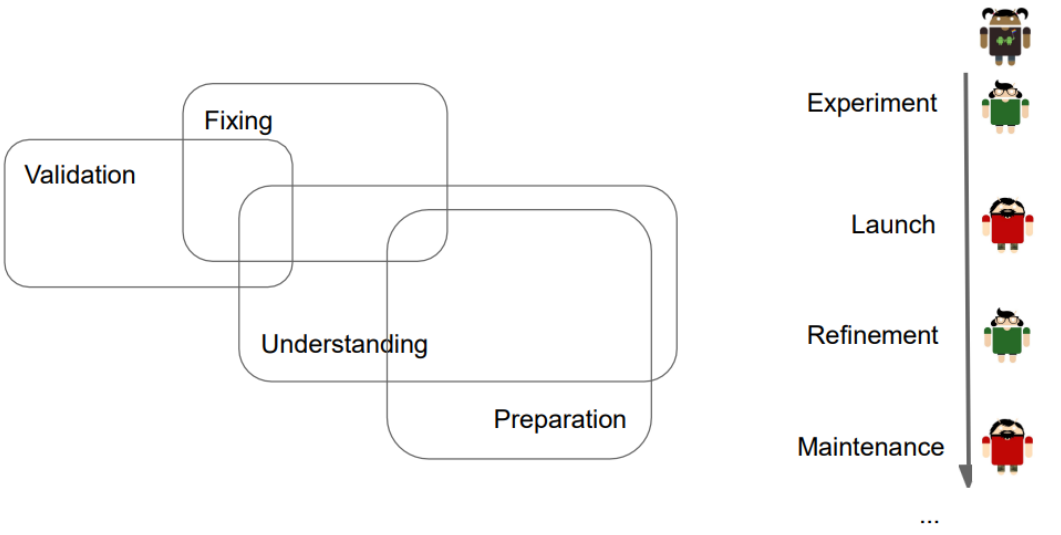
Gli approcci cambiano da do it yourself a software as a service. Cioè al posto di fare tutto da zero, usiamo dei software che ci assistono. Il trade off è dato dal tempo, il costo, la facilità di utilizzo, il controllo sui dettagli.

Il primo aspetto del data acquisition è il bias, è inserito nei dati. I dataset che contengono bias possono essere inappropriati per il training. Per esempio un dataset delle telecamere dell'università che riprende persone bianche, non riesce a riconoscere le persone asiatiche. Quindi il bias può essere già all'interno dei dati, da dove vengono raccolti.

Diventa quindi importante capire la provenienza dei dati, spiegando anche come e perché vengono presi.



Serve anche una nuova figura, oltre al software engineer e il machine learning expert. Il site reliability engineer si occupa dell'efficienza, della velocità del modello.



Data understanding

è lo step iniziale dell'analisi dei dati. é utile farla graficamente, per rendere i dati più immediati e chiari. Si possono fare analisi univariate o multivariate, scoprendo le relazioni tra le variabili, le distribuzioni, la struttura del dataset.

Serve fare data understanding all'inizio, ma anche di continuo, perché i dati cambiano. Continuamente bisogna verificare che i dati nuovi sono simili a quelli iniziali, perché se c'è un data drift significativo questo può essere un problema.