1. Abstract

Al giorno d’oggi i dati hanno un ruolo sempre più centrale nella società; gran parte di questi sono dati sensibili e un loro diretto utilizzo potrebbe portare a dei problemi per questioni di privacy, ma un loro utilizzo è fondamentale per andare a capire nel profondo e prevedere fenomeni sociali e scientifici. Una soluzione frequentemente utilizzata in ambito statistico è quella di generare dati sintetici che preservino le proprietà dei dati originali, senza però andare ad intaccarne la riservatezza.

Il nostro progetto si basa su metodi di statistica bayesiana nonparametrica (BNP), come per esempio simulazione dati da Polya Tree.

1. Step 1

Non avendo un dataset, per prima cosa abbiamo generato dei dati da una distribuzione nota tra [0,1] come una beta o un’uniforme.

Presa la nostra prior come un Polya Tree di profondità m=10 (con parametri iniziali alpha inizialmente tutti posti pari a c\*m^2, con c fissato) ed essendo il Polya Tree un modello coniugato, tramite quei dati abbiamo aggiornato i parametri del nostro PT.

Dopodiché abbiamo fatto sampling da tale posterior, estraendo 10 distribuzioni per ogni valore di c (5, 100, 1000), ripetendo il tutto sia con dati iniziali campionati da un’uniforme su [1/4,3/4] che da una Beta(2,2)

Predictive: ragionando e provando a calcolare la predittiva, abbiamo pensato che fare sampling da lì non fosse altro che fare sampling da una distribuzione multinomiale sui 2^m possibili intervalli creati.

