Anonymous Learning via Look-Alike Clustering: A Precise Analysis of Model Generalization

Adel Javanmard (University of Southern California)
Vahab Mirrokni (Google Research)

lookalike clustering for anonymous learning and model generalization

- ***Problem:*** We are given a supervised learning task and would like to protect a set of sensitive features during the training phase.

- ***Questions:***
  - How to protect privacy, while still making personalized prediction?
  - What is the measure of privacy?
  - Is privacy protection in conflict with model generalization?
  - If yes, how does this trade-off shape under different problem parameters? (e.g., overparameterization, signal-to-noise-ratio, data quality, etc)

Model

- Linear regression:

$$Y = x_s^T \theta_s + x_{ns}^T \theta_{ns} + z$$

Sensitive features

non-Sensitive features

noise

SNR = (strength of θ_s)/(noise std) $= \frac{||\theta_s||}{\sigma}$

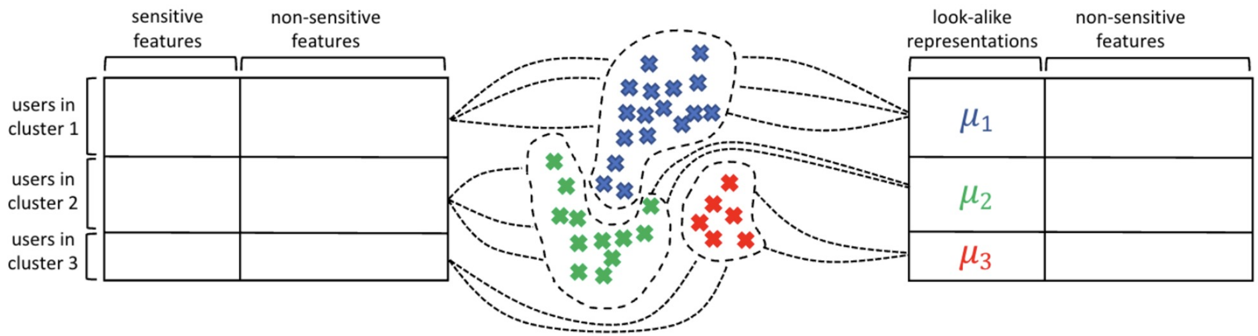
- n: size of training data,
- p: dimension of sensitive features
- $d - p$: dimension of non-sensitive features

We focus on high-dimensional asymptotics, where the size of training data, number of sensitive/ insensitive features grow in proportion.

$(\frac{d}{n} \to \Psi_d, \frac{p}{n} \to \Psi_p,$ as $n \to \infty)$

Lookalike clustering for anonymous learning and model generalization

- ***Our approach:*** We follow a natural technique called `look-alike clustering’



1. Cluster users based on non-private information
2. Within each cluster, replace users’ sensitive features with a common representation (center of cluster)

Privacy measure? We obtain k-anonymity on sensitive features if min size cluster is at least k.

Privacy- model generalization tradeoff

Common belief: protecting users privacy conflicts with model generalization

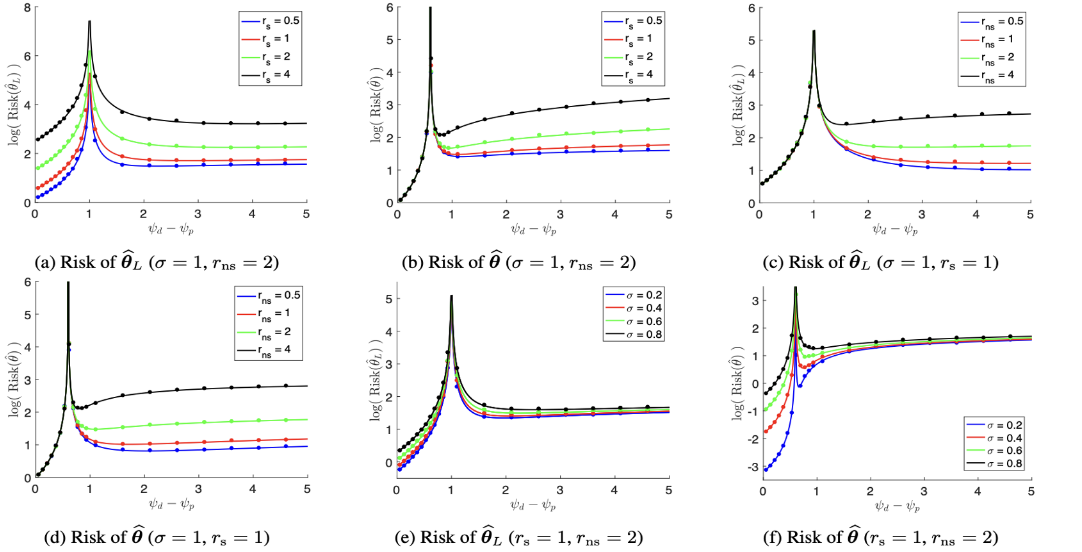
- Modern deep NNets have remarkable generalization property, and they often perform in overparameterized regime (memorize/interpolate training data)
- Similar behavior is observed for random forests, Adaboost, kernel methods
- For mixture of subpopulation data, [Feldman 20] shows label memorization is necessary for optimal generalization, under long-tailed distribution
- It is beyond label memorization: [Brown et. al 21] studies setting where optimal generalization requires memorizing high-entropy / high-dimensional covariates information

We present a different picture for anonymity via lookalike modeling!

Precise characterization of model generalization

➢ We provide a precise characterization of model generalization, using techniques from Convex-Gaussian minimax-theorem (CGMT) [*Thrampoulidis et a. 2015*]

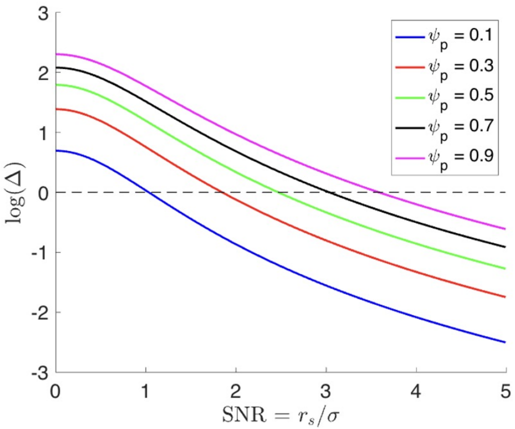
➢ Our theoretical analysis allows to understand role of different parameters (e.g. size/number of clusters, cluster separation, overparameterization, SNR) on model generalization.

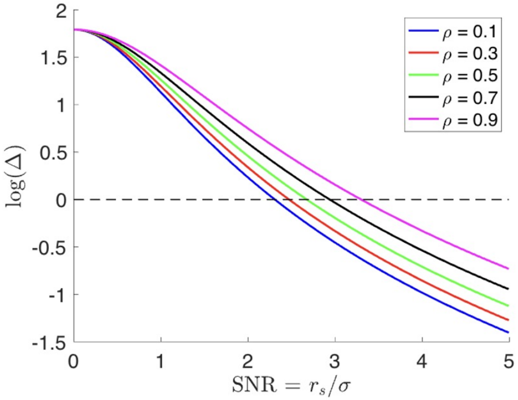


(a) Risk of $\widehat{\boldsymbol{\theta}}_L$ ($\sigma = 1, r_{\rm ns} = 2$)
(b) Risk of $\widehat{\boldsymbol{\theta}}$ ($\sigma = 1, r_{\rm ns} = 2$)
(c) Risk of $\widehat{\boldsymbol{\theta}}_L$ ($\sigma = 1, r_{\rm s} = 1$)
(d) Risk of $\widehat{\boldsymbol{\theta}}$ ($\sigma = 1, r_{\rm s} = 1$)
(e) Risk of $\widehat{\boldsymbol{\theta}}_L$ ($r_{\rm s} = 1, r_{\rm ns} = 2$)
(f) Risk of $\widehat{\boldsymbol{\theta}}$ ($r_{\rm s} = 1, r_{\rm ns} = 2$)

When does lookalike clustering improve model generalization?

Δ = (generalization error of non-lookalike) / (generalization error of lookalike)





SNR = (strength of θ_s)/(noise std) $= \frac{||\theta_s||}{\sigma}$

In low SNR look-alike-clustering <u>improves generalization</u>, in addition to k-anonymity.

Intuition on better generalization in low SNR?

- At low SNR, noise is comparable with the heterogeneity within cluster.
- By replacing sensitive features with cluster center, look-alike clustering acts as a regularization to avoid overfitting.