RMIA? NOTES

The attack principle of RMIA (Robust Membership Inference Attack) is based on statistical reasoning and contrastive learning. Its goal is to determine whether the target is a member of the training set by statistically estimating the behavioral differences between the target sample and the reference non-member sample.

Principle:

1. Hypothesis Testing Statistical Inference
   Binary hypothesis: Hin/Hout (member/non-member)
   The goal is to determine which hypothetical world x belongs to.

2. Comparative framework: Comparison based on sample pairs (x vs. z)
   Instead of judging x individually, x is compared with each non-member sample z~π in the reference set, and its membership is inferred by "relative advantage".

3. Scoring function Score_MIA(x)

   $$\mathrm{LR}_\theta(x, z) = \frac{\Pr(\theta|x)}{\Pr(\theta|z)},$$

   This probability is essentially how many non-member samples z outperform x in pairwise comparisons.

4. Likelihood Ratio (LR)
   Pr(x|θ): softmax/logits output of target model θ (e.g. true class probability)
   Pr(z|θ): corresponding value of reference non-member samples
   The scoring function is a rank-based comparison indicator.

5. Attack Signals, The Core Features includes:
   Softmax: The softmax prediction of the true categories
   Softmax_relative: top1 - top2 difference
   Taylor_softmax: Taylor expansion approximation of softmax
   Margin: top1 - top2 logits difference
   Cross_entropy: Loss function value
   Entropy: Prediction Distribution Entropy

6. Reference Models
   Usually by using multiple shadow models, and excluding the x while training.
   To estimate the background distribution of non-member behaviors, multi-model averaging is performed (to improve robustness);
   Use multiple reference models or multiple data augmentation versions.

7. Attack Evaluation Metrics

TPR / FPR Curve
ROC Curve
AUC
TPR(The attack can find many members) @ low FPR(Don't hurt non-members) (TPR high -> good, FPR low -> good) ()

The core principle of RMIA is to compare the target sample with multiple non-member samples in pairs and calculate their advantage ratios as the basis for membership inference.

RMIA, or Robust Membership Inference Attack, is a technique used to figure out whether a specific data point was part of a model's training set. Instead of looking at that sample in isolation, RMIA compares it to a bunch of reference samples that are definitely not in the training set—usually taken from the test set or some holdout data. For each comparison, it checks how confident the model is about the target sample versus the reference ones, using things like softmax probabilities or logit margins. Then, it calculates how often the target sample looks more like a training sample than the references. If this "win rate" is high enough—above some threshold—the sample is likely a training member. The whole process is based on hypothesis testing and doesn't require any access to the model's internals, just its predictions.