

Separability

Using classification models to evaluate the usefulness of privatized data without losing simplicity, explainability or generality.

Executive Summary

Metric Overview

Real World Use Case

Metric Definition

Formal Metric Definition

Explanation of Metric Parameters

Snapshot and Deep Dive Modes

Snapshot

Deep Dive

Metric Defense

Discriminative Power

Description of Coverage

Scalability / Feasibility

Generalizability

Theoretical Background

VC Dimension

PAC Learning and Empirical Risk Minimization

0-1 loss function

Theoretical Properties

Separability is between 0 and 1

Separability bounds the change in behavior on private and ground truth

Executive Summary

Metric Overview

Separability combines two key ideas:

Idea 1: If a classifier can effectively distinguish between ground truth data and privatized data, we show it is possible to use privatized data to draw inaccurate conclusions.

Idea 2: The strength of classifiers should be measured with respect to the complexity of the function class used to build them.

By normalizing the accuracy of a classifier differentiating between private and ground truth datasets by the accuracy of a classifier attempting to fit noise, we can uncover structural differences in privatization algorithms without worrying about the relative ability of a given classifier to overfit. While this approach generalizes to any family of classifiers, in this submission, we are promoting the use of **Logistic Regression** and **Fixed-Depth Decision Trees**.

We see four key benefits to separability under logistic regression.

First, the metric is **easy to interpret**. If a privatized data release has high separability, then the difference in error of classification models from that function class with respect to our privatized and ground truth datasets can also be high.

Next, the metric is **simple to implement**. Separability relies on concepts that data scientists are familiar with — namely binary classification. By building our classifiers with simple explainable models, the metric can be implemented using standard data science packages. There are no new concepts to learn, no new packages to install and no specialized hardware needed.

```
from sklearn.linear_model import LogisticRegression
import pandas as pd
import random

def compute_separability(df1, df2):
    ground_truth['CLASS'] = 'Ground Truth'
    privatized['CLASS'] = 'Privatized'

    combined = pd.concat([ground_truth, privatized])

    X = combined.drop(columns=['CLASS'])
    y = combined['CLASS']
    clf = LogisticRegression().fit(X, y)

    y_random = y.copy().values
    random.shuffle(y_random)
    clf_random = LogisticRegression().fit(X, y_random)
    return clf.score(X, y) / clf_random.score(X, y_random) - 1
```

Third, separability is **explainable**. Separability benefits from the explainability of the chosen function class. If a privatized dataset has high separability under logistic regression or decision trees, data scientists can rely on well-known and standard techniques to uncover underlying issues. For example, later in this document, we show how analyzing coefficients enabled us to diagnose a structural problem in the privatized data for Maryland police calls.

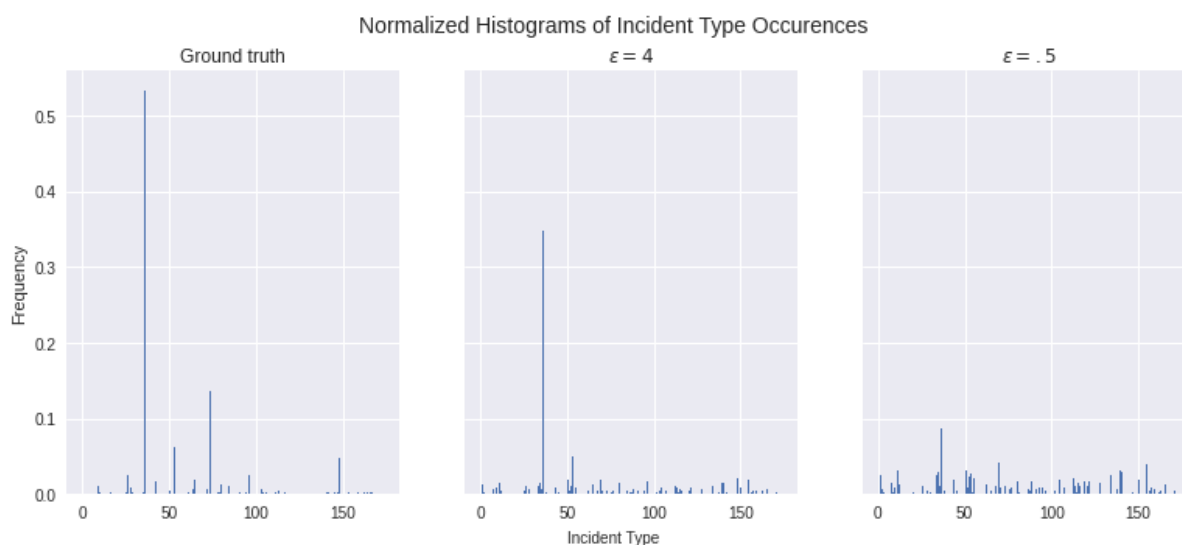
Finally, separability offers a **theoretically robust and modular framework**. While this submission suggests the use of logistic regression and decision trees, these results also apply to random forests or any other classification architecture. In fact, for function families that can solve the XOR problem (like decision trees), separability provides a strict upper bound on the difference in error on ground truth and privatized datasets, thereby providing a guarantee of data utility.

Real World Use Case

In applying our separability metric under logistic classification to the privatized Baltimore 911 Call and Police Incident Data provided in the sample competition pack, we find that the privatization methods meaningfully change the structure of the algorithm's output.

The data consists of 1.5 million 911 calls, along with the respective incident type, neighborhood, time, and caller ID for each call. The algorithm outputs of interest are histograms of incident types at the neighborhood-month level.

Below is a plot of the normalized histograms for one neighborhood-month, with the ground truth on the left, and privatized versions in the middle and on the right. A quick visual inspection shows that the ground truth and privatized histograms are distinguishable. Smaller values of ϵ , a parameter in the Laplace mechanism, correspond to higher magnitudes of noise and stronger privacy guarantees. Unsurprisingly, adding more noise makes the histogram look more obviously different from the ground truth.



Based on the significant visual differences in the histograms, we should expect our metric to assign this privatized output a bad score. In our case, this means that the separability of our ground truth and privatized datasets should be very high.

As hoped, the separability of this dataset is incredibly high at .8. In our Snapshot and Deep Dive section we analyze the coefficients of our classifier and its results on specific chunks of the dataset to determine why separability is high, and provide recommendations for what steps would need to be taken for this score to improve.

Metric Definition

Formal Metric Definition

We consider a setting in which an algorithm is run on a given dataset. For example, the output of the algorithm may be a table of summary statistics or a histogram of the outcomes. Denoting the ground truth algorithm result by D and its privatized version by $P = M(D)$, we propose the following procedure:

1. Combine D and P into one larger dataset X
2. Label rows of X as 1 if the row is from D and 0 if the row is from P , creating target vector Y
3. Train a classifier C on X to predict Y and compute the accuracy of C on X
4. Randomly shuffle Y to obtain Y_{random}
5. Follow the same procedure as in step 3 to train a classifier C_{random} on X to predict Y_{random} and compute the accuracy of C_{random}
6. Return $SEP(D, P) = \text{Accuracy of } C / \text{Accuracy of } C_{random} - 1$

$SEP(D, P)$ is always a number between 0 and 1, and lower numbers mean that the privatized data release is a better representation of your data.



We provide a more detailed proof of why separability is between 0 and 1 as well as the formal guarantees that separability provides later in the document

Explanation of Metric Parameters

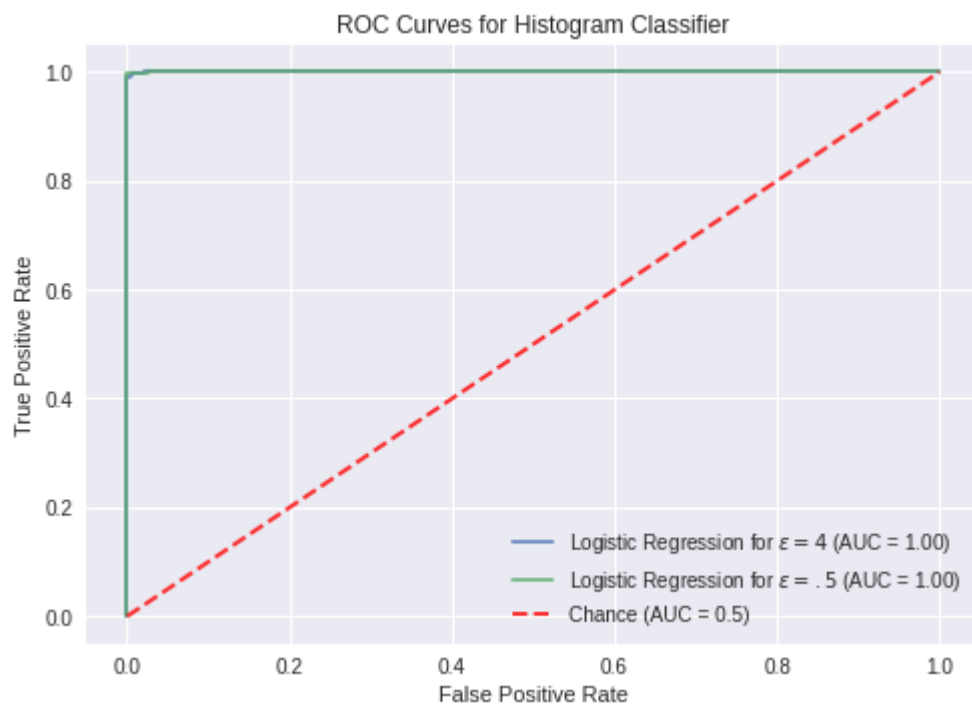
The one meaningful choice when implementing separability is the choice of classification functions. While we suggest the use of logistic regression or decision trees, this choice is entirely at the discretion of the model implementer.

Snapshot and Deep Dive Modes

Separability can be used to create a snapshot of the overall utility of a privatized data release as well as do a deep dive into particular subsets of the data. We explore these properties through a more detailed analysis of the Baltimore 911 Call and Police Incident Data.

Snapshot

There are 174 types of incidents, so histograms are 174-dimensional vectors. Labeling the ground truth histograms with 1 and the privatized histograms with 0, we fit a logistic regression model.

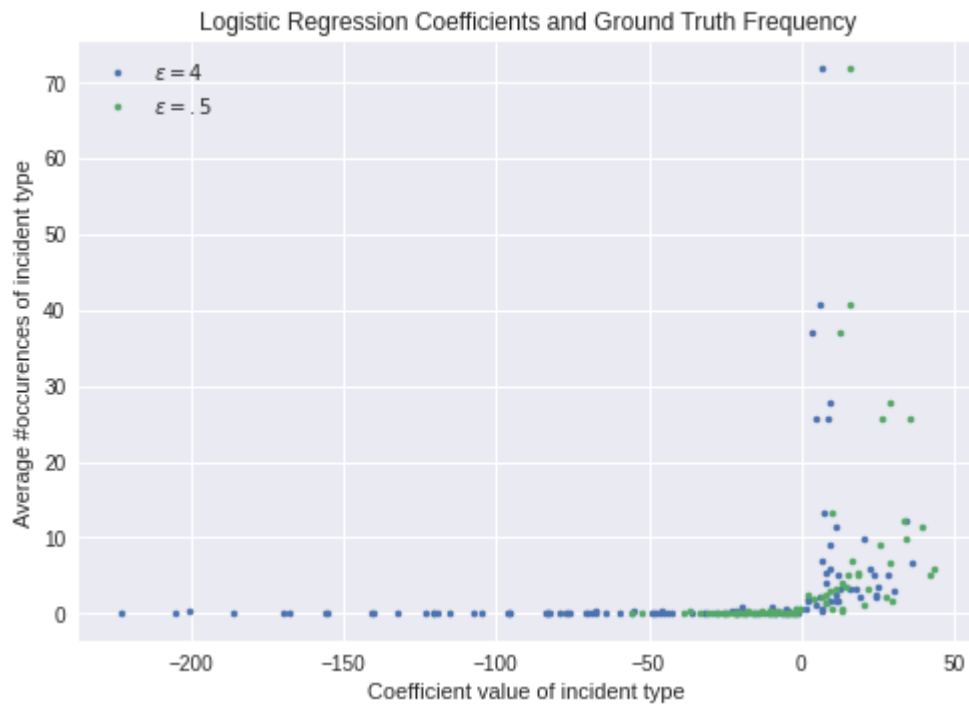


While our classifier perfectly distinguishes between the ground truth and privatized histograms, as can be seen in the receiver operating characteristic (ROC) curves above, it is not able to effectively split our randomized dataset with accuracy at just 55%. This gives us a separability of .8. This high separability score suggests that the provided privatized histograms should **not** be used for downstream analysis, aligning with our intuition.

We believe that sparsity is a major driver of these discrepancies. Adding noise to sparse outputs makes the output no longer sparse and makes the privatized output noticeably different from the ground truth. In the use case, for example, the true histograms contained mostly zeros, but adding noise made the histograms not sparse. This issue can be avoided at the upstream algorithm design stage. Rather than using a vanilla Laplace mechanism, which does not preserve sparsity, one can use alternative differential privacy mechanisms that preserve sparsity. One such procedure for histograms is described [here](#). Sparsity-preserving procedures should perform better under our metric and generate results more appropriate for downstream use.

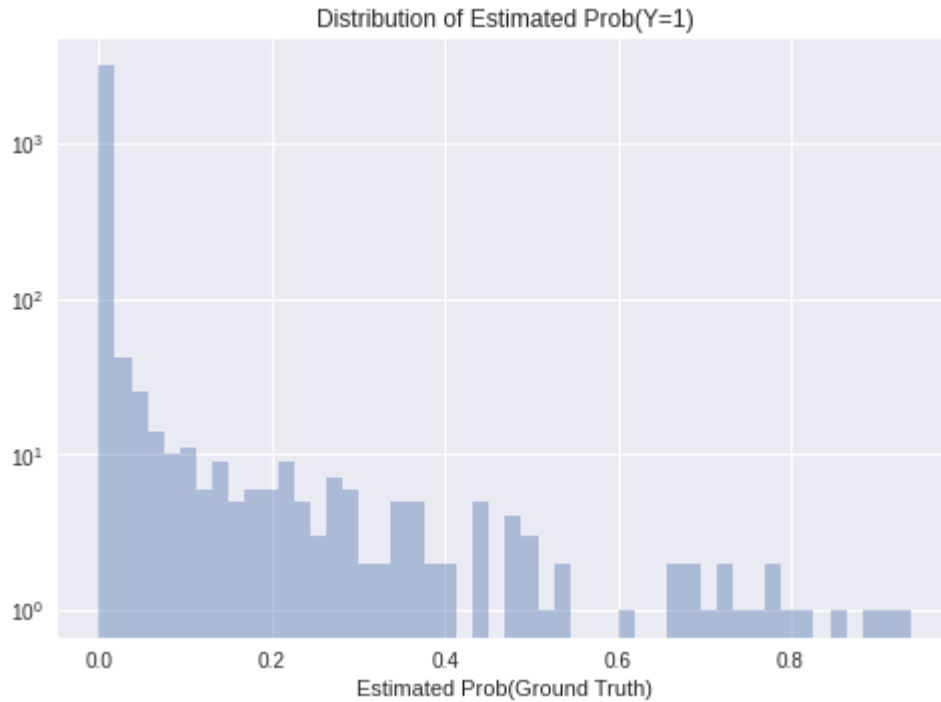
Deep Dive

Because we are using logistic classifiers, we can interpret our model to learn why it is so easy to distinguish between the ground truth and the privatized data at the dataset level. Looking at the model's parameters, we see that coefficients are negative for incident types that do not occur in the ground truth histogram and positive for incident types that occur more frequently. These negative coefficients mean that the presence of certain incident types is highly indicative of the histogram being privatized.



This shows that the significant number of calls in categories that do not exist in our ground truth dataset has meaningfully changed the conclusions that can be drawn from the privatized dataset.

The classifier yields further insights about the quality of the data at the neighborhood-month level. Given the high separability, evaluating the model for a privatized histogram gives the probability the classifier assigns to that histogram being from the ground truth (i.e. $\text{Prob}(Y=1)$). Higher probabilities are desirable and indicate that the privatized histogram looks like a ground truth. We compute these probabilities for each of the histograms in the $\epsilon = 4$ dataset and plot the distribution.



The estimated probabilities have a mean of 1.4% and a variance of 0.00567, which corresponds to a standard deviation of about 7.5%. For the vast majority of cases, the classifier is able to easily determine that the algorithm output is not from the ground truth.

We decompose the variance in the estimated probabilities to their intra-group and inter-group group components, where the groups are spatial (neighborhood) or temporal (month).

Group	Within Group	Between Group	Total Variance
Neighborhood	0.00145	0.00435	0.00567
Month	0.00568	1e-5	0.00567

Spatially, most of the variation is between neighborhoods. That is, some neighborhoods consistently have better estimated probabilities than others. Temporally, most of the variation is within-month, so seasonal trends in the estimated probabilities are not a primary concern. Rather, efforts should be directed toward understanding why some neighborhoods look more like the ground truth than others after privatization. This may be due to the randomness of the Laplace mechanism or due to systematic differences between the neighborhoods.

Metric Defense

Discriminative Power

One of the benefits of this metric is that the relative capabilities and drawbacks of the discriminative power of separability is identical to the relative capabilities and drawbacks of chosen method of binary classification. For example, while logistic regression effectively captures how changes in the magnitude of features impact the likelihood of being a particular class, it actively does not capture higher order effects like decision trees.

Description of Coverage

This metric is specifically designed to measure the impact of privatization on the accuracy of yes and no questions. Importantly, it does not provide corresponding bounds for regression problems.

Scalability / Feasibility

The training time for a logistic regression is $O(nd)$ and the training time for a fixed-depth decision tree is $O(n \times \log(n) \times d)$, where d is the number of dimensions of the dataset and n is the number of samples. Computing the Linear Inseparability on the provided Baltimore 911 Call and Police Incident Data for both logistic regression and decision trees completed within a few seconds on a Macbook Pro.

Generalizability

We believe that low scores on our metric are an important prerequisite to releasing privatized data. If a privatized data release has high separability, downstream data analysts can correctly use that information to draw inaccurate conclusions.


We recommend using separability both:

1. As a practical tool to check histograms before releasing to the public or to other researchers
2. As a research tool to analyze how differential privacy algorithms distort results

Importantly, **this metric is effective for any dataset that lends itself to binary classification tasks**. For example, researchers might want to determine...

- Which party a particular area voted for by analyzing privatized demographic information
- Whether or not a particular town has a high risk of health issues based on the existence of certain types of industry
- Whether a particular patient has a disease

Theoretical Background

 This section relies on some relatively complicated math as well as concepts from statistical learning theory. While these are in no way required to implement separability, they are required to understand its generalized error bounds.

VC Dimension

VC dimension is a concept from computational learning theory that bounds the "richness" of a function class F . For our uses, the only important result is if a function class has finite VC dimension, then as the number of samples with random labels in a dataset grows, the maximum accuracy of binary classifiers from F will approach .5.

The vast majority of functions we encounter (including logistic regression and fixed-depth decision trees) have finite VC dimension. We show below how separability behaves for function classes with finite VC dimension.

PAC Learning and Empirical Risk Minimization

PAC (Probably Approximately Correct) learning and empirical risk minimization are also fundamental concepts in computational learning theory. In our application, we will use the following adaptation of the definition of Agnostic PAC learnability: for a binary classification problem that is PAC-learnable, given a sufficient period of time, we can use empirical risk minimization to determine a classifier that is arbitrarily close to the optimal classifier, for arbitrarily high confidence levels.

The Fundamental Theorem of Statistical Learning Theory tells us that if a function class F has finite VC dimension, it is PAC learnable.



The combination of these two statements is the foundation of our subsequent work.

For function classes with finite VC dimension, we have a deterministic protocol that allows us to compute the optimal classifier to our level of desired accuracy and confidence.

0-1 loss function

For a classifier C , for a value x , the 0 – 1 loss function is 1 if C classifies x correctly and 0 otherwise. For a given dataset X , we define the loss $L_C(X)$ as the average loss of the 0 – 1 loss of C over X .

$$L_C(X) = \sum_{x \in X} \frac{L_C(x)}{|X|}$$

Theoretical Properties

Separability is between 0 and 1

Under this setup, the classifier $f(X) = 0$ will have accuracy .5. Therefore, without loss of generality, we can assume that the accuracy of C and C_{random} are bounded below by .5.

By the definition of accuracy, C and C_{random} are also bounded above by 1. This guarantees that $0 \leq SEP(D, P) \leq 1$.

Separability bounds the change in behavior on private and ground truth



This proof makes the assumption that if a classifier $C \in F$, then the loss function associated with it is in F as well. This result holds for function families that are closed under XOR, like decision trees.

This proof is quite technical! The key takeaway from it is the following:

For sufficiently large datasets and sufficiently complex function classes, the difference in loss of classifiers on our ground truth and privatized data is less than or equal to the separability.

For sufficiently large datasets, if we are using a function family F with finite VC dimension, the accuracy of our classifier C_{random} converges to .5.

Therefore, $SEP(D, P) \rightarrow 2 \times Accuracy(C) - 1$.

And if we assume that our function class has finite VC dimension, then we can assume that by using empirical risk minimization, the Fundamental Theorem of Statistical Learning tells us that $Accuracy(C) \rightarrow \sup_{C^* \in F} Accuracy(C^*)$.

Therefore we can say that to arbitrary levels of confidence

$$(1) \quad SEP(D, P) \rightarrow 2 \times \sup_{C^* \in F} Accuracy(C^*) - 1$$

Now consider a binary classifier C trained on our privatized dataset, and define $L_C(D)$ and $L_C(P)$ as value of the 0 – 1 loss functions on our ground truth and privatized datasets respectively. Define $\gamma = 2 \times \sup_{C^* \in F} Accuracy(C^*) - 1$, our theoretical separability.

By the definition of accuracy:

$$\max_{C \in F} Accuracy_C(D, M(D)) = \max_{f \in F} \frac{\sum_{x \in D} f(x) + \sum_{x \in D} 1 - f(M(x))}{2 \times |D|}$$

Simplifying

$$\max_{C \in F} Accuracy_C(D, M(D)) = \max_{f \in F} \frac{|D| + \sum_{x \in D} f(x) - f(M(x))}{2 \times |D|}$$

$$2 \max_{C \in F} Accuracy_C(D, M(D)) - 1 = \max_{f \in F} \frac{\sum_{x \in D} f(x) - f(M(x))}{|D|}$$

Replacing in for our definition of γ from (1)

$$(2) \quad \gamma = \max_{f \in F} \frac{\sum_{x \in D} f(x) - f(M(x))}{|D|}$$

Now, assume towards contradiction that there exists a classifier C such that

$$L_C(D) - L_C(P) > \gamma$$

Replacing in for the definition of L_C , this in turn implies

$$\frac{\sum_{x \in X} L_C(x) - L_P(M(x))}{|D|} > \gamma$$

However, this is a direct contradiction of our re-expression of γ as shown in (2), as L_C is a binary function in F . Therefore we are guaranteed that for all classifiers C , the difference in accuracy on the private and ground truth datasets is bounded by γ .