Adult Tabular Experiment: Information-Theoretic Fidelity Evaluation

Ground Truth vs. Simulated vs. Synthetic (SDV CTGAN/TVAE)

November 5, 2025

1 Overview

This experiment quantifies the *information fidelity* of synthetic and simulated tabular data against a ground-truth reference, using the UCI Adult/Census Income dataset. We evaluate three data sources for the same phenomenon:

1. GT: ground-truth rows drawn from the real dataset (held-out split).
2. SIM: simulated rows generated from feature-wise fitted distributions.
3. SYN: synthetic rows generated by a learned model (SDV's CTGAN or TVAE) fit on the training split of the real dataset.

All sources are evaluated in a *common numeric representation*: continuous features are standardized, and categorical features are one-hot encoded. Metrics include per-feature Jensen–Shannon (JS) divergence for continuous marginals, and global distributional tests: Maximum Mean Discrepancy (MMD) with an RBF kernel and a Classifier Two-Sample Test (C2ST) AUC.

2 Repository Layout and Execution

Key paths

```
common/                          # shared utilities (I/O, metrics, viz, seeding)
experiments/adult_tabular/
  run.py                         # CLI entry point
  data.py                        # download/clean/split + encoders (scaler, OHE)
  simulate.py                    # per-feature fitting + (optional) Gaussian copula
  synth_sdv.py                   # SDV CTGAN/TVAE training + sampling
  evaluate.py                    # JS, MMD, C2ST + figures
  figures/                       # output histograms (created at runtime)
  data_cache/                    # dataset cache (created at runtime)
  results.json                   # machine-readable results for a run
```

How to run (from repo root)

```
python -m experiments.adult_tabular.run \
  --seed 42 \
  --n_eval 5000 \
```

```
--sim_mode gaussian_copula \
--synth ctgan \
--epochs 300 \
--batch_size 500 \
--pac 10 \
--out results_ctgan.json
```

For TVAE, use `--synth tvae` (the `--pac` flag is ignored).

3 Data, Splits, and Preprocessing

Let $X \in \mathbb{R}^{n \times d}$ denote features and $Y \in \{0, 1\}$ the `income` label. The pipeline:

1. Download and clean the Adult dataset; drop rows with missing values.
2. Split into train/eval (e.g., 75/25, stratified by Y).
3. Fit preprocessing on the *train* split only:
   - StandardScaler on continuous columns: $x \mapsto (x - \mu)/\sigma$.
   - One-Hot Encoder on categorical columns
4. Transform the *eval* split using the fitted transformer; this provides the GT reference sample for metrics.

We evaluate only in the standardized + one-hot space; labels are not used by the core metrics.

4 Simulated Data (SIM)

4.1 Per-Feature Fitting

For each raw (unencoded) feature, we detect a plausible family and estimate parameters:

- **Binary**: Bernoulli(p) via $\hat{p} = \frac{1}{n} \sum \mathbf{1}\{x_i = 1\}$.
- **Categorical**: empirical multinomial over observed categories.
- **Counts** (nonnegative integers): Poisson(λ) if $\widehat{\mathrm{Var}} \approx \hat{\lambda}$, else Negative Binomial.
- **Continuous**: AIC-based model selection among Normal, Lognormal, Gamma, Exponential, Student-t Parameters are fit by maximum likelihood.

5 Synthetic Data (SYN)

We train a single-table synthesizer on the raw train split:

- **CTGAN**: Conditional GAN for tabular data with mode collapse mitigation. Important hyperparameters include epochs, batch size, and the "pack" size (*pac*), with the constraint that `batch_size % pac = 0`.
- **TVAE**: Variational autoencoder variant for tabular synthesis.

6 Evaluation Metrics

Let P, Q denote distributions over a common D-dimensional numeric space (standardized+one-hot). Let $\{x_i\}_{i=1}^n \sim P$ and $\{y_j\}_{j=1}^m \sim Q$ be i.i.d. samples.

6.1 Per-Feature Jensen–Shannon Divergence (Continuous Marginals)

For a single continuous feature X with histograms p and q (using the same bins and ranges), define the JS divergence

$$\mathrm{JS}(p\|q) \;=\; \tfrac{1}{2}\,\mathrm{KL}\big(p \,\big\|\, \tfrac{p+q}{2}\big) \;+\; \tfrac{1}{2}\,\mathrm{KL}\big(q \,\big\|\, \tfrac{p+q}{2}\big)\,, \tag{1}$$

where $\mathrm{KL}(p\|q) = \sum_b p_b \log \frac{p_b}{q_b}$ is the discrete (binned) KL. We report *per-feature* JS for the continuous columns, comparing pairs $(\mathrm{GT}, \mathrm{SIM})$, $(\mathrm{GT}, \mathrm{SYN})$, and $(\mathrm{SIM}, \mathrm{SYN})$. Lower is better, and $\mathrm{JS} = 0$ iff histograms match exactly.[1]

6.2 Maximum Mean Discrepancy (MMD) with RBF Kernel

Let $k(x, y)$ be a positive definite kernel. The squared MMD between P and Q in the RKHS $\mathcal{H}$ induced by k is

$$\mathrm{MMD}^2(P, Q) \;=\; \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 \;=\; \mathbb{E}_{x,x'}[k(x, x')] + \mathbb{E}_{y,y'}[k(y, y')] - 2\,\mathbb{E}_{x,y}[k(x, y)]. \tag{2}$$

With samples, an unbiased estimator is

$$\widehat{\mathrm{MMD}}^2 \;=\; \frac{1}{n(n-1)}\sum_{i\neq i'} k(x_i, x_{i'}) \;+\; \frac{1}{m(m-1)}\sum_{j\neq j'} k(y_j, y_{j'}) \;-\; \frac{2}{nm}\sum_{i,j} k(x_i, y_j). \tag{3}$$

We use the RBF kernel $k(x, y) = \exp(-\gamma\,\|x - y\|^2)$ with $\gamma = \frac{1}{2\sigma^2}$ and σ chosen by the median heuristic on a subsample. Lower MMD indicates closer joint distributions. We report MMD for $(\mathrm{GT}, \mathrm{SYN})$, $(\mathrm{GT}, \mathrm{SIM})$, and $(\mathrm{SIM}, \mathrm{SYN})$.

6.3 Classifier Two-Sample Test (C2ST) AUC

The C2ST trains a binary classifier to distinguish P-samples from Q-samples. Let $f : \mathbb{R}^D \to [0, 1]$ be a probabilistic classifier trained on labeled examples $\{(x_i, 0)\}_{i=1}^n \cup \{(y_j, 1)\}_{j=1}^m$. A strong classifier suggests a detectable difference between P and Q. We report the ROC–AUC on the combined sample:

$$\mathrm{AUC} \;=\; \mathbb{P}\big(f(Y^+) > f(X^-)\big), \tag{4}$$

where Y^+ is a positive (from Q) and X^- is a negative (from P) draw. An AUC of 0.5 indicates indistinguishability by the classifier (chance), while 1.0 indicates perfect separability. We use logistic regression (linear decision boundary) for stability and interpretability.

[1] Because JS here is computed from histograms, values depend on the binning scheme. We use a fixed number of bins and shared ranges for fairness.

Interpretation summary

- **Per-feature JS (continuous)**: marginal fidelity; sensitive to univariate shifts/tails.
- **MMD (RBF)**: global multivariate discrepancy; captures joint geometry across all features.
- **C2ST AUC**: discriminability; how easy it is to tell the two samples apart.

Combining these provides complementary views: JS for marginal alignment, MMD for joint structure, and C2ST for practical detectability.

7 Outputs and Artifact Schema

Each run writes:

- **Figures**: histograms for each continuous feature comparing GT/SIM/SYN.
- **JSON**: a single machine-readable `results.json` containing provenance and metrics.

JSON (abridged shape)

```
{
  "experiment": { "name": "adult_tabular", "timestamp": "...", "seed": 42 },
  "data": {
    "ground_truth": { "type": "external", "source": "...", "n_train": ..., ... },
    "simulated":    { "generator": "per_feature_gaussian_copula", "n": ... },
    "synthetic":    { "library": "SDV", "model": "CTGAN" | "TVAE", "train": {...} }
  },
  "metrics": {
    "continuous_js": { "<feature_name>": { "gt_vs_sim": ..., "gt_vs_syn": ..., ... }, ... },
    "global": {
      "mmd_rbf_gt_syn": ..., "mmd_rbf_gt_sim": ..., "mmd_rbf_sim_syn": ...,
      "c2st_auc_gt_syn": ..., "c2st_auc_gt_sim": ..., "c2st_auc_sim_syn": ...
    }
  },
  "figures": ["hist_age.png", "hist_fnlwgt.png", ...],
  "notes": "...",
  "versions": { "python": "...", "platform": "..." }
}
```

8 Reproducibility and CLI Parameters

Seeds A single `--seed` is propagated to NumPy/PyTorch (where applicable) and used for simulators and sampling.

Key flags

- `--sim_mode` $\in$ {`independent`, `gaussian_copula`} controls dependence modeling in SIM.
- `--synth` $\in$ {`ctgan`, `tvae`} chooses the SDV model.
- `--epochs`, `--batch_size` control training; CTGAN also uses `--pac` and requires `batch_size \% pac = 0`.
- `--n_eval` sets the sample size for SIM/SYNand the reference size from GT (size-matched).

9 Practical Interpretation Guide

When comparing $(\mathrm{GT}, \mathrm{SYN})$ and $(\mathrm{GT}, \mathrm{SIM})$:

- If **JS** is low across most continuous features, marginal distributions are well matched.
- If **MMD** is low while JS is low, the joint distribution is also close.
- If **C2ST AUC** is near 0.5, a linear classifier cannot easily distinguish the samples; large AUC indicates detectable differences.