InfoGram and Admissible Machine Learning

Deep Mukhopadhyay

deep@unitedstatalgo.com

NIST AI Bias Meeting

ML 1.0: Predictive Machine Learning [1960 –]

- First-Generation "Predictive" ML: Developed over the last 60 years—since the early 1960s, and produced a bundle of powerful (accurate & flexible) algorithms like svm, gbm, random forest, deep neural net, etc.

- Success story: Enormous, especially in tech and eCommerce industry.

- `AutoML`: Builds high-performance ML-algos by automating away a lot of mundane tasks like learner selection, feature engineering, and hyperparameter optimization.

The Emerging Regulatory Environment

> *Faced with the profound changes that AI technologies can produce, pressure for "more" and "tougher" regulation is probably inevitable.*
>
> — 100-Year Study on AI, Stanford (2019)

- Development ≠ Deployment: While substantial progress has been made toward developing more powerful ML 1.0 algorithms, the widespread adoption of these technologies currently facing regulatory roadblock, especially in safety-critical areas that directly affect human lives.
- Burning question: how to *systematically* build regulatory compliant algorithms by balancing fairness, interpretability, and accuracy in the best manner possible?

Machine Learning (2022) 111:205–242
https://doi.org/10.1007/s10994-021-06121-4

Check for updates

InfoGram and admissible machine learning

Subhadeep Mukhopadhyay[1]

Received: 3 December 2020 / Revised: 27 October 2021 / Accepted: 28 October 2021 /
Published online: 10 January 2022

Abstract
We have entered a new era of machine learning (ML), where the most accurate algorithm with superior predictive power may not even be deployable, unless it is *admissible* under the regulatory constraints. This has led to great interest in developing fair, transparent and trustworthy ML methods. The purpose of this article is to introduce a new information-theoretic learning framework (admissible machine learning) and algorithmic risk-management tools (InfoGram, L-features, ALFA-testing) that can guide an analyst to *redesign* off-the-shelf ML methods to be regulatory compliant, while maintaining good prediction accuracy. We have illustrated our approach using several real-data examples from financial sectors, biomedical research, marketing campaigns, and the criminal justice system.

Keywords Admissible machine learning · InfoGram · L-Features · Information-theory · ALFA-testing · Algorithmic risk management · Fairness · Interpretability · COREml · FINEml

ML 2.0: Admissible Machine Learning [2021 –]

Executive Summary

`AdmissibleML` offers new statistical learning principles and algorithmic risk-management tools that can guide a ML-developer to *quickly build better* algorithms that are less-biased, more-interpretable, and sufficiently accurate.

Application 1: Algorithmic Fairness

The Census Income Data. It is extracted from 1994 United States Census Bureau database, which contains $n = 45,222$ records involving personal details on:

$y_{n\times 1}$: $\mathbb{1}(\texttt{income} > \$50\text{k/yr})$

$\mathbf{S}_{n\times q}$: Sensitive vars; $\{\texttt{Age}, \texttt{Gender}, \texttt{Race}, \texttt{Marital\_Status}\}$

$\mathbf{X}_{n\times p}$: 10 attributes; $\{\texttt{Education level}, \texttt{Occupation}, \ldots\}$

Goal: Predict whether a person makes \$50k per year while minimizing unfair discrimination based on protected classes.

ML 1.0: Pure Prediction Algorithm

Step 1. Choose a ML algorithm.

Step 2. Train the ML classifier only on $\mathbf{X}$ (i.e, without sensitive attributes)

$$\boxed{\texttt{ML}(y \sim \mathbf{X})}$$

Step 3. Deploy the most **accurate** ML_0.

Gradient Boosting Machine (GBM)

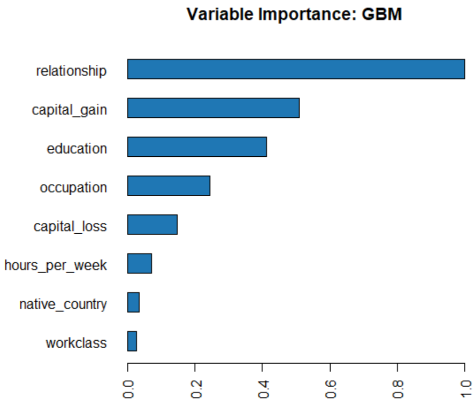


Figure: Shows relevance-index R_j. The top feature `relationship` represents the respondent's role in the family—i.e., whether the earning member is husband, wife, child, or other relative. Avg. test accuracy: 85.65% (on 15% test set, repeated 50 times).

Is it Deployable?

- Obviously, it shouldn't be deployed without assessing whether the model is admissible under discrimination laws based on protected characteristics.
- Achieving high predictive-accuracy is as important as ensuring regulatory compliance and transparency.
- So, how should we proceed now?

Current Framework

Good news: Significant research efforts in the last 4-5 years led to some concrete AI toolkits:

- IBM's Fairness 360 [developed in 2018]
- Microsoft's FairLearn [developed in 2020]

They provide two core facilities:

1. Fairness assessment through different metrics.
2. Different unfairness mitigation methods.

Assessment Strategies: Limitations

Too many numbers with too little information. Dashboard full of fairness metrics: IBM 360 Fairness tool currently produces **77** fairness related metrics!

1. The Troubling Part: These fairness measures are mutually incompatible and cannot be satisfied simultaneously. How to reconcile these large collections of self-contradictory metrics to make a confident decision? **Not clear**.
2. Marginal assessment: These methods ask user to choose (i) one single discrete sensitive variable (e.g., race, gender, or marital_status) and computes a series on numbers. Recall: our `Income` dataset has 4 sensitive variables.
3. What happens if a sensitive feature is continuous (e.g., age)? **Not clear**. What happens if $\mathbf{S}$ is multivariate: **Not clear**.

Note 1.

Cataloging a huge library of inherently contradictory model validation metrics is hardly going to help ML-engineers to search for a deployable model. Instead of *searching in a dark*, we need some other methodical & prudent strategy.

Note 2.

We need an "*Explanatory*" Risk Management (**XRM**) framework that can provide explanation and insights into *what* (are the key sources of bias) and *how* (to combat unwanted bias) for accelerating the model-search.

Mitigation Strategies: Limitations

Step 1. Choose one particular fairness metric from a big pool.

Step 2. Choose one of the following three strategies:

- *Pre-processing*: Re-weights or re-labels the original data to minimize the given fairness measure.
- *In-processing*: Optimizes hyperparameters of a blackbox ML by imposing the given fairness measure as constraint.
- *Post-processing*: Controls the given (un)fairness metric by artificially changing the classification thresholds for each protected group.

Note 3.

All 3 unfairness mitigation strategies carry serious legal compliance risk: Because either they undertake (i) data massaging/manipulation; or (ii) they use protected attributes during model training or decision making.

What practitioners actually do? **A top AI-practitioner**:

"*I ran 40,000 different random forest models with different features and hyper-parameters to search a fair model.*"

Note 4.

Non-constructive Approach: No wonder, this ad-hoc random process often ends up being a wild-goose chase, resulting in a spectacular waste of computation and time.

ML 2.0: Infogram and Admissible Machine Learning

- **Theory-side**: The paper lays out the *core principles* for designing `AdmissibleML` which is grounded in fundamental information-theoretic and nonparametric statistical ideas.
- **Utility-side**: Provides concrete algorithmic tools to *aid the development* of regulatory compliant fair and transparent AI systems—essential for earning trust of customers/public.

Key Concepts and Tools

- Infogram
- L-Features
- ALPHA-testing
- AdmissibleML: COREtree, COREglm, FINEtree, FINEglm.

InfoGram: Practice

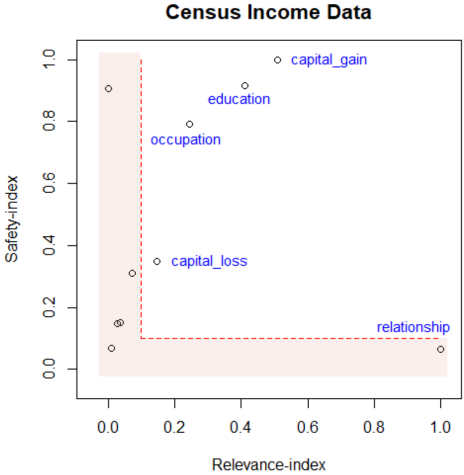


Figure: Infogram maps variables in a two dimensional effectiveness vs. safety diagram. It is an exploratory tool for risk-benefit analysis that provides insights into 'what and how'.

Safety-Index: Definition and Interpretation

Definition. Define the safety-index for variable X_j as

$$F_j \; = \; \mathrm{MI}\left(Y, X_j \mid \{S_1, \ldots, S_q\}\right), \quad j = 1, \ldots, p.$$

Interpretation.

- It quantifies how much extra information X_j carries for Y that is not acquired through the sensitive variables $\mathbf{S}$.
- Variables with "small" F-values will be called *inadmissible*, as they possess little or no informational value beyond their use as a dummy for protected characteristics.

InfoGram is an acronym for information diagram, which is a scatter plot of $\{(R_1, F_1), \ldots, (R_p, F_p)\}$.

- The variable `relationship` is highly predictive, yet a proxy for the sensitive attributes.
- **A dangerous consequence**: Most unguided predictive ML algorithms will include in their models, even though it is quite unsafe.
- Admissible ML models should avoid[1] using variables like `relationship` to reduce unwanted bias.

Note 5.

Without a formal automated method, it is a hopeless task (for model developers and regulators) to identify these innocent-looking hidden proxy variables for modern-day large-scale problems.

[1] At least should be assessed by experts to determine its appropriateness.

Admissible ML: FINEtree

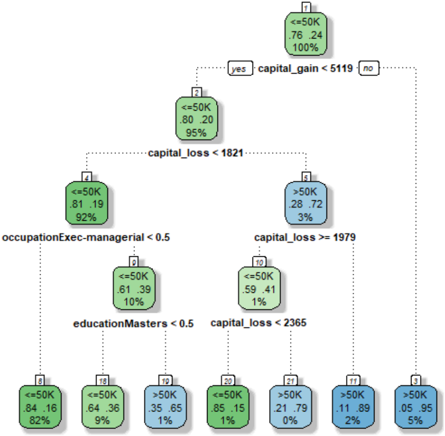


Figure: `FINE` = An Admissible ML-model that balances **F**airness, **IN**terpretability, and **E**fficiency. Accuracy: 83.5%.

Theory: Outline

The foundation of AdmissibleML relies on information-theoretic principles and nonparametric statistical methods. The key ideas and results are presented in Section 2 of my paper.

It has four connected parts:

- $\boxed{\text{Formulation}}$
- $\boxed{\text{Interpretation}}$
- Estimation
- Inference

Information-theoretic Formulation

Notation.

- $Y \in \{1, \ldots, k\}$ is the response variable.
- $\mathbf{X} = (X_1, \ldots, X_p)$: p-dimensional feature matrix
- $\mathbf{S} = (S_1, \ldots, S_q)$: q-dimensional sensitive attributes.

Definition. Conditional mutual information (CMI) between Y and $\mathbf{X}$ given $\mathbf{S}$ is defined as:

$$\mathrm{MI}(Y, \mathbf{X}|\mathbf{S}) = \iiint_{y,\mathbf{x},\mathbf{s}} \log\left(\frac{f_{Y,\mathbf{X}|\mathbf{S}}(y,\mathbf{x}|\mathbf{s})}{f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s})}\right) f_{Y,\mathbf{X},\mathbf{S}}(y,\mathbf{x},\mathbf{s})\, \mathrm{d}y\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{s}.$$

Usual Interpretation #1

Under conditional independence:

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}$$

the following decomposition holds for all $y, \mathbf{x}, \mathbf{s}$

$$f_{Y,\mathbf{X}|\mathbf{S}}(y, \mathbf{x}|\mathbf{s}) \; = \; f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s}).$$

CMI *quantifies* the conditional dependence: the average deviation of the ratio

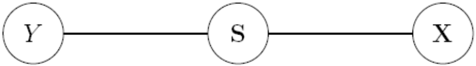
$$\frac{f_{Y,\mathbf{X}|\mathbf{S}}(y, \mathbf{x}|\mathbf{s})}{f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s})},$$

Property and Graphical Model

An imp property: CMI possesses the necessary and sufficient condition as a measure of conditional independence

$$\mathrm{MI}(Y, \mathbf{X}|\mathbf{S}) = 0 \text{ if and only if } Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}.$$

Conditional independence can be described graphically, where each node is a random variable (or random vector).



NOTE: The edge between Y and $\mathbf{X}$ passes through the $\mathbf{S}$.

More Useful Interpretation #2

The conditional entropy $H(Y|\mathbf{S})$ is defined as

$$H(Y \mid \mathbf{S}) = \int_{\mathbf{s}} H(Y \mid \mathbf{S} = \mathbf{s})\, \mathrm{d}F_{\mathbf{s}},$$

which measures how much uncertainty remains in Y *after* knowing $\mathbf{S}$, on average.

Theorem 1. $\mathrm{MI}(Y, \mathbf{X}|\mathbf{S})$ can be expressed as the difference between two conditional-entropy statistics:

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; H(Y \mid \mathbf{S}) \;-\; H(Y \mid \mathbf{S}, \mathbf{X}) \tag{1}$$

Interpretation. This alternative representation of CMI (1) allows us to interpret it from a new angle: $\mathrm{MI}(Y, \mathbf{X}|\mathbf{S})$ measures the net impact of $\mathbf{X}$ in reducing the uncertainty of Y, given $\mathbf{S}$.

Nonparametric Estimation: Theory

Theorem 2. Let Y be a discrete random variable taking values $1, \ldots, k$, and $(\mathbf{X}, \mathbf{S})$ be a `mixed` pair of random vectors. Then the conditional mutual information can be rewritten as

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; \mathbf{E}_{\mathbf{X},\mathbf{S}}\Big[\mathrm{KL}\big(p_{Y|\mathbf{X},\mathbf{S}} \parallel p_{Y|\mathbf{S}}\big)\Big], \tag{2}$$

where Kullback-Leibler (KL) divergence from $p_{Y|\mathbf{X}=\mathbf{x},\mathbf{S}=\mathbf{s}}$ to $p_{Y|\mathbf{S}=\mathbf{s}}$ is defined as

$$\mathrm{KL}\big(p_{Y|\mathbf{X},\mathbf{S}} \parallel p_{Y|\mathbf{S}}\big) \;=\; \sum_{y} p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x},\mathbf{s}) \log\left(\frac{p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x},\mathbf{s})}{p_{Y|\mathbf{S}}(y|\mathbf{s})}\right).$$

Interpretation #3. Eq. (2) $\Rightarrow$ CMI measures how much information is shared only between $\mathbf{X}$ and Y that is not contained in $\mathbf{S}$. This viewpoint is used throughout my paper.

Nonparametric Estimation: Algorithm

Given: n i.i.d samples $\{\mathbf{x}_i, y_i, \mathbf{s}_i\}_{i=1}^n$.

Theorem 2 immediately leads to the following estimator of CMI that works for large(n, p, q) settings:

$$\widehat{\mathrm{MI}}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \log \frac{\widehat{\Pr}(Y = y_i|\mathbf{x}_i, \mathbf{s}_i)}{\widehat{\Pr}(Y = y_i|\mathbf{s}_i)}. \tag{3}$$

Algorithm. Choose a ML classifier (e.g., SVM, rf, gbm, deep neural net, etc.) and train the following two models:

$$\begin{aligned} \texttt{ML.train}_{y|\mathbf{x},\mathbf{s}} &\;\leftarrow\; \texttt{ML}_0\big(Y \sim [\mathbf{X}, \mathbf{S}]\big) \\ \texttt{ML.train}_{y|\mathbf{s}} &\;\leftarrow\; \texttt{ML}_0\big(Y \sim \mathbf{S}\big) \end{aligned}$$

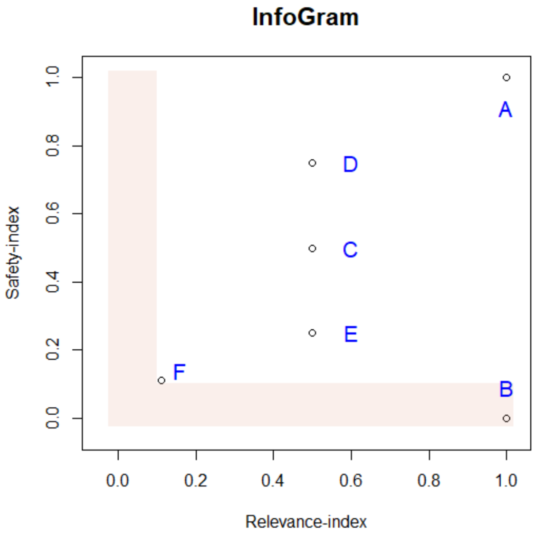
to the conditional probability estimates of (3).

Three Practical Benefits

Our style of nonparametric estimation of $\widehat{\mathrm{MI}}(Y, \mathbf{X} \mid \mathbf{S})$ comes with some important practical benefits:

- Flexibility: Requires neither the knowledge of the exact parametric form of high-dimensional $F_{X_1,\ldots,X_p}$ nor the knowledge of the conditional distribution of $\mathbf{X} \mid \mathbf{S}$
- Applicability: The method can be safely used for *mixed* $\mathbf{X}$ and $\mathbf{S}$—i.e, *any combination* of discrete, continuous, or even categorical variables.
- Scalability: The procedure is scalable for *high-dimensional big datasets* with large(n, p, q).

InfoGram: Graphical Interpretation



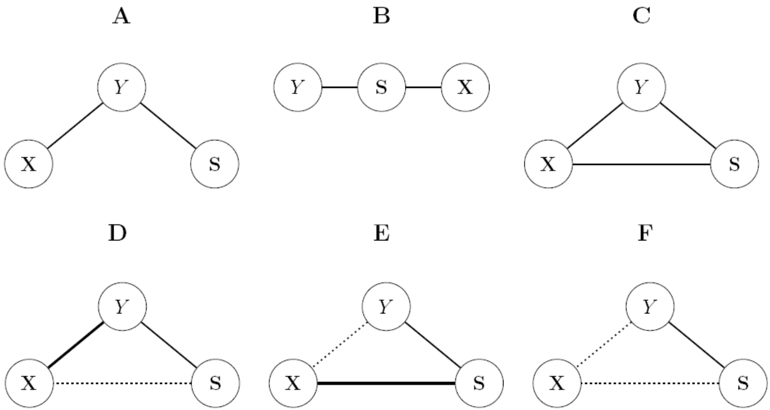


Figure: Fairness is not a yes/no concept, but a matter of degree, which is quantified via safety-index—indicated by the varying edge thicknesses between **S** and **X**. `InfoGram` provides the necessary guardrails for constructing algos that can retain as much predictive accuracy as possible, while defending against unforeseen biases.

Application 2: Digital Marketing Campaign Data

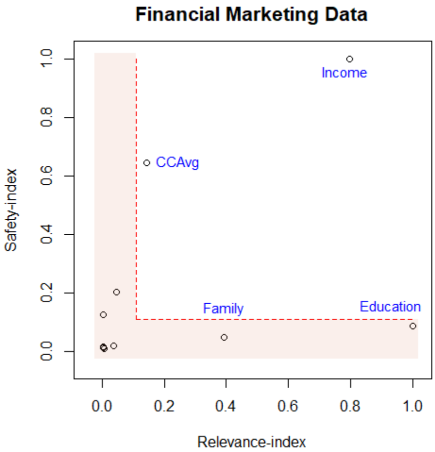


Figure: Goal is to develop an AI tool for *automatic and fair* digital marketing campaign that will maximize the targeting effectiveness of the ad campaign while minimizing the harmful effects on protected groups. $\mathbf{S} = \{\texttt{age, zip code}\}$ and $p = 10$ additional features.

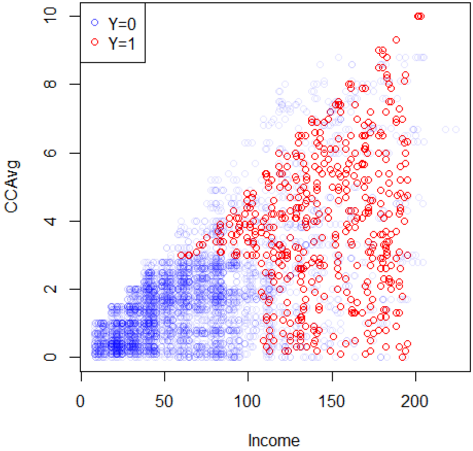


Figure: Infogram runs a ‘combing operation’ to distill down a large, complex problem to its *core* that holds the bulk of the “admissible information.” The useful information is mostly concentrated into two variables—`Income` and `CCAvg`, as seen in the scatter diagram; the color blue and red indicate two different classes.

Customer Targeting using AdmissibleML: FINEglm

- Extracting a simple model: We train a logistic regression model based on the two admissible features, leading to the following model:

$$\text{logit}\{\mu(x)\} = -6.13 + .04\,\texttt{Income} + .06\,\texttt{CCAvg},$$

where $\mu(x) = \Pr(Y = 1|\mathbf{X} = \mathbf{x})$. This simple model achieves 91% accuracy. It provides a clear understanding of the 'core' factors that are driving the model's recommendation.

- Infogram-assisted ML: An efficient, interpretable, and equitable algorithmic recommendation system—which ensures that we are making '*responsible*' decisions using such algorithm.

Application 3: Algorithmic Interpretability

Breast Cancer Wisconsin Data. It contains $n = 569$ malignant and benign tumor cell samples. The task is to build an accurate ML classifier based on $p = 31$ features extracted from cell nuclei images.

ML 1.0. Gbm and random forest attain accuracy in the range of $95 - 97\%$. Quite impressive!

Is it deployable? Will an oncologist or hospital use this AI-technology to make decisions about their patients? Probably not since the core algorithmic "logic" is incomprehensible to medical experts. In Science *why* is as important as *what*.

Revised goal: ML 2.0. We need to design an *admissible* (interpretable and accurate) learning algorithm.

Infogram

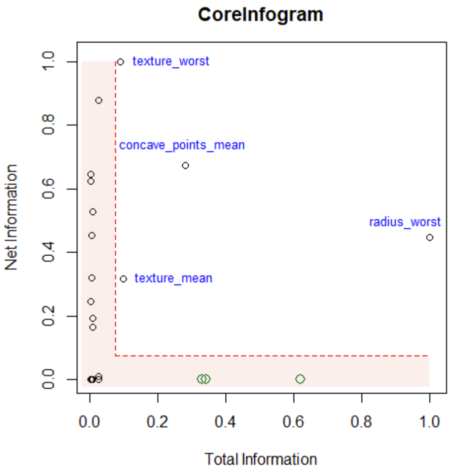


Figure: L-features: The highlighted L-shaped area contains features that are either irrelevant or redundant. Predictive Features $\neq$ CoreSet

Theory

- Identifying `CoreSet` is a much more difficult undertaking than merely selecting the most predictive ones.

- To enable refined characterization of the vars, we've to add more dimension to the classical ML feature importance tools.

Definition. Net-predictive information (NPI) of a feature X_j given all the rest of the variables $\mathbf{X}_{-j} = \{X_1, \ldots, X_p\} \backslash \{X_j\}$ is defined in terms of conditional mutual information:

$$C_j \;=\; \mathrm{MI}(Y, X_j \mid \mathbf{X}_{-j}), \quad \text{for } j = 1, \ldots, p.$$

- The joint plot of $\{(C_1, R_1), \ldots, (C_p, R_p)\}$ aims to discover the *core variables* that are driving the outcome.

CORE scatter plot

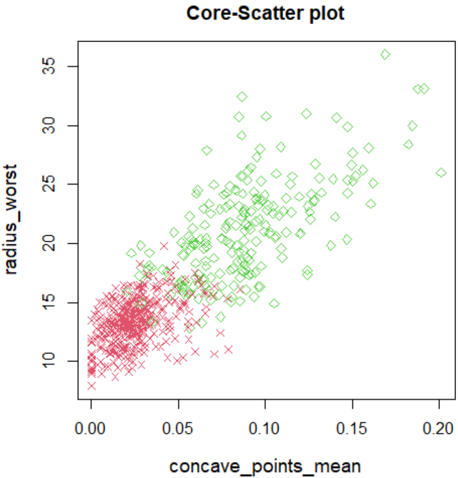


Figure: Reveals where the crux of the information is hidden and *how* they separate the malignant and benign tumor cells.

Admissible ML: COREglm

The simplest possible model that one could build is a logistic regression based on those admissible “core” features.

The output of `glm`() R-function:

```
#COREglm Model: UCI breast cancer data
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -29.42361    3.85131  -7.640 2.17e-14 ***
concave_points_mean 96.48880   16.11261   5.988 2.12e-09 ***
radius_worst         0.99767    0.16792   5.941 2.83e-09 ***
texture_worst        0.30451    0.05302   5.744 9.27e-09 ***
```

This infogram-guided 3-variable simple model turns out to be surprisingly accurate 96.50%—as accurate as complex black-box ML methods, yet highly transparent and interpretable.

Final Remarks

- ML 1.0: `PredictiveML` culture, where the expectation from a Stat-model is to produce the most accurate prediction.
- ML 2.0: `AdmissibleML`, where the expectation from a Stat model is to aid understanding and safe decision-making.
- ML 1.0: Long history since 1960s: knn, kernel methods, CART, random forest, GBM, and recent deep learning.
- ML 2.0: Going through its infancy; Slow progress—designing statistical mechanism for 'Responsible AI' is much HARDER than developing another ML 1.0 method.
- **My claim**: The next decade will see rapid progress in *fundamental ideas and related tools* required to establish a strong foundation for ML 2.0. But this will need adequate support and funding from Government & Industry.