統計的機械学習　（応用計量分析２）第１回

本講義について

最新の情報はCLASSを参照してください

本資料

- 統計的機械学習　（応用計量分析２）（9987F10）
- 担当教員：谷本　啓
  - 非常勤講師 / 株式会社Preferred Networks リサーチャー
  - 連絡先：a31607@rs.tus.ac.jp
  - 評価方法：レポートまたは到達度評価
    - 補助的に小テストを活用（25%程度）
- 最新情報はCLASSを参照のこと
- 本日の出欠コード：**1298**

本講義の主題

出欠コード
XXXX

データに基づく意思決定のための因果推論（+α）

因果推論とは

介入

…意思決定向け分析に用いる
例) 広告すれば売上がいくら上がる？

←**本講義の主題**

もしこう{ した / していた }ならどう{ なる / なっていた }かを推量する分析技術

反実仮想

…(ほぼ) 振り返り的分析に用いる
例) この女性がもし男性だったら
面接で採用されていたか？

世の中にたえて桜のなかり**せば**
春の心はのどけから**まし**
在原業平

出欠コード
XXXX

『因果機械学習』ドラフト＋補助参考書

- 谷本『因果機械学習』ドラフト
  - https://akira-tanimoto.netlify.app/pdfs/causalMLbook_r18.pdf
  - 誤りを見つけて報告してくれた人は謝辞にて感謝します
- 補助参考書
  - 各トピックにさらに興味が出た場合には下記の本を勧めます
  - ギルボア『不確実性下の意思決定理論』
  - パール『統計的因果推論　モデル・推論・推測』
  - モーガン＆ウィンシップ『反事実と因果推論』

内容（予定）

進度と内容は様子を見ながら調整します

- **1. ガイダンス・因果推論と機械学習の概論**
- 2. 意思決定理論（期待効用理論）の復習、因果推論との関係
- 3. 潜在結果モデルに基づく因果推論の枠組み
- 4. 平均因果効果の推定法
- 5. 条件付き平均因果効果（CATE）の推定法１：メタ学習器
- 6. CATEの推定法２：二重機械学習
- 7. CATEの推定法３：決定木と決定森
- 8. 構造方程式モデルとバックドア基準
- 9. 因果探索
- 10. 発展的な因果推論手法：フロントドア調整、操作変数法、回帰不連続デザイン、代理変数法
- 11. 発展的な意思決定理論
- 12. 強化学習
- 13. オフライン強化学習
- 14. バンディット
- 15. まとめ

出欠コード
XXXX

因果推論と機械学習

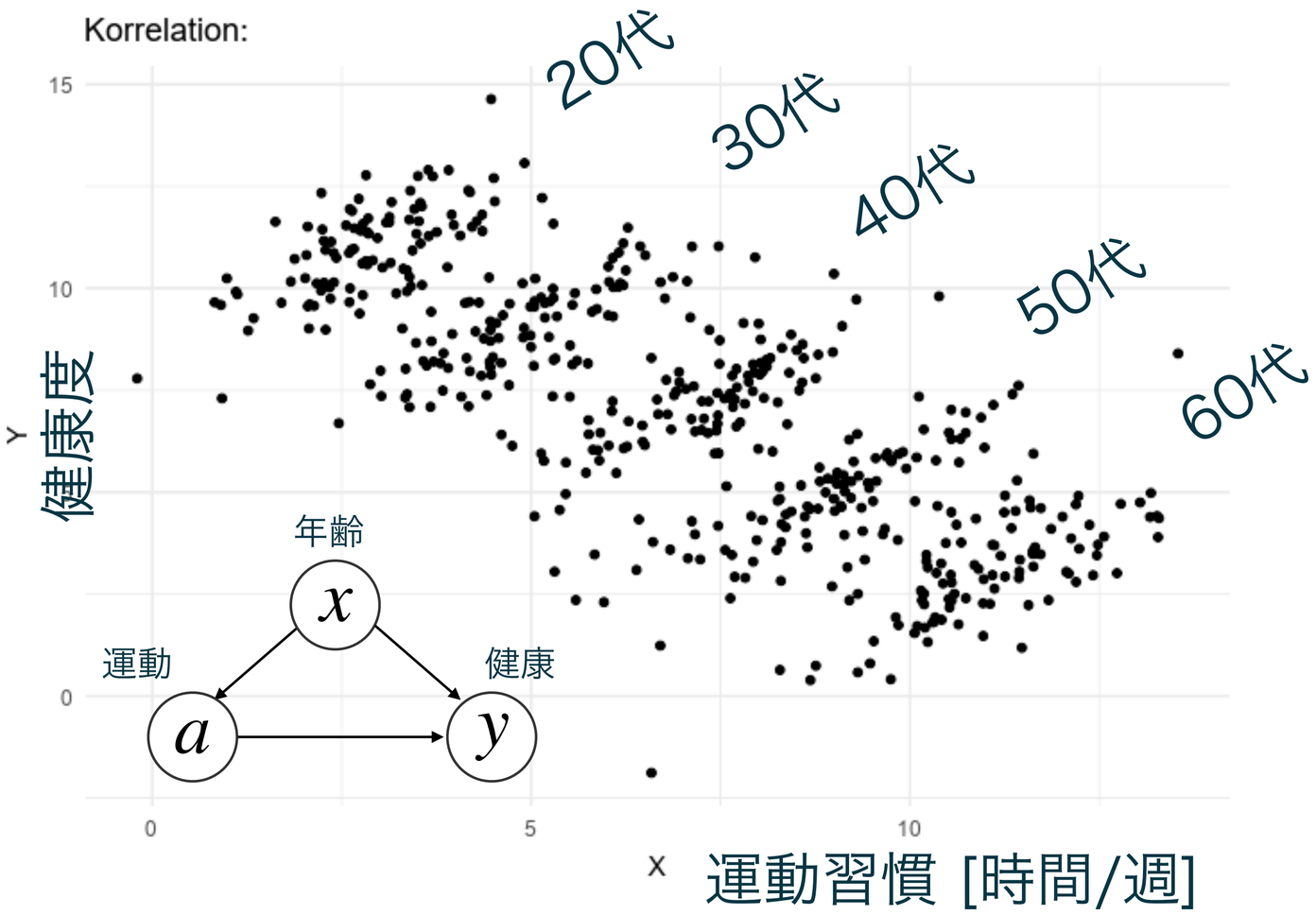
- 1. ガイダンス・因果推論と機械学習の概論
  - 因果推論の動機
  - 教師あり学習問題と一般のデータに基づく意思決定問題の違い
  - 教師あり機械学習の概説

因果推論と意思決定

動機：因果と相関の違い

シンプソンのパラドックス：データの見方で結論が変わる

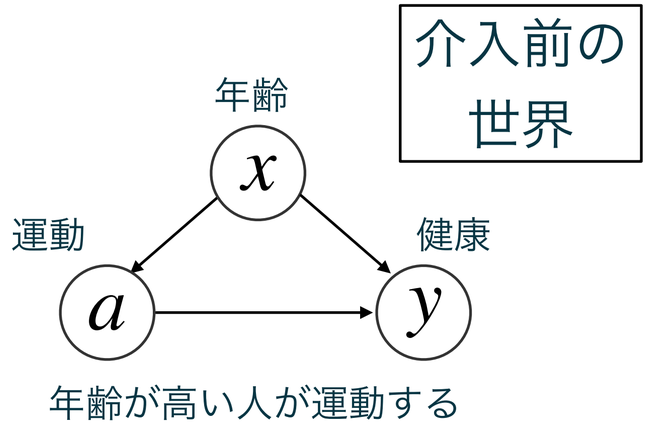
- どちらが正しい？
  - 全体で見ると運動する人ほど不健康
  - 年代別に見ると運動する人ほど健康
- →どちらも正しい
- しかし
  「運動する人ほど健康でない」ことは
  「運動すると健康でなくなる」ことを意味しない
- データを何に使うのか（意思決定）を考えることが重要



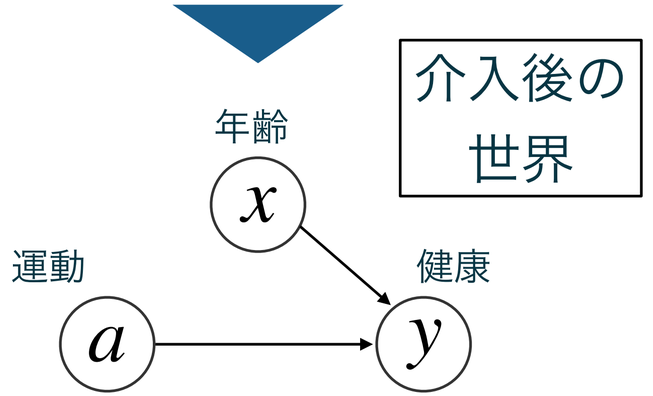
意思決定に資する分析とは

意思決定の仕方を変えても不変の規則性＝因果性

- データから意思決定に資する規則性を抽出するには？
  - 意思決定の仕方を変えても不変の規則性が必要
- 世界はモジュール的である（と仮定）
  - ある部分の意思決定の仕方を変えても
    （例：年齢が若くても運動する）
    他の部分の因果機序（$y = f(x, a)$）は不変
- 世界のモジュール性に沿ったモデリングをする
  ＝因果推論



年齢が高い人が運動する



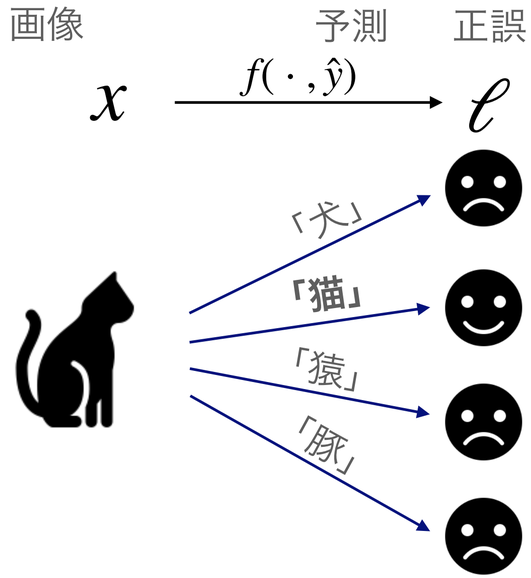
年齢が若くても運動する

教師あり機械学習（普通の予測）との違い：例

出欠コード
XXXX

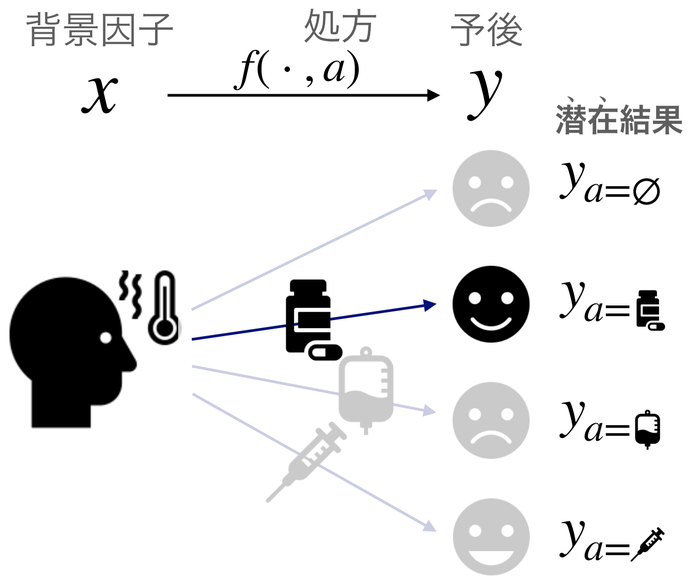
意思決定問題では一般に反事実的結果が欠損

教師あり学習



- 画像認識
  - ラベル以外は間違い

因果推論



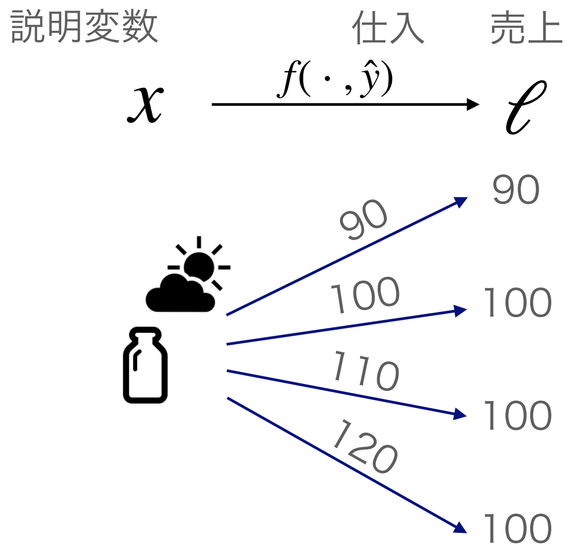
- 医療措置予後予測
  - 実際に行った措置の結果のみ観測

教師あり機械学習（普通の予測）との違い：実践的な例

出欠コード
XXXX

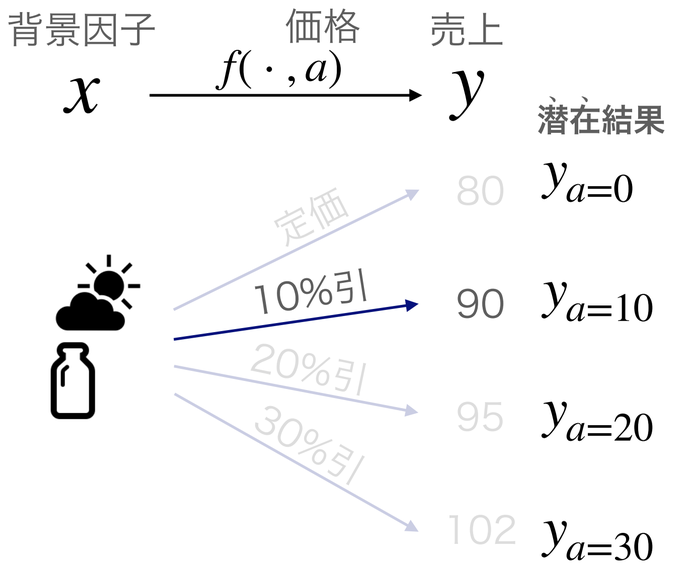
発注最適化＝(ほぼ)教師あり、価格最適化＝部分観測

教師あり学習



- 需要予測
  - 売れた量＝需要と仮定

因果推論



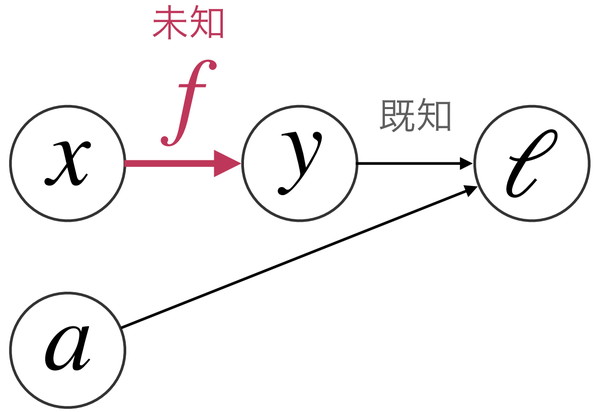
- 価格弾力性推定
  - 実売価格に対する需要のみ観測

教師あり機械学習（普通の予測）との違い：一般化

出欠コード
XXXX

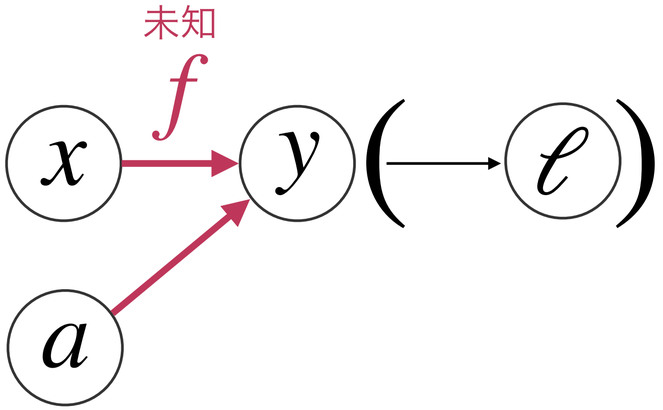
最終的な意思決定における変数が推定対象に直接関与するか

教師あり学習



- 意思決定変数 a と未知関数が分離

因果推論



- 意思決定変数 a を入力に含む関数が未知

（教師あり）機械学習の概説

教師あり学習を応用して因果推論を行うので重要

機械学習とは

出欠コード
XXXX

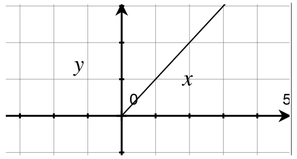
データから規則性を抽出し、予測や意思決定を行う技術

- **訓練データ**Dから**予測モデル**（関数） $\hat{f}$ を推定する**学習器** A を考える
  - $\hat{f} = A(D)$
- 予測モデル $\hat{f}$ は変数 x をもとに変数 y を予測する： $y \fallingdotseq \hat{f}(x)$　と期待される
- 学習器 A は通常何らかの**損失関数** L （予測誤差の近似）を最小化する関数として定義
  - $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{L}(f; D)$　：$\mathcal{F}$の中でデータDに基づく損失関数$\hat{L}$を最小化する関数を$\hat{f}$とする
    - $\mathcal{F}$ (花文字のF) は**仮説クラス**と呼ばれる。ハット ˆ は推定量を表す
- モデルクラス$\mathcal{F}$・損失関数L・最適化アルゴリズムを別々に考え、組み合わせられることが多い
  - 非線形で柔軟なモデルを考え、それによる個別化された意思決定を志向する傾向
  - ※とはいえ統計学との重なりも大きい

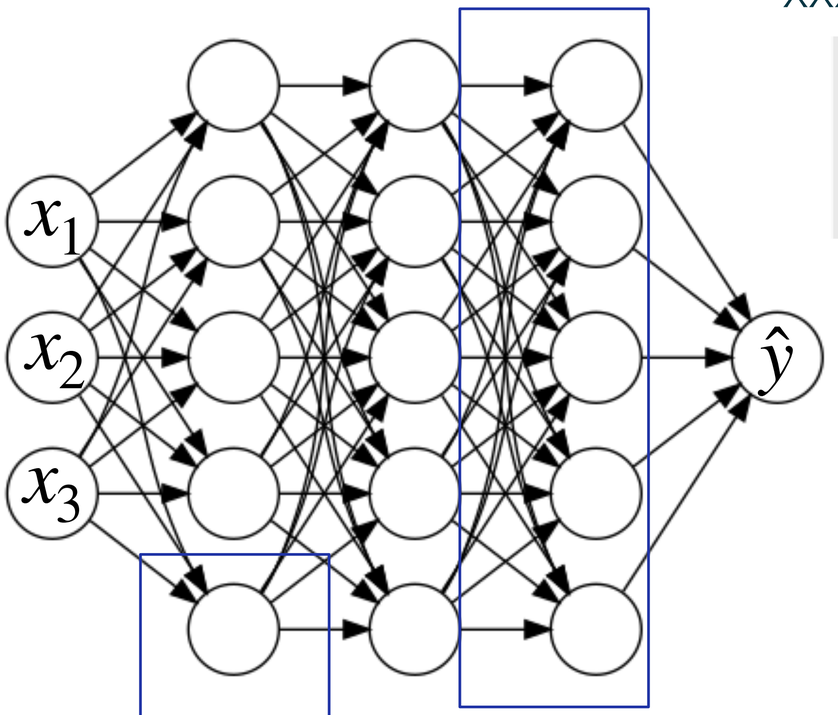
具体例（モデルクラス）

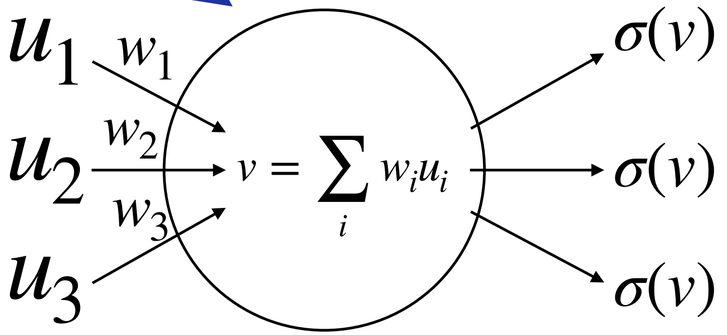
多層パーセプトロン

- 深層モデルの基本形
- ベクトルの線形変換と要素ごとの非線形変換 σ を繰り返し適用
  - $f(x) = W_L\sigma\left(\ldots W_2\sigma\left(W_1x\right)\ldots\right)$
- $\theta = \left(W_1, \ldots, W_L\right)$ がモデルパラメタ（推定対象）
- 非線形関数はRectified Linear Unit $\sigma(v) = \max(0,v)$ など



- 判別問題の場合は最終層にsoftmax関数 $g(u)_i = \dfrac{e^{u_i}}{\sum_j e^{u_j}}$ を適用



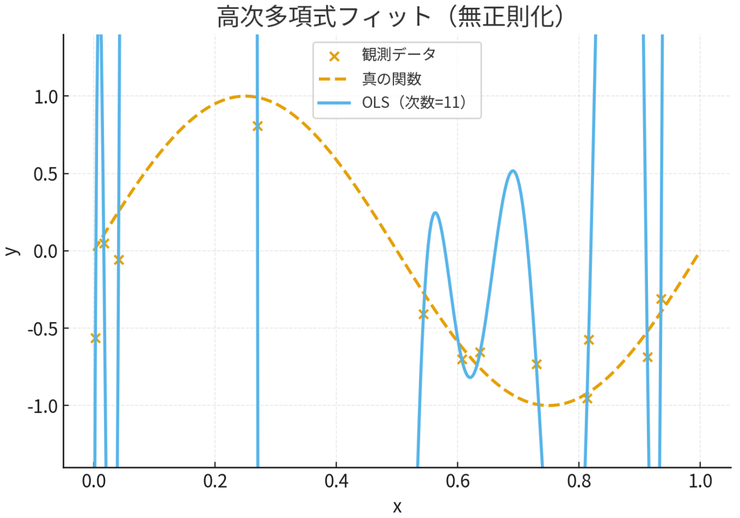


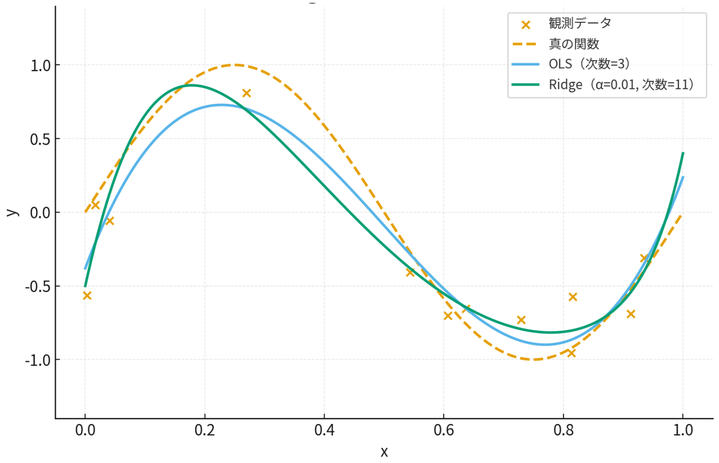
具体例（損失関数）

出欠コード
XXXX

予測精度が高く、簡潔な関数を選ぶ

- 個々のデータ点に対する損失ℓの典型は
  - 回帰は２乗誤差：$\ell(f; o^i) = (y^i - f(x^i))^2$
  - 二値判別は交差エントロピー：
    $\ell(f; o^i) = -y^i \log f(x^i) - (1 - y^i)\log(1 - f(x^i))$
- モデルの“複雑さ”に対する罰則を加える場合が多い
  - $\hat{L}_{\text{reg}}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell(f; o^i) + \alpha R(f)$
  - 罰則はパラメタの２乗など $R(f) = \|\theta\|_2^2$　（Ridge正則化と呼ばれる）
  - サンプルサイズに対してモデルクラスが複雑すぎると**過学習**するため
    （訓練データに過剰に適合し、未知のデータに対する精度が下がる）



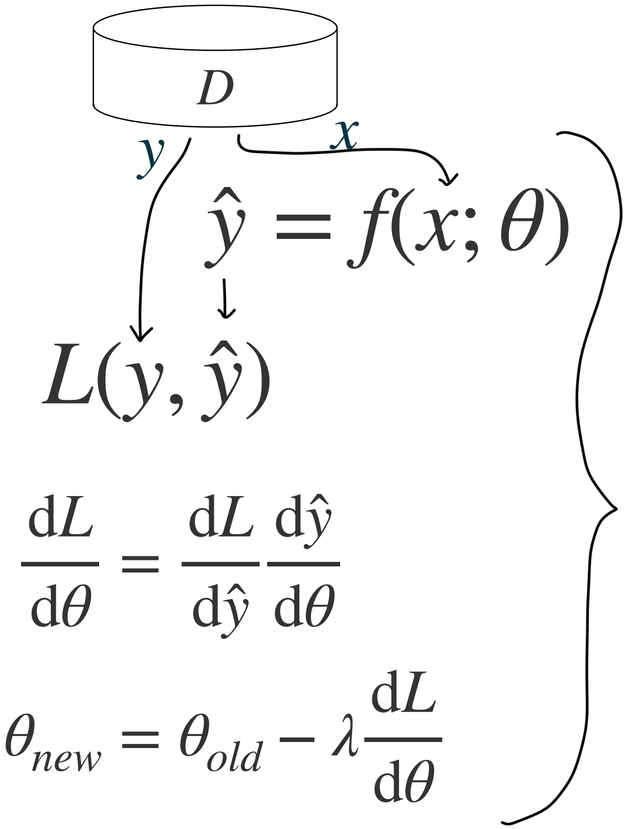


具体例（最適化アルゴリズム）

出欠コード
XXXX

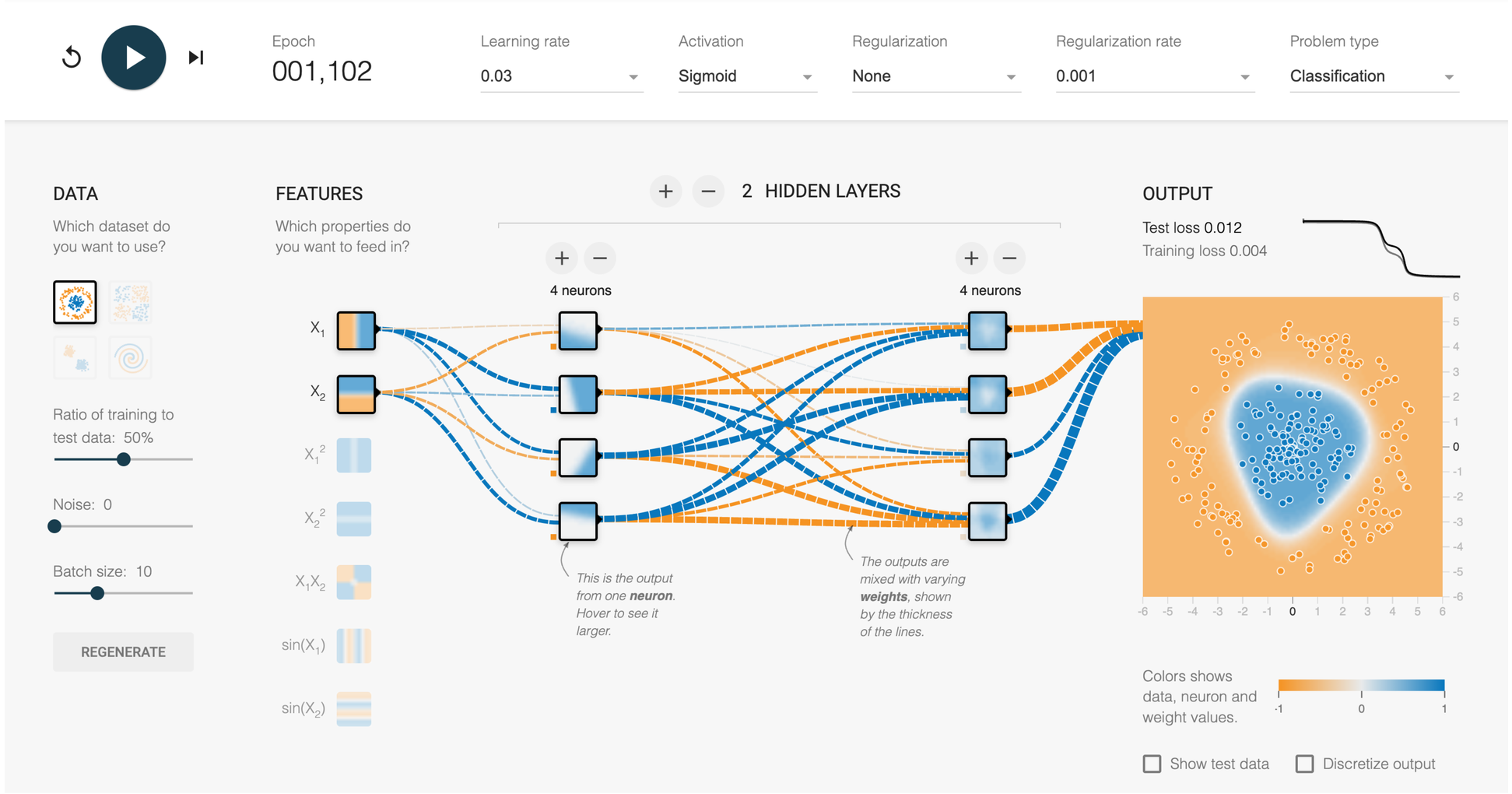
（確率的）勾配降下法

1. モデルクラス $\mathcal{F}$、損失関数L、データDを入力
   - Dは入力変数 x と予測対象 y を含む
2. データの一部をランダムに取得
3. L を最小化するように勾配方向に向けて
   少しだけ θ を更新
4. 上記2、3を繰り返す
   - 2の抽出を非復元で行って全データを学習した
     1回分をepochと呼び、epochをさらに繰り返す
   - 訓練サンプルとは別に取っておいた検証用サンプル
     での精度が悪化し始めたら終了
5. パラメタ θ を出力



※ ベクトルでの微分は要素ごとの微分のベクトル

多層パーセプトロンの学習例



A Neural Network Playground

（参考）機械学習の理論保証

おそらくだいたい正しい（PAC）ことを保証する

- 0-1損失を用いた二値判別を$d-1$次元線形モデルで行うとき、
- モデルクラス$\mathcal{F}$の中で最善のモデルに比べて高々ε程度しか悪くない
- ような結果をもたらすデータが出る確率が$1-\delta$以上、つまり
- $$p\left(\mathbb{E}[\ell(\hat{f};o)]-\min_{f\in\mathcal{F}}\mathbb{E}[\ell(f;o)]\leq\varepsilon\right)\geq 1-\delta$$
- となるのに必要なサンプルサイズ$N_{\mathcal{F}}(\varepsilon,\delta)$は、定数$C_1$、$C_2$が存在して、以下が成立
- $$C_1\frac{d+\log(1/\delta)}{\varepsilon^2}\leq N_{\mathcal{F}}(\varepsilon,\delta)\leq C_2\frac{d+\log(1/\delta)}{\varepsilon^2}$$