AAAI2019読み会
「特徴量選択を教師付き学習する！」

Human-in-the-Loop Feature Selection

Alvaro Correia, Freddy Lecue

読み手：Hisashi Kashima (KU/AIP)

論文の概要：
特徴量選択を学習する問題を考えた

- 背景：予測モデルに用いる特徴量の選択は、学習効率・精度の向上だけでなく、モデルや予測の解釈にも有効
- 貢献：新たな問題設定
  - 訓練データにおいて、入力（特徴量）と出力（ラベル）に加えて、どの特徴量が重要かという補助情報が与えられている
  - 出力を予測するだけでなく、特徴量を選択するモデルを学習する

予測タスクの例：
各地区の不動産価格（数値）の予測

- ボストンの各地区における不動産の平均価格データ

犯罪率	酸化窒素濃度	部屋数	1940年以前築	高速へのアクセス	固定資産税率	教師数と生徒数の比	有色人種の率	社会的地位の低い	価格
0.00632	0.538	6.575	65.2	1	296	15.3	396.9	4.98	24
0.02731	0.469	6.421	78.9	2	242	17.8	396.9	9.14	21.6
0.02729	0.469	7.185	61.1	2	242	17.8	392.83	4.03	34.7
0.03237	0.458	6.998	45.8	3	222	18.7	394.63	2.94	33.4
0.06905	0.458	7.147	54.2	3	222	18.7	396.9	5.33	36.2
0.02985	0.458	6.43	58.7	3	222	18.7	394.12	5.21	28.7
0.08829	0.524	6.012	66.6	5	311	15.2	395.6	12.43	22.9
0.14455	0.524	6.172	96.1	5	311	15.2	396.9	19.15	27.1
0.21124	0.524	5.631	100	5	311	15.2	386.63	29.93	16.5
…	…	…	…	…	…	…	…	…	…

入力 $\mathbf{x}$ → 出力 y

  - 犯罪率や部屋数など9個の変数から価格を予測（予測）したい

- 特徴選択の問題：価格に影響する変数は何か？
  - 特に特徴量が多い（$\mathbf{x}$の次元が高い）と大変

問題設定：
訓練データの各事例において、重要な特徴も教示される

（入力・出力に加えて）

- 入力：訓練データ集合 $\{(\mathbf{x}_i, \mathbf{q}_i, y_i)\}_i$ （通常は$\{(\mathbf{x}_i, ~~\mathbf{q}_i~~, y_i)\}_i$）
  - $\mathbf{x}_i \in \mathbb{R}^D$：$i$番目の例の入力特徴ベクトル
  - $y_i \in \{0,1\}$： i番目の例の出力ラベル

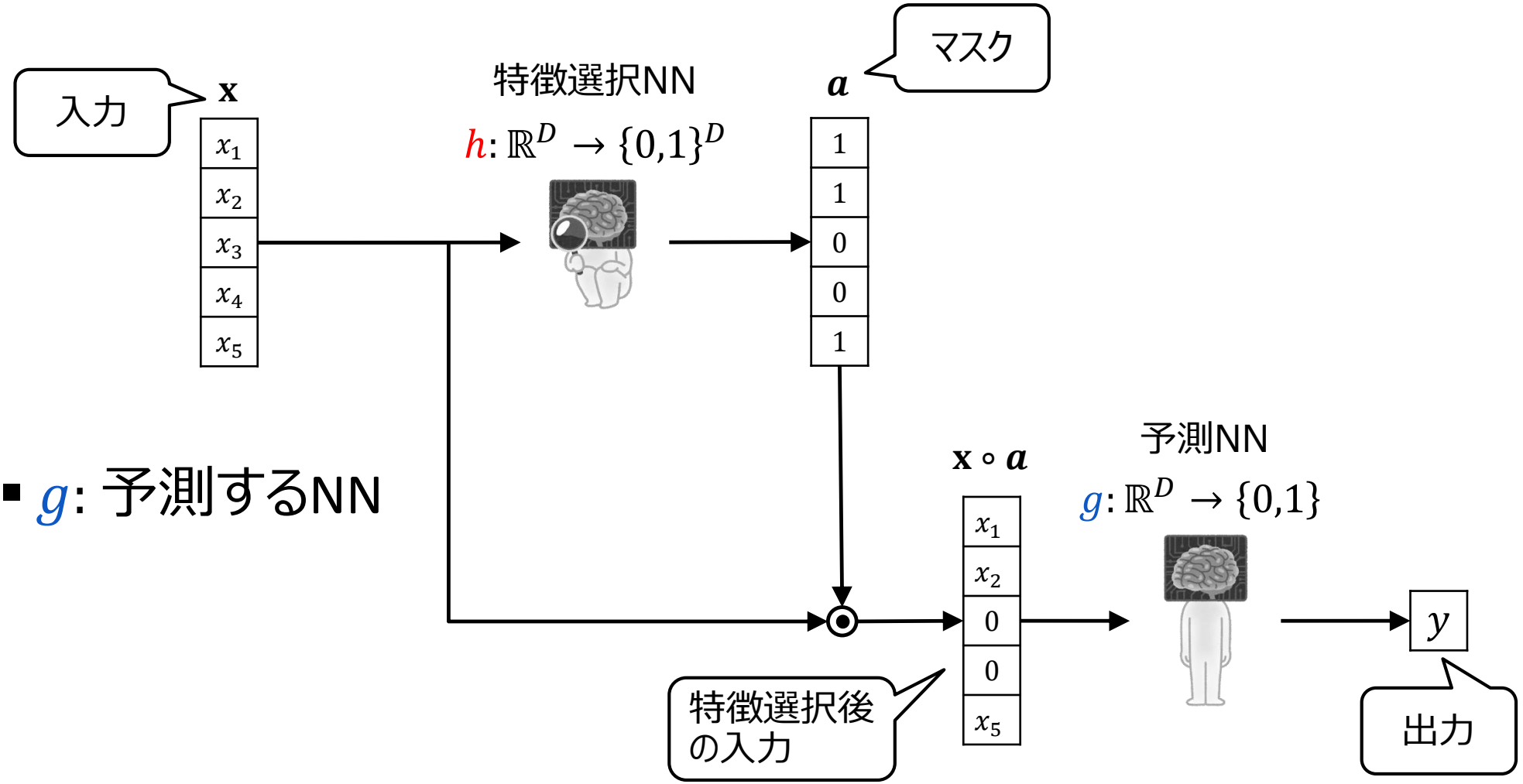
  } ここまでは いつもと同じ

  - $\mathbf{q}_i \in \{0,1\}^D$：$i$番目の例で、どの特徴が重要かを表すベクトル
    - 各次元の値は1だと重要な特徴、0だと不要を意味する

- 出力：予測モデル $f: \mathbb{R}^D \to \{0,1\}$
  - こちらは通常と同じ

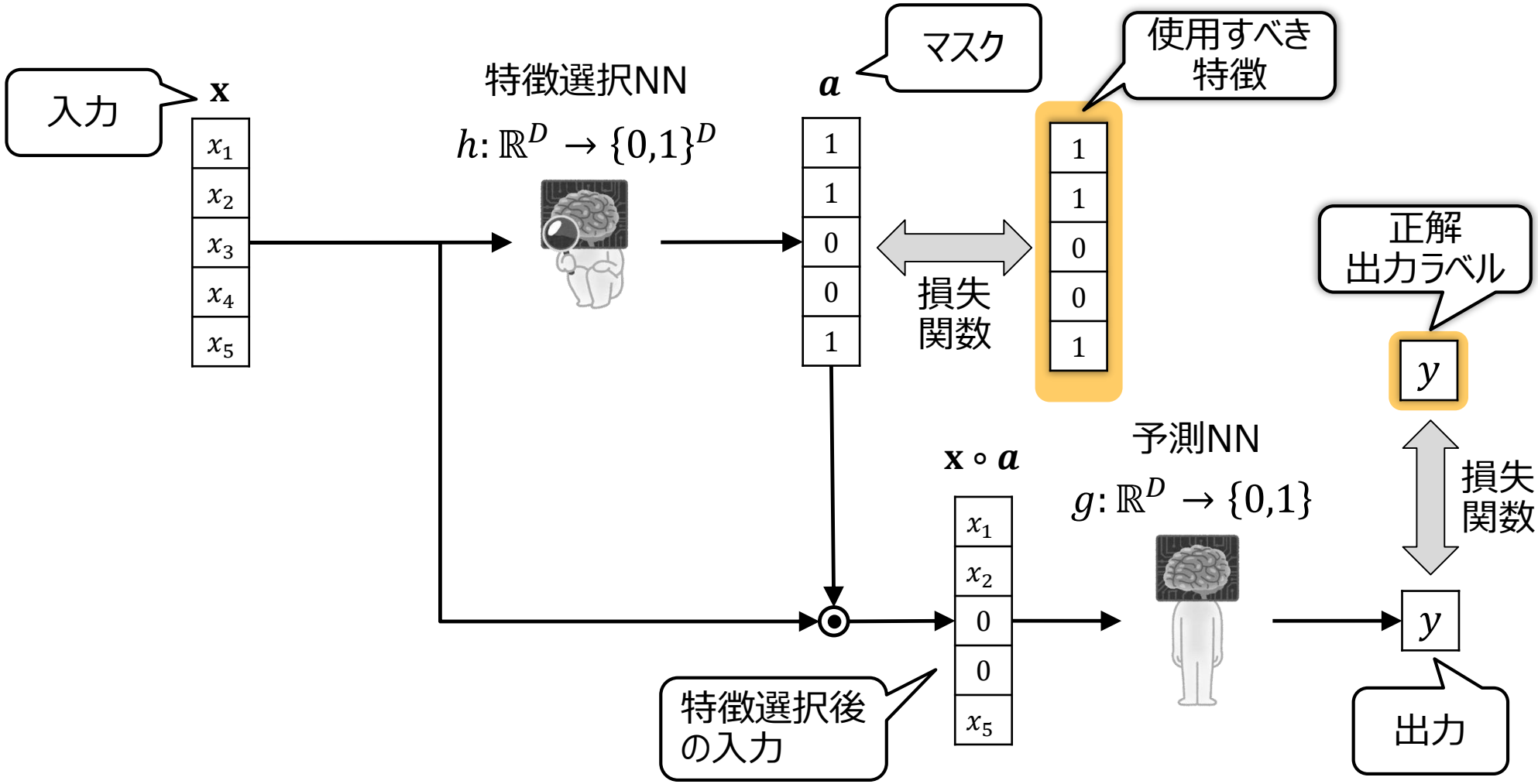
モデル：
特徴を選択するモデル＋選択された特徴で予測するモデル

- h: 特徴選択するニューラルネットワークモデル（NN）



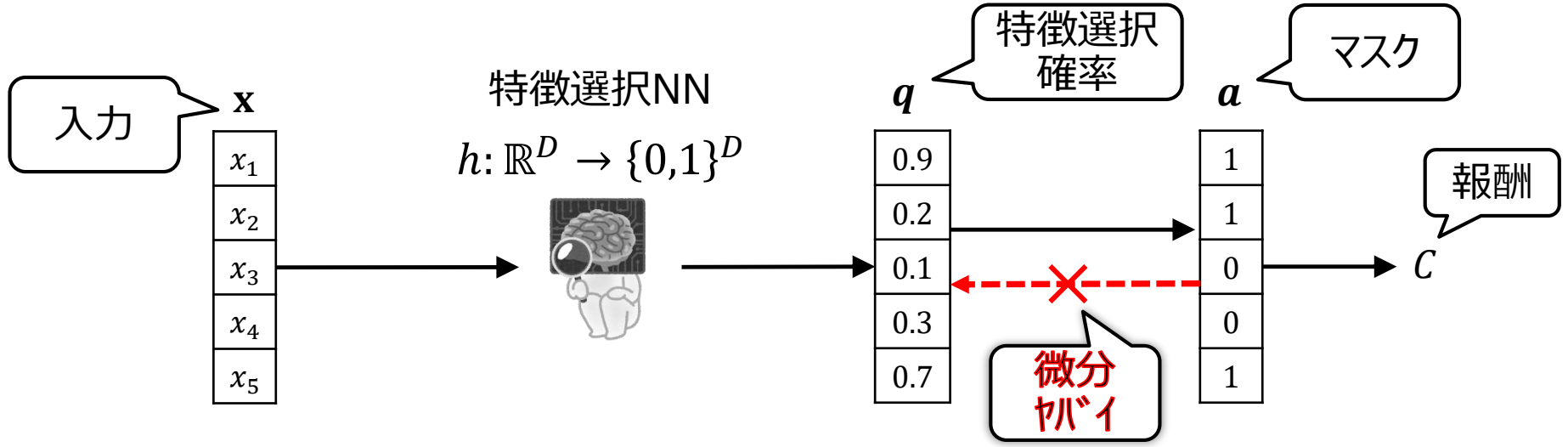
- g: 予測するNN

2種類の教師信号：
正解出力ラベルと併せて使用すべき特徴も与えられる



技術的な問題：
特徴選択が確率的な閾値処理なので誤差逆伝播困難

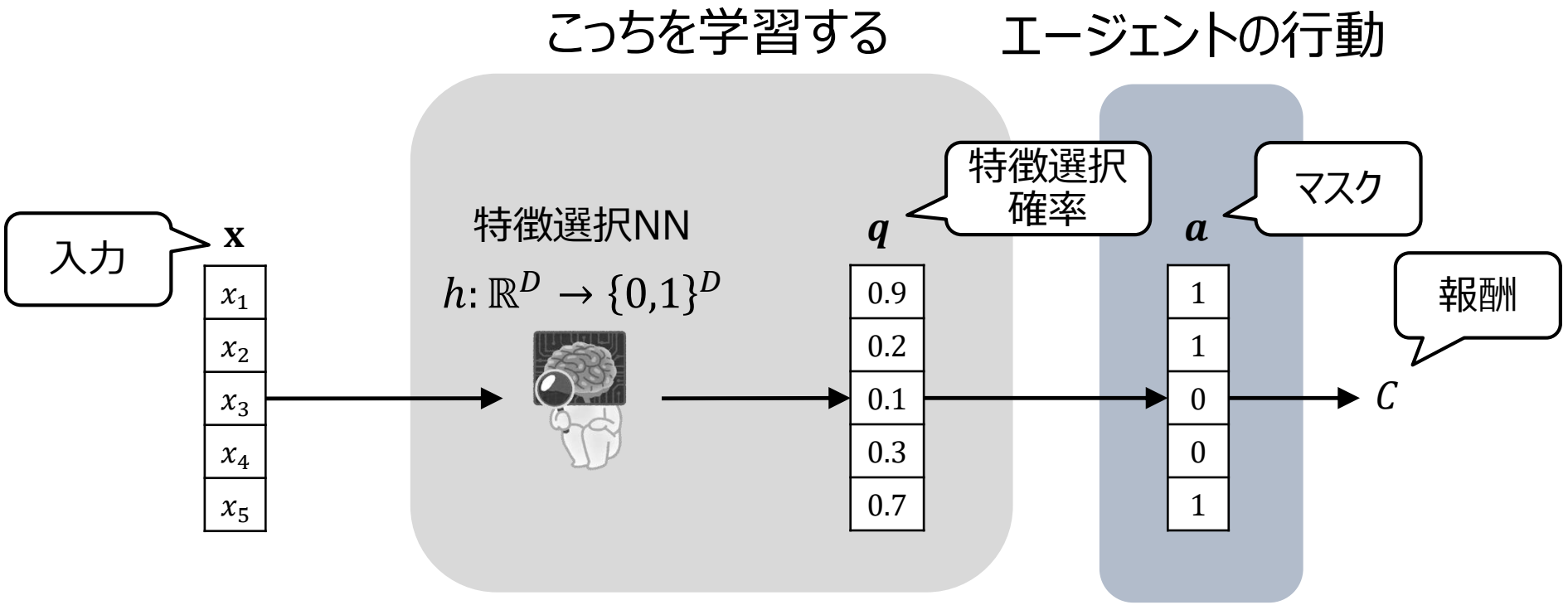
- 実際には特徴選択確率にもとづくサンプリングでマスクが決まる



- 問題点：訓練時の誤差逆伝播で微分が意味をもたない
  （閾値的な処理なので）
- 解決法（2案）：
  1. 即時報酬強化学習（REINFORCE）：特徴選択の報酬を最大化
  2. 滑らかにする（Gumbel softmax）

解決案①：
即時報酬強化学習にする

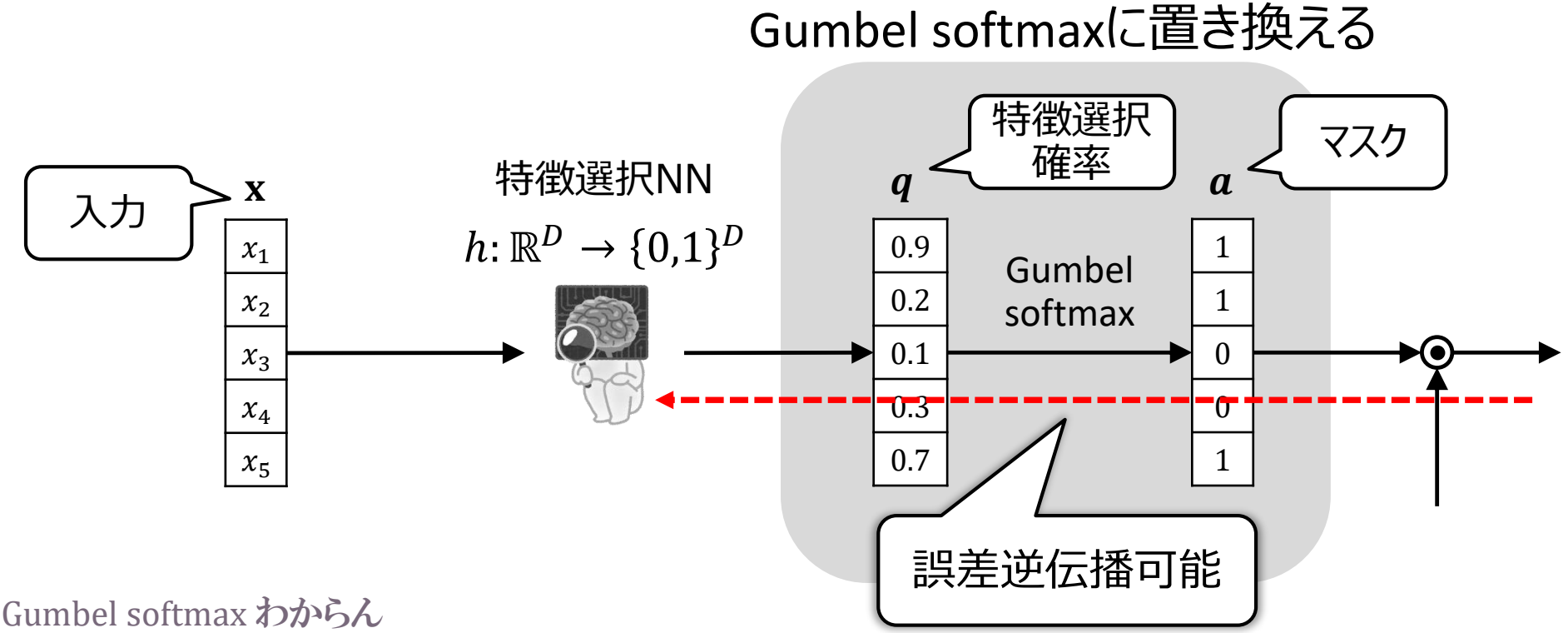
- 即時報酬強化学習（REINFORCE）として考える
- 特徴選択を行動として損失関数（報酬）を最小化（最大化）



※ 予測器と特徴選択器を並列に学習するのかend-to-endでやるのかはちょっとわかんなかった（前者の気がする）

解決案②：
滑らかにして誤差逆伝播（微分）できるようにする

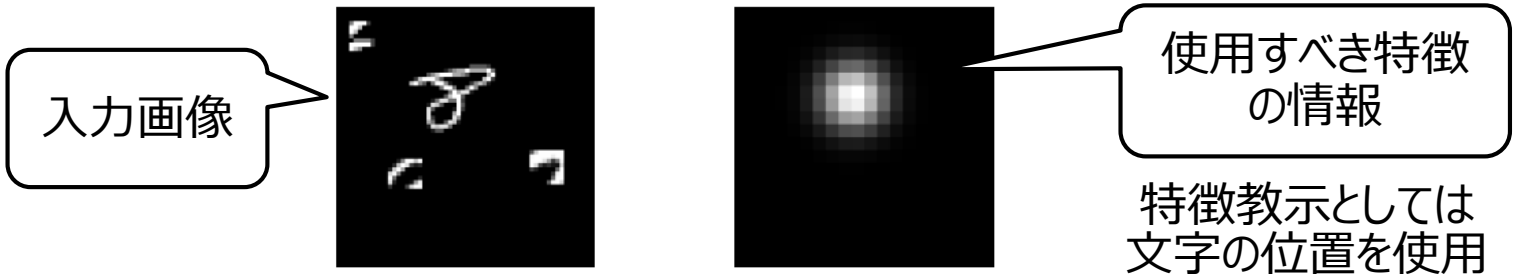
- Gumbel Softmaxで滑らか近似
  - 離散分布のサンプルの微分可能な近似表現（極限で一致）
- 誤差逆伝播できるようになる



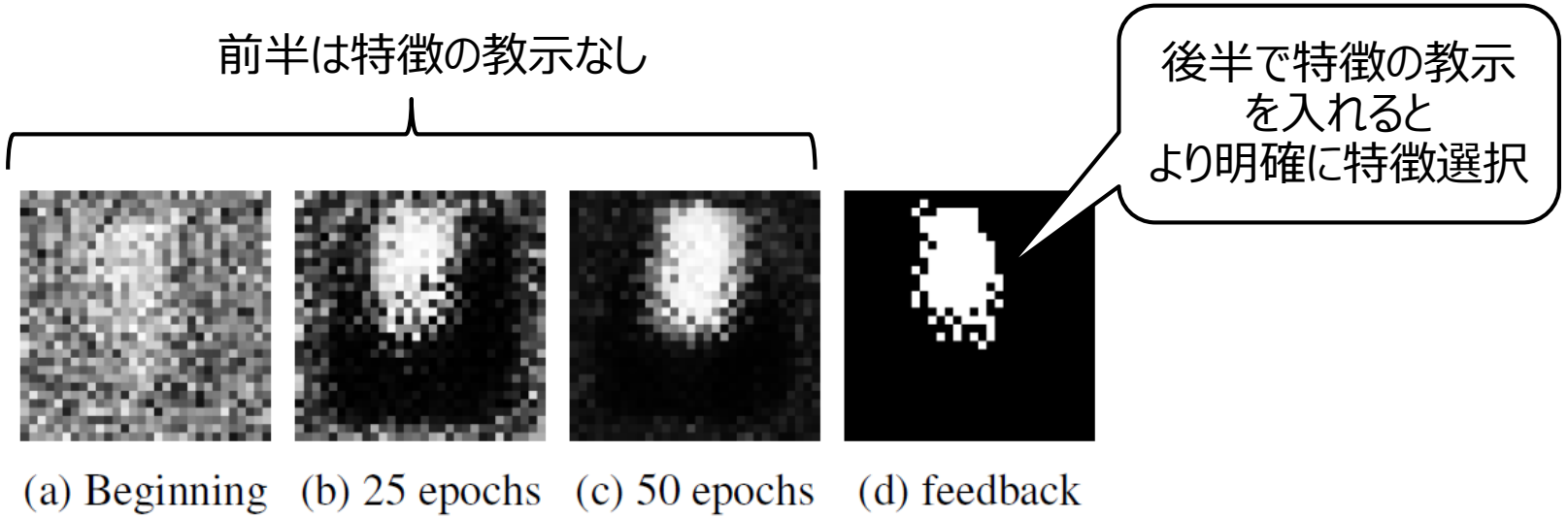
※ Gumbel softmax わからん

数値例①：
特徴の教示によって特徴選択効果⇈

- 手書き文字データセット（MNIST）で実験



- 特徴の教示を入れると、より鮮明に特徴選択されるようになる



数値例②：
予測精度では強化学習ベースのほうがよさそう

- 全体的には強化学習ベースのほうが予測精度は高そう

特徴教示なし

特徴選択の損失関数（cos or MSE）

文字認識データ

Table 1: Estimators Impact on Accuracy (%).

Feedback / Estimator	SF	PD
Before Feedback	85.35	**85.70**
Cosine Feedback	**92.30**	88.40
MSE Feedback	91.16	89.61

強化学習

Gumbel

プロジェクトリスク予測データ

Table 4: Accuracy (%) on PRC Test Set with Each Estimator.

Feedback / Estimator	SF	PD
Before Feedback	29.53	**29.99**
Cosine Feedback	**82.49**	77.51
MSE Feedback	80.11	78.44

- では、Gumbel softmaxは要らない??
  - 特徴の教示が$\{0,1\}$でなくより細かいレベル（$\{0,1,\ldots,K\}$）に分かれている場合にも自然に拡張可能

論文の概要：
特徴量選択を学習する問題を考えた

- 背景：予測モデルに用いる特徴量の選択は、学習効率・精度の向上だけでなく、モデルや予測の解釈にも有効
- 貢献：新たな問題設定
  - 訓練データにおいて、入力（特徴量）と出力（ラベル）に加えて、どの特徴量が重要かという補助情報が与えられている
  - 出力を予測するだけでなく、特徴量を選択するモデルを学習する