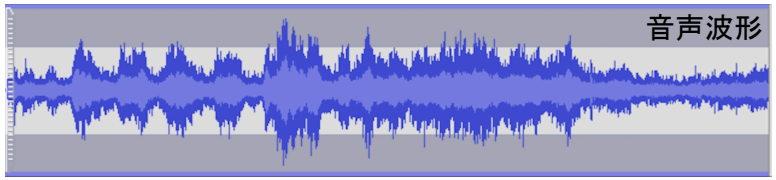
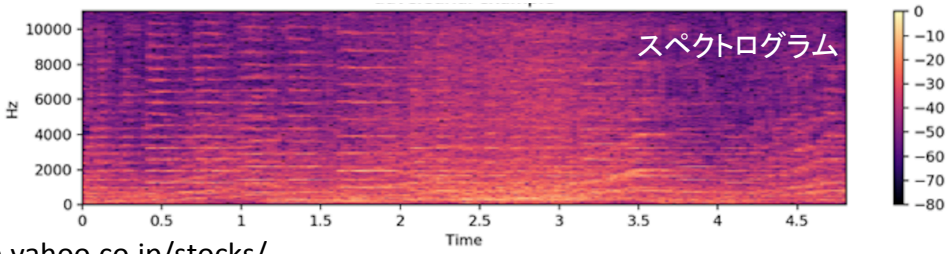
機械学習(10)
ニューラルネットワーク2 RNNとGAN

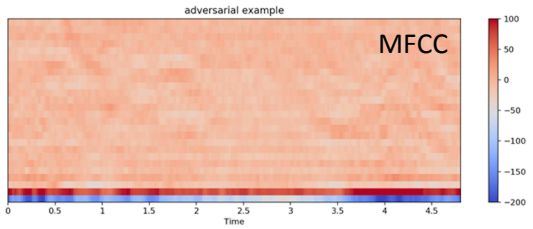
情報科学類　佐久間　淳

時系列データ

- 時間に連れて変化する観測値
  - $x_1, x_2, \ldots, x_t, \ldots$
  - 実数値でも離散値でも
  - x_tはそれ以前のデータ(x_{t-1}, x_{t-2},…)に依存することを想定
- 株価
- テキスト:「文書における特徴ベクトルの要素は文書中の単語です」
- 音声







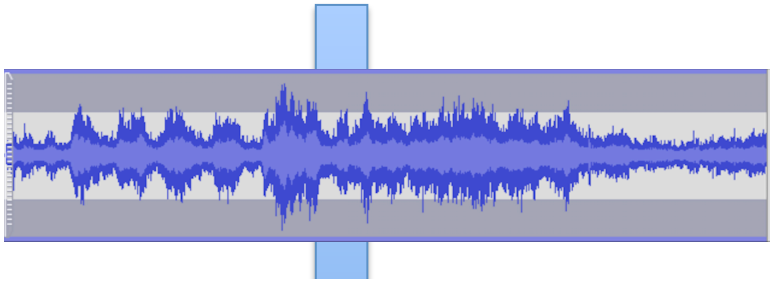
https://stocks.finance.yahoo.co.jp/stocks/

時系列データにおける予測(音声認識)

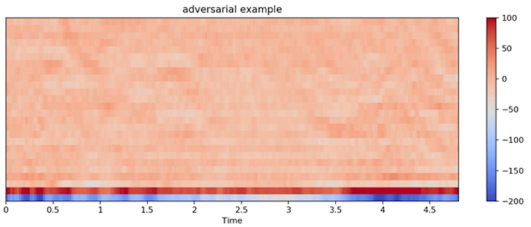
- 訓練データ (x_1,y_1), (x_2,y_2),…(x_t,y_t),…
  - X_1: “Open the front door”という発声の音声波形
  - y_1: “Open the front door”
- 訓練データ集合でxからyへの写像を学習
- テスト時 :
  - x: 未知の音声波形
  - y=f(x): “Turn on the light”

ニューラルネットワークによる音声認識

音声データ



特徴ベクトル
(MFCC)



ニューラルネットワーク

文字生起確率

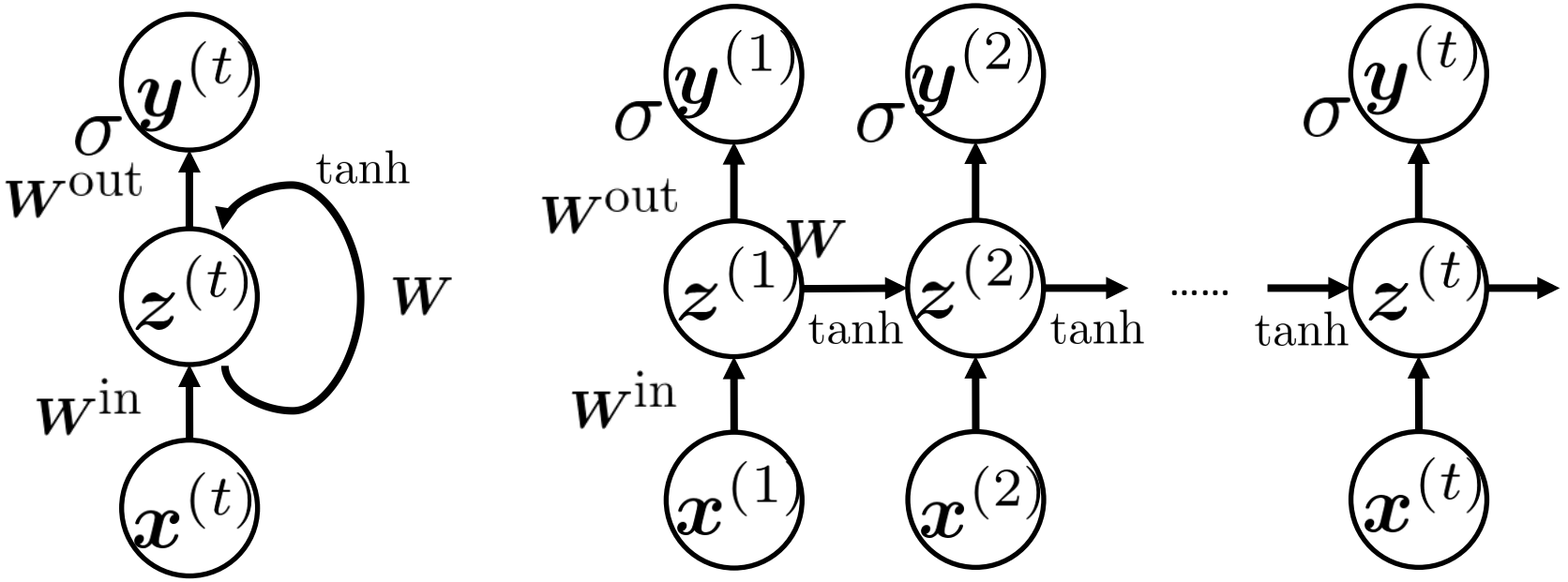
sh ix hv eh dcl jh ih dcl d ah kcl k

言語モデル

テキスト

She had your dark…

リカレントNNの構造と順伝播(RNN)

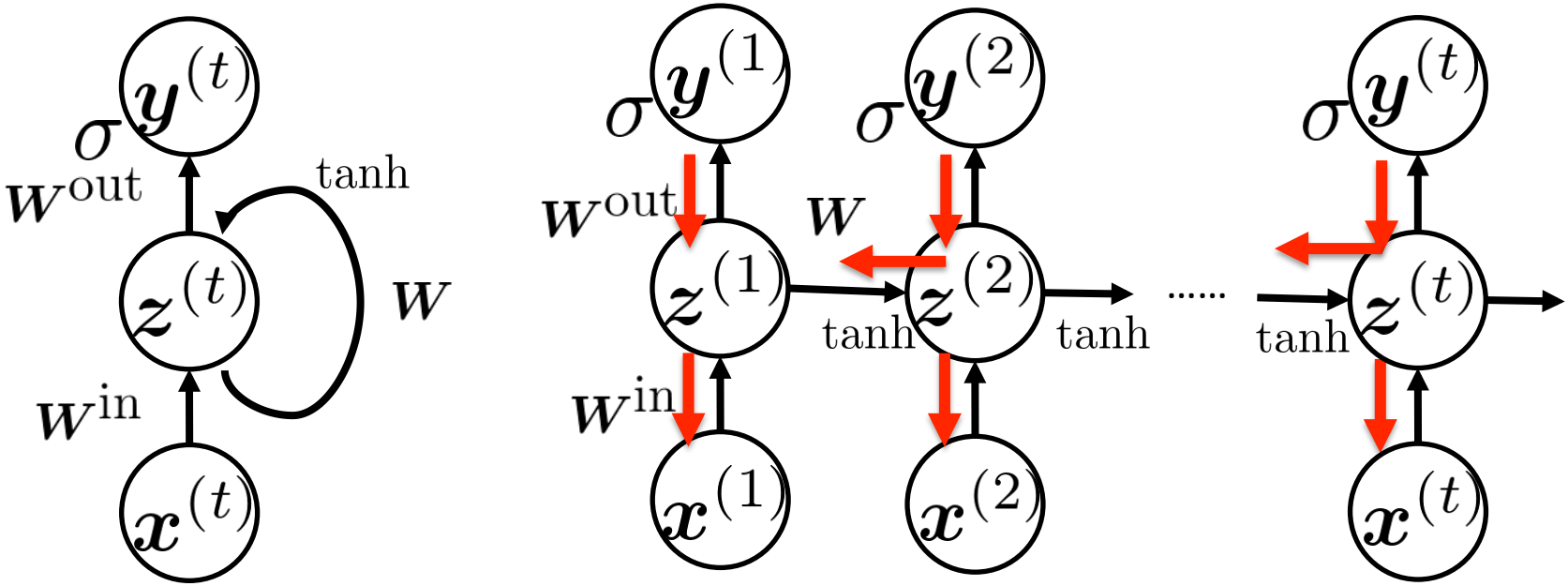


- 各時刻tにおいて入力 $x^{(t)}$ を受け取り、出力 $y^{(t)}$ を返す
- 時刻tにおける中間層の第jノードへの入力 $u_j^{(t)}$ は、時刻tの任意の入力 $x_i^{(t)}$ と時刻t-1の中間層の任意のノード(第j'ノード)の出力 $z_{j'}^{(t-1)}$

$$u_j^{(t)} = \sum_i w_{ij}^{(\text{in})} x_i^t + \sum_j w_{jj} z_{j'}^{(t-1)}$$

- 時刻tにおける中間層の第jノードの出力 $z_j^{(t)}$ $z_j^{(t)} = f(u_j^{(t)})$
- 時刻tにおける出力層の出力 $y_j^{(t)} = g(\sum_k w_{jk}^{\text{out}} z_k^{(t)})$

RNNの逆誤差伝播法



- 最終層の損失関数 $E(\boldsymbol{W}) = \sum_n \sum_t \sum_k d_{nk}^{(t)} \log y_k^{(t)}(\boldsymbol{x}_n; \boldsymbol{W})$

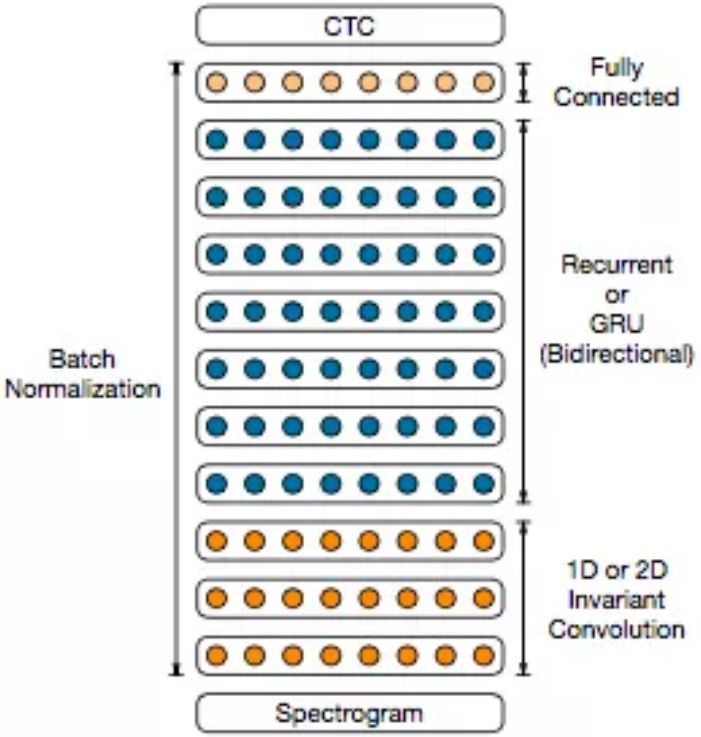
- BPTT (back propagation through the time)
  - RNNを時刻方向に展開し、閉路のないNNに変換
  - t=Tからはじめ, t=t-1, t-2,…,1と順番にδを計算し逆誤差伝播

応用

- ロボット制御
- 時系列予測
- 音声認識
- 音楽生成
- 手書き文字認識
- 行動認識
- 手話認識
- 翻訳
- テキスト読み上げ, etc.

Deep speech 2

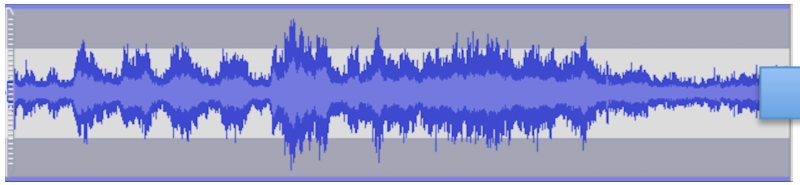
- End-to-endの音声認識
- 言語/環境に依存しない
  - 英語, 中国語
  - ノイズの有無
- 人間の音声認識を超える精を達成

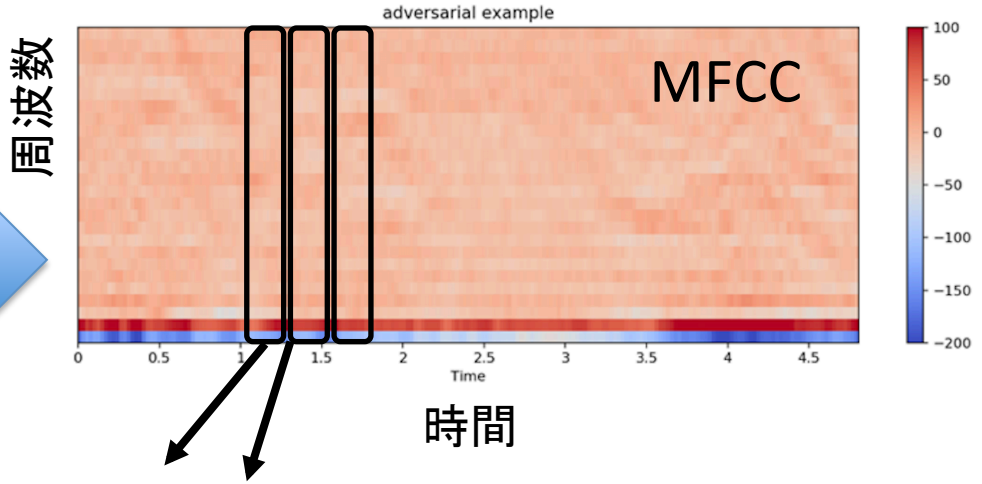


Amodei et al., Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin

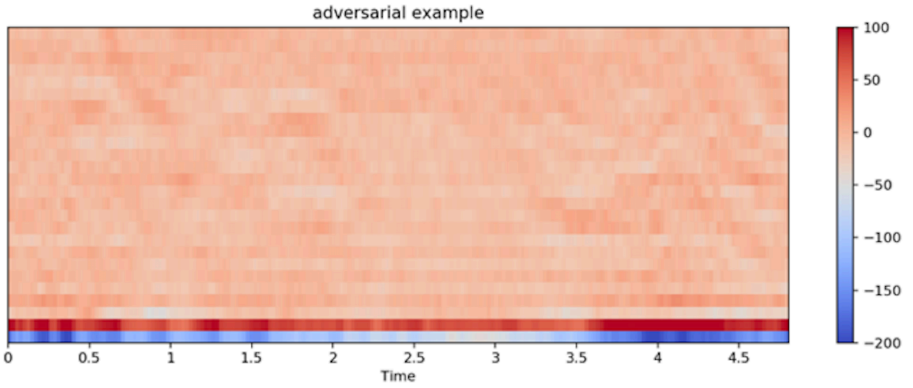
音声認識における畳み込み層

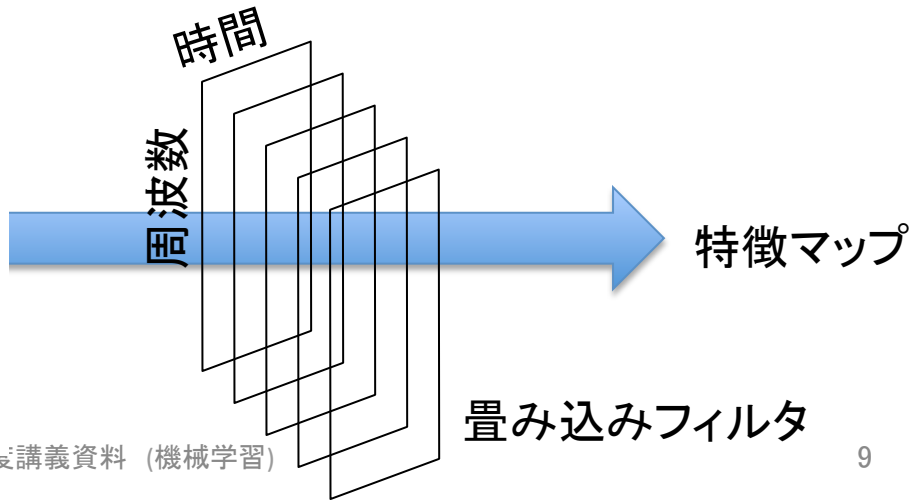
音声波形



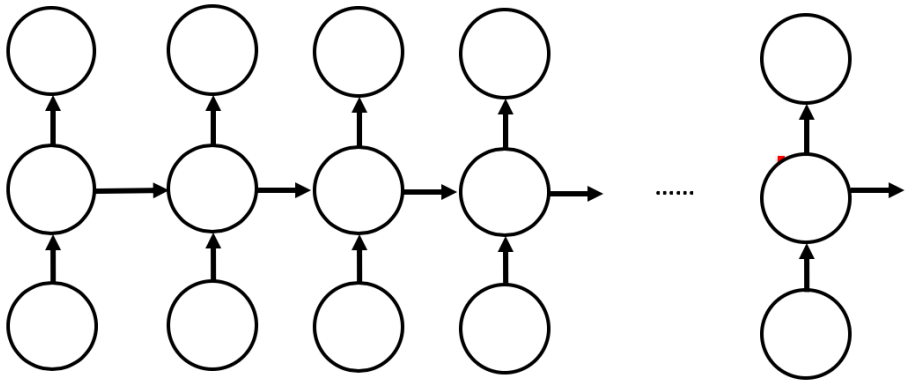


特徴ベクトル $x^1, x^2, \ldots, x^t, \ldots$





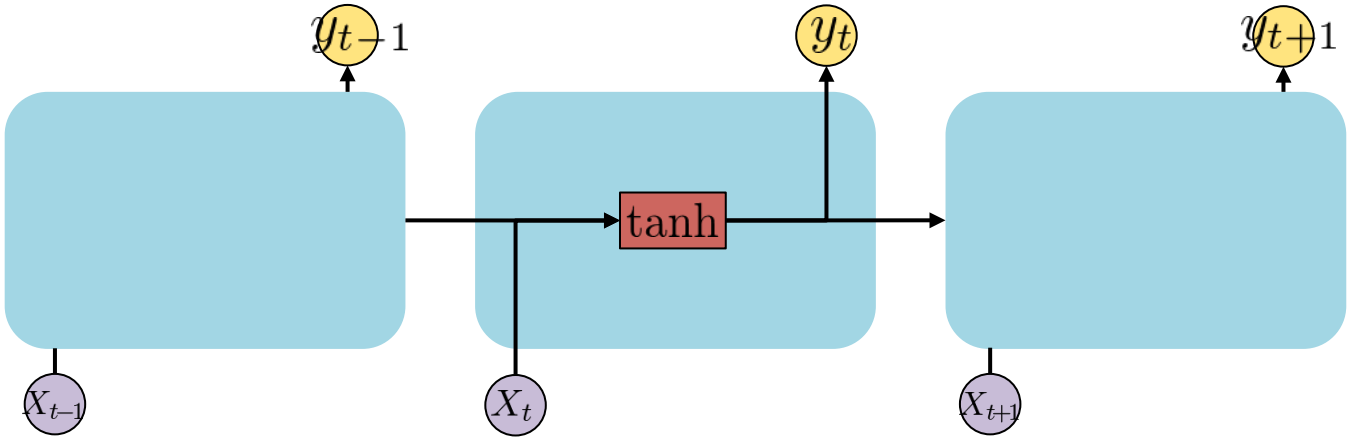
RNNにおける勾配消失



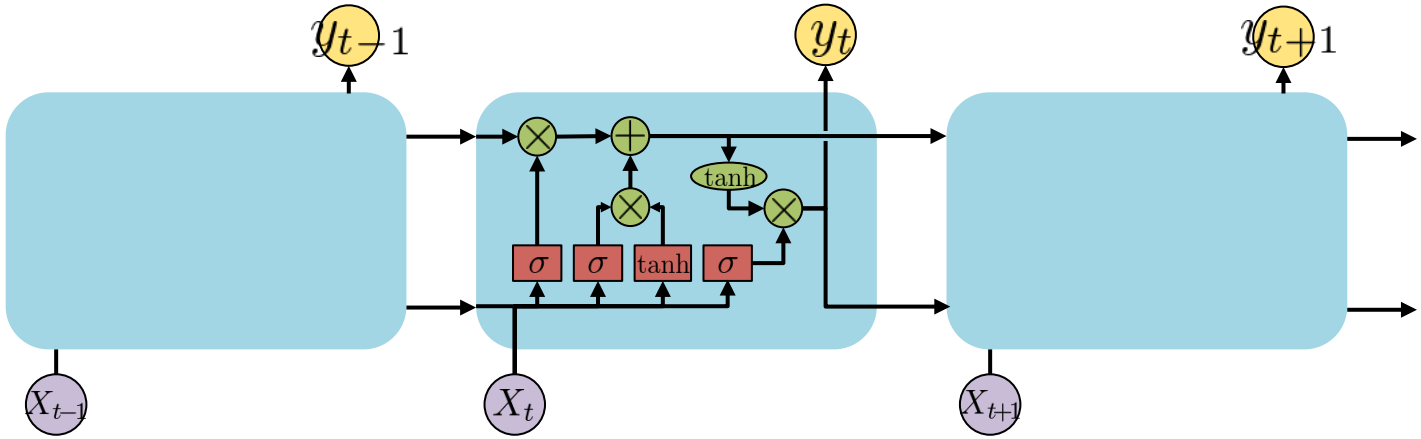
- RNN
  - 原理的には$y^{(t)}$の出力に$x^{(t)}$, $x^{(t-1)}$,…, $x^{(1)}$が全て関与
  - 実際には、遠く過去に遡った入力は出力にほぼ関与できない
- BPTTは時間をさかのぼる方向に勾配を伝播
  - (t=1に近づくにつれ)σ関数の勾配を多数回乗算
  - 勾配が「爆発」または「消滅」
  - 長期依存の系列の学習が正しく行われない

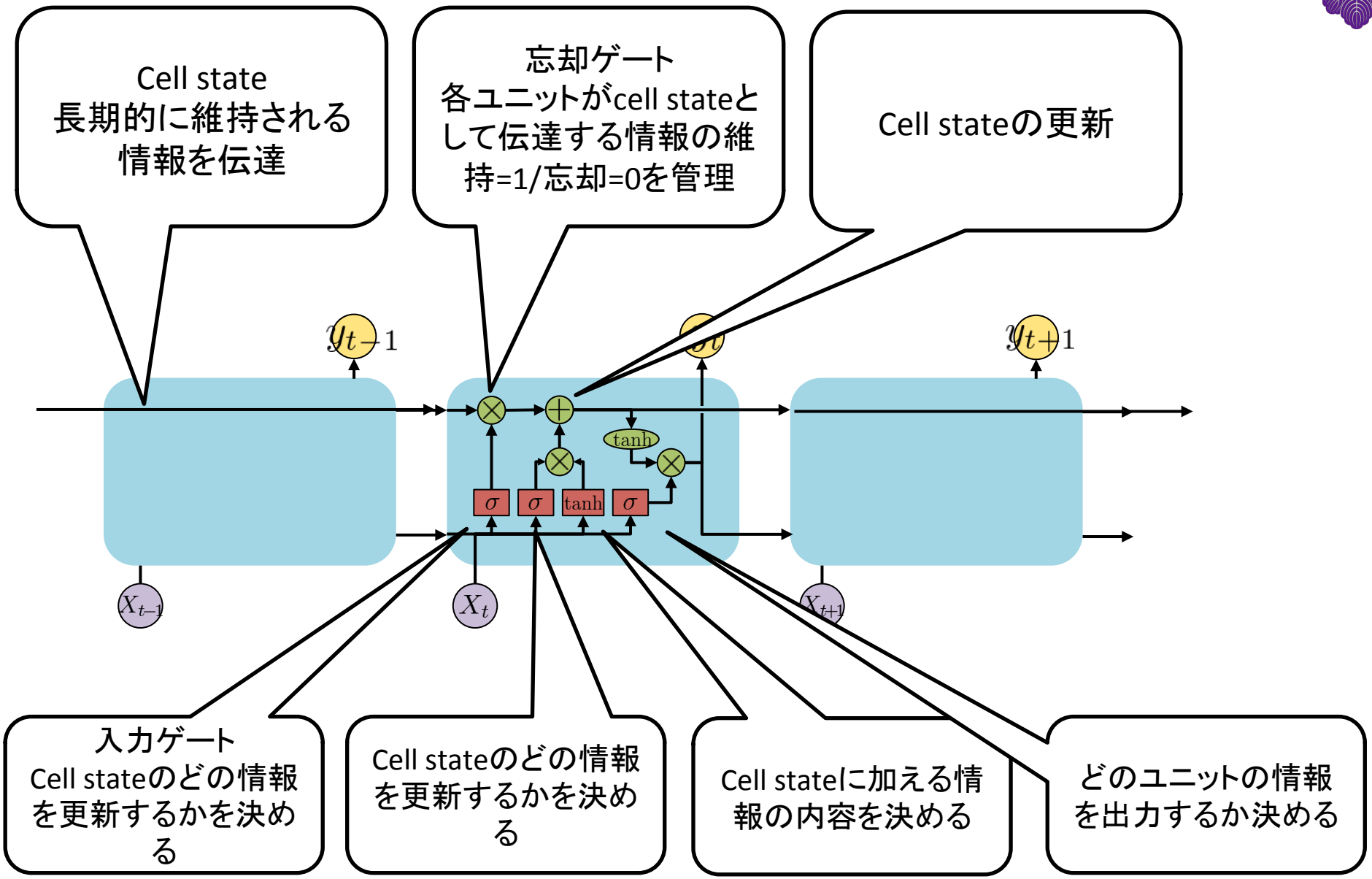
RNNとLSTM (long short term memory)

RNN



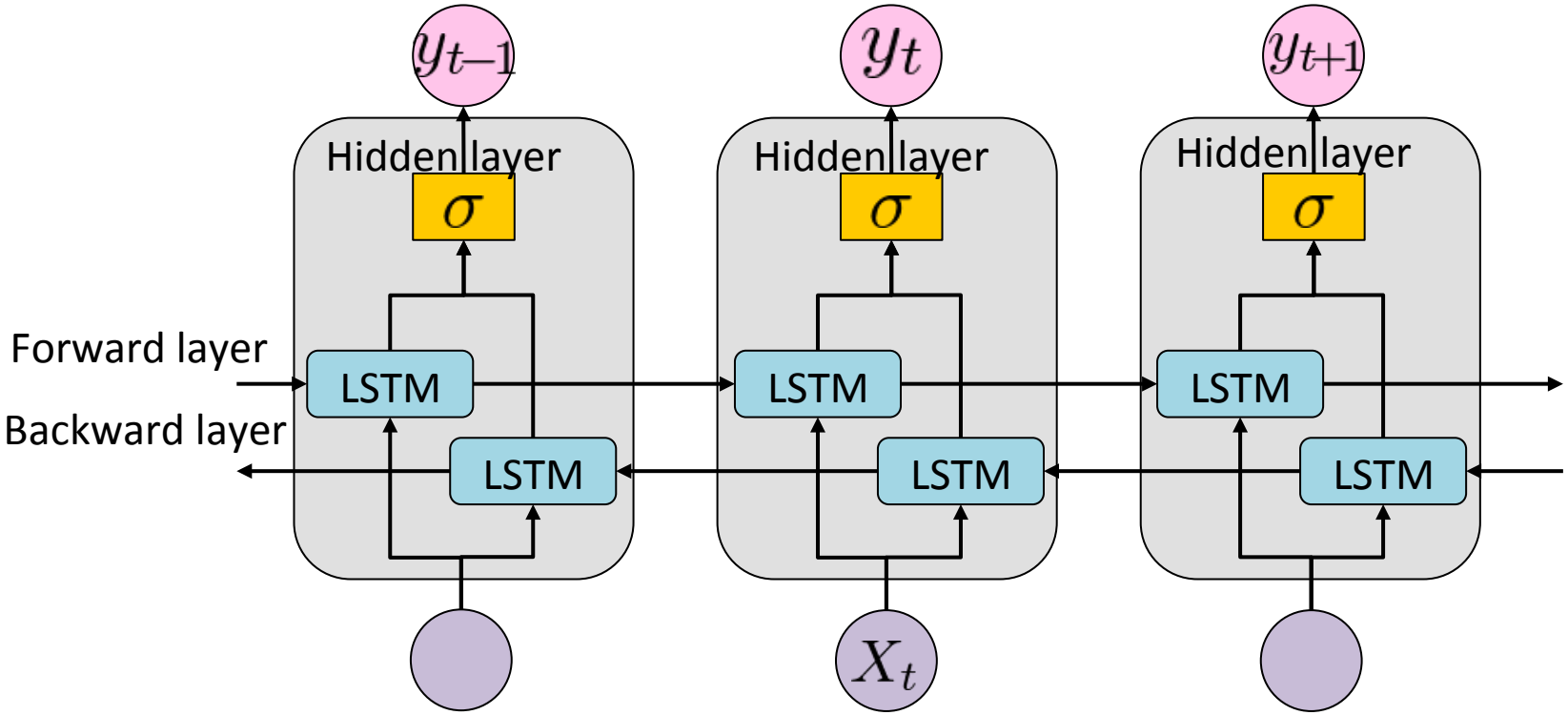
LSTM





Deep Speech 2で利用しているGRUはLSTMの簡略化されたバージョン

Bidirectional RNN

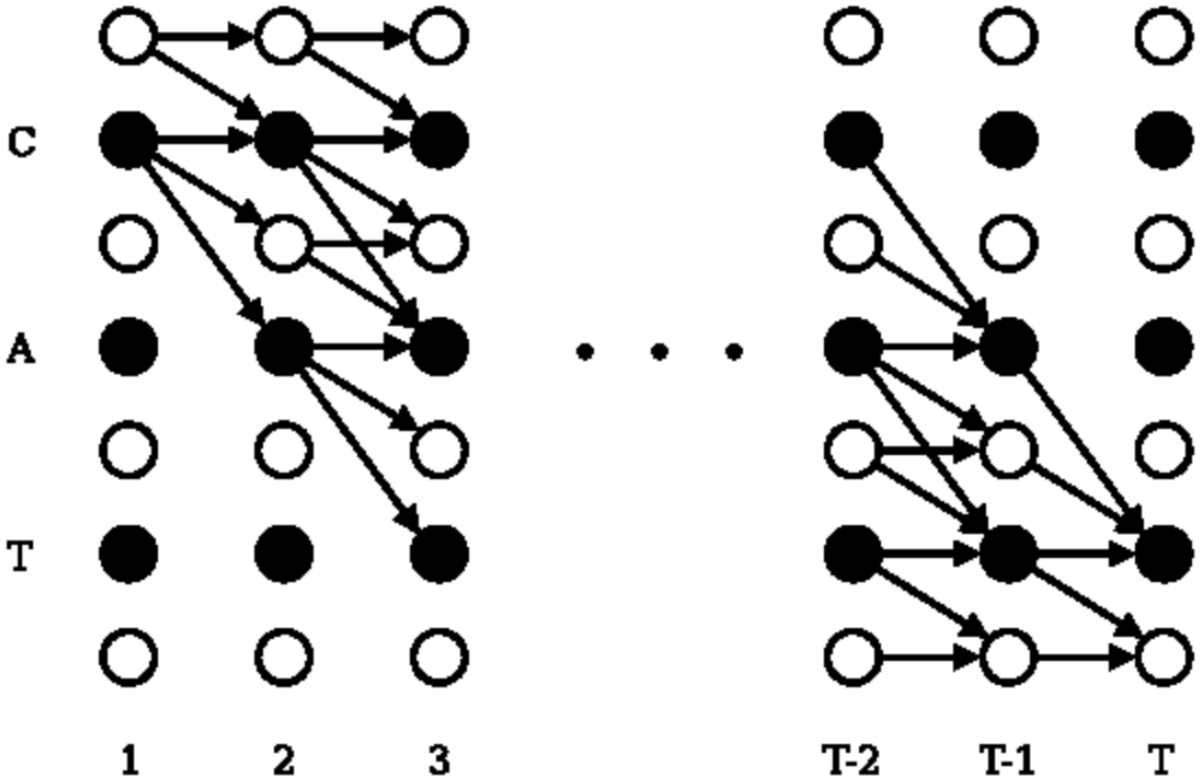


- RNN/LSTM 現在までの入力から未来の出力を予測
- Bidirectional RNN/LSTM 過去の入力と未来の入力から現在の出力を予測
- オフライン予測には使えるがオンライン予測には使えない
- Look-ahead convolution: あらかじめ指定したτステップ先までの情報のみ使う

Connectionist temporal classification (CTC)

- RNN/LSTMは1入力に対して1予測を出力
- RNN/LSTM入力と予測対象のフレームは一致しない
  - 音声認識器の入力音声(MFCC) のタイムフレーム(e.g., 20ms)と、認識結果(文字列)の一文字
- CTC: 入出力の系列長が異なる場合の損失関数
  - NNは時刻tの発声がある文字である確率を予測
    $\mathbf{y}^{(t)}=(\Pr(y^{(t)}=\text{“A”}|x^{(t)}), \Pr(y^{(t)}=\text{“B”}|x^{(t)}), \ldots, \Pr(y^{(t)}=\text{“Z”}|x^{(t)}))$
  - 予測$\mathbf{Y}=\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots \mathbf{y}^{(5)}$と正解”CAT”の損失=予測$\mathbf{Y}$から”CAT”が生成される確率$\Pr(\text{“CAT”}|\mathbf{Y})$

CTCの説明



- 空白文字を*とすると、CATを5出力で表す表現は
  - **CAT, *CCAT, *CAAT, *CATT, CC*AT, CCT*T, …
  - これら全てが**Y**から生成される確率を足し合わせ、これを損失関数とする

Graves et al., Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Deep Speech 2による音声認識結果

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

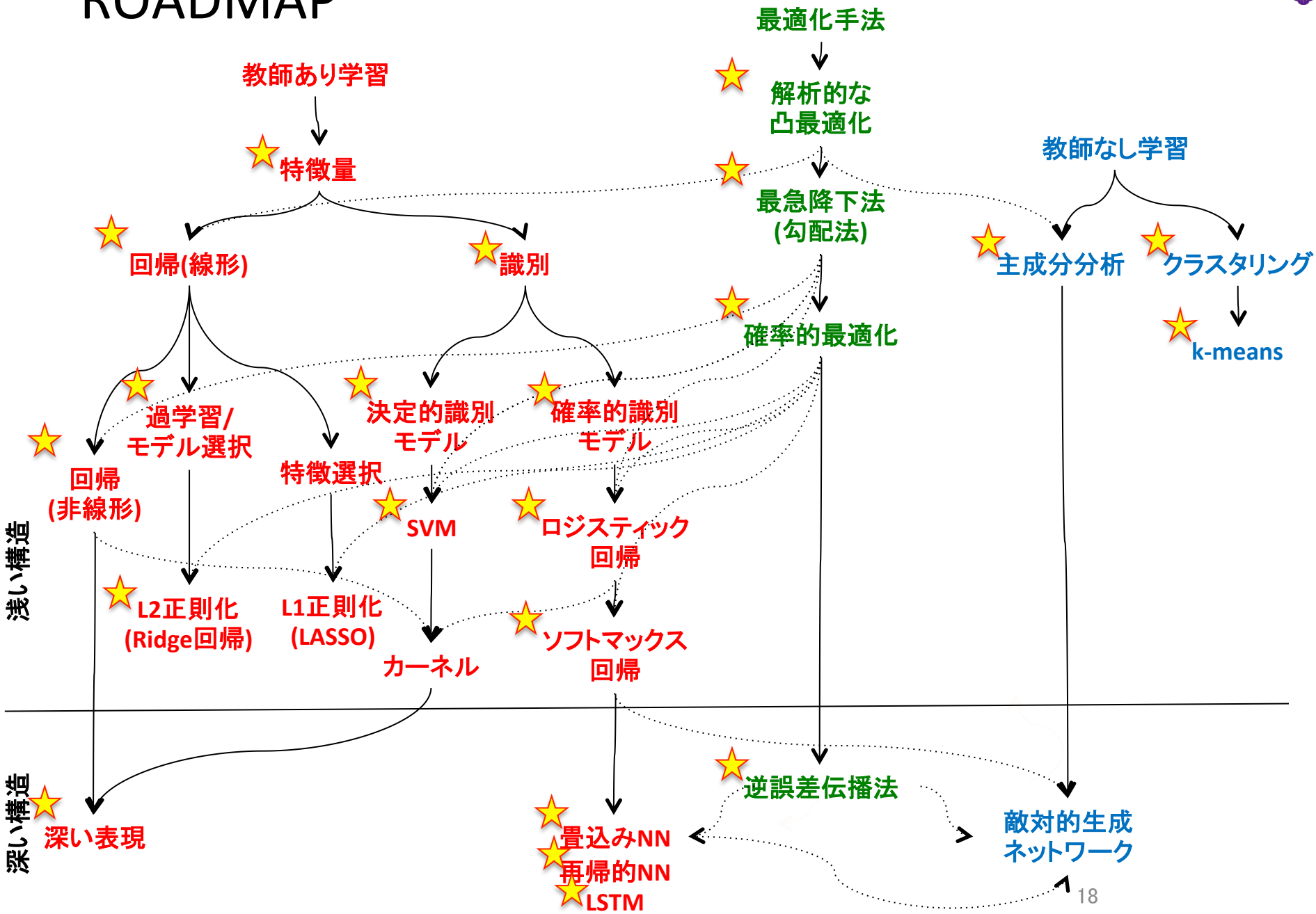
Accented Speech			
Test set	DS1	DS2	Human
VoxForge American-Canadian	15.01	7.55	4.85
VoxForge Commonwealth	28.46	13.56	8.15
VoxForge European	31.20	17.55	12.76
VoxForge Indian	45.35	22.44	22.15

Noisy Speech			
Test set	DS1	DS2	Human
CHiME eval clean	6.30	3.34	3.46
CHiME eval real	67.94	21.79	11.84
CHiME eval sim	80.27	45.05	31.33

RNNまとめ

- 時系列データを扱うNN: RNN, LSTM
- 時刻tの中間層出力が時刻t+1の中間層入力に追加される
- 順伝播は通常のNNと同様
- 逆誤差伝播は閉路のないNNに変換して行う
- 目標値の系列長と予測出力の系列長が一致するなら、通常の損失関数を使う
- 一致しないなら、CTCを使う

ROADMAP


最適化手法
教師あり学習
解析的な
凸最適化
特徴量
最急降下法
(勾配法)
教師なし学習
回帰(線形)
識別
主成分分析
クラスタリング
確率的最適化
k-means
過学習/
モデル選択
決定的識別
モデル
確率的識別
モデル
回帰
(非線形)
特徴選択
SVM
ロジスティック
回帰
浅い構造
L2正則化
(Ridge回帰)
L1正則化
(LASSO)
ソフトマックス
回帰
カーネル
逆誤差伝播法
深い構造
深い表現
畳込みNN
敵対的生成
ネットワーク
再帰的NN
LSTM

再掲

教師なし学習

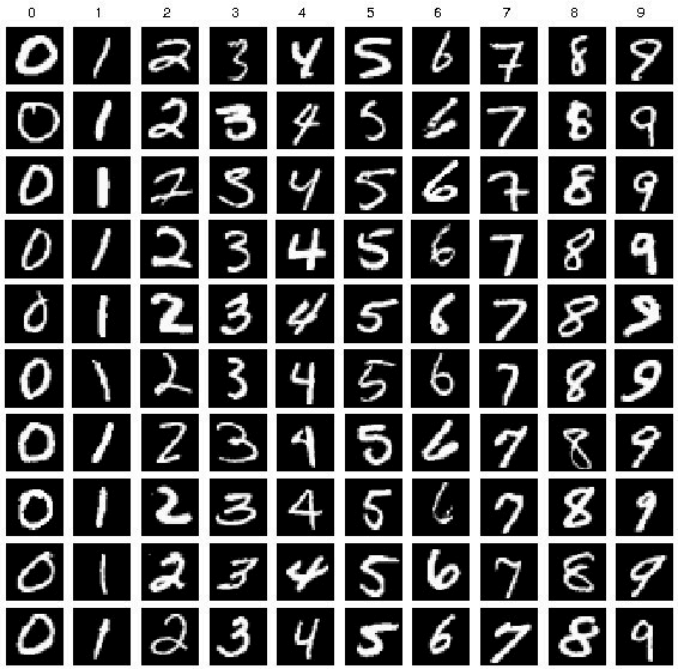
- これまでは…教師あり学習をやってきました
  - 訓練事例$\{(x_i, y_i)\}$から予測機f:X→Yを学習
  - 未知事例$\{x_j\}$のラベルy_jを予測
  - Yが…
    - 連続値Rなら回帰,
    - 条件付き確率[0,1]なら確率的識別モデル(e.g., ロジスティック回帰)
    - 離散値なら決定的識別モデルによる分類(k-NNとかSVMとか)
- 教師なし学習では事例$\{x_i\}$しか与えられない
  - 事例の集合からわかることを調べる．たとえば…
  - 潜在的な変数の推定(クラスタリング，PCA etc)
  - 潜在的なパターンの抽出(相関ルール etc)

$P(y|x)$ に興味アリ

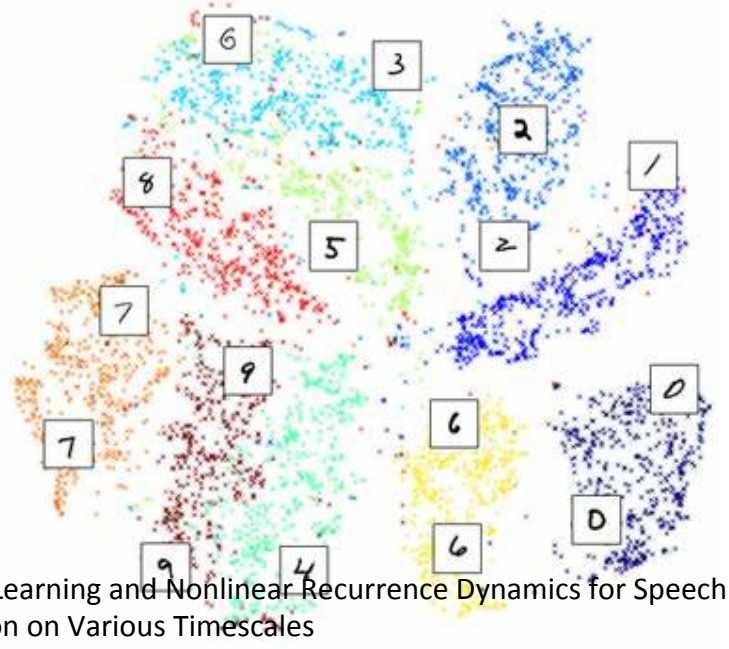
$P(x)$ に興味アリ

高次元で複雑な分布の推定

- データが基礎的な分布(e.g., 正規分布)に従う場合は、その観測値から分布が推定できる (e.g., 最尤推定)
- 高次元で複雑な分布に従うデータ(e.g., 数字画像)は、統計的手法で推定ができない



28*28次元モノクロ画像をt-SNEで2次元表示

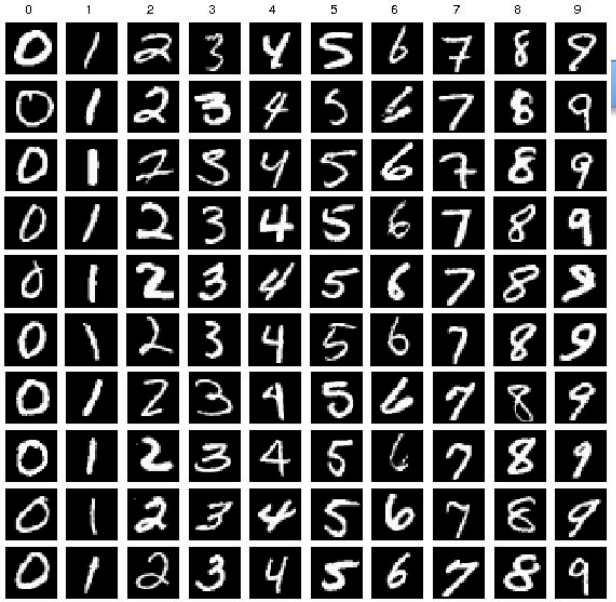


Manifold Learning and Nonlinear Recurrence Dynamics for Speech Emotion Recognition on Various Timescales

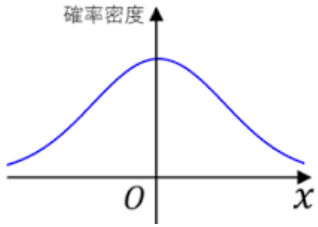
敵対的生成ネットワーク(GAN)

- 高次元で複雑な分布(e.g., 人間が書く”8”の画像の分布)の推定は難しい問題
- その代わり、その分布に従うサンプルを生成できるモデル(生成モデル)を学習することを考える

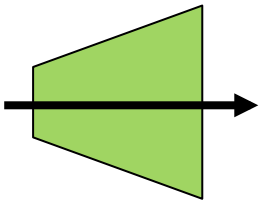
人間が生成した手書き数字画像



生成モデルの学習

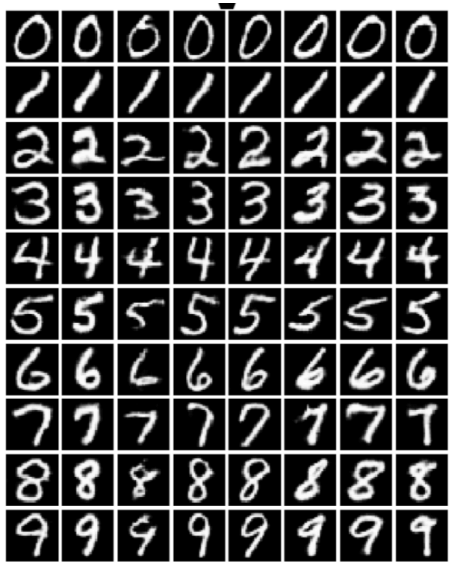


単純な乱数



NN

生成モデルが生成した手書き風数字画像



GANの応用

- インテリアデザイン、服飾デザイン
- 画像からの3次元モデル生成
- 加齢後の顔画像先生
- デジタル地図のスタイル変換
- セリフに合わせた顔画像先生
- ファッションモデルの画像生成
- 宇宙画像の生成（ダークマター研究のための重力レンズのシミュレーションに利用）
- ビデオゲームのステージ生成

Caption to picture

Figure 3. Example results by our proposed StackGAN, GAWWN [20], and GAN-INT-CLS [22] conditioned on text descriptions from CUB test set. GAWWN and GAN-INT-CLS generate 16 images for each text description, respectively. We select the best one for each of them to compare with our StackGAN.

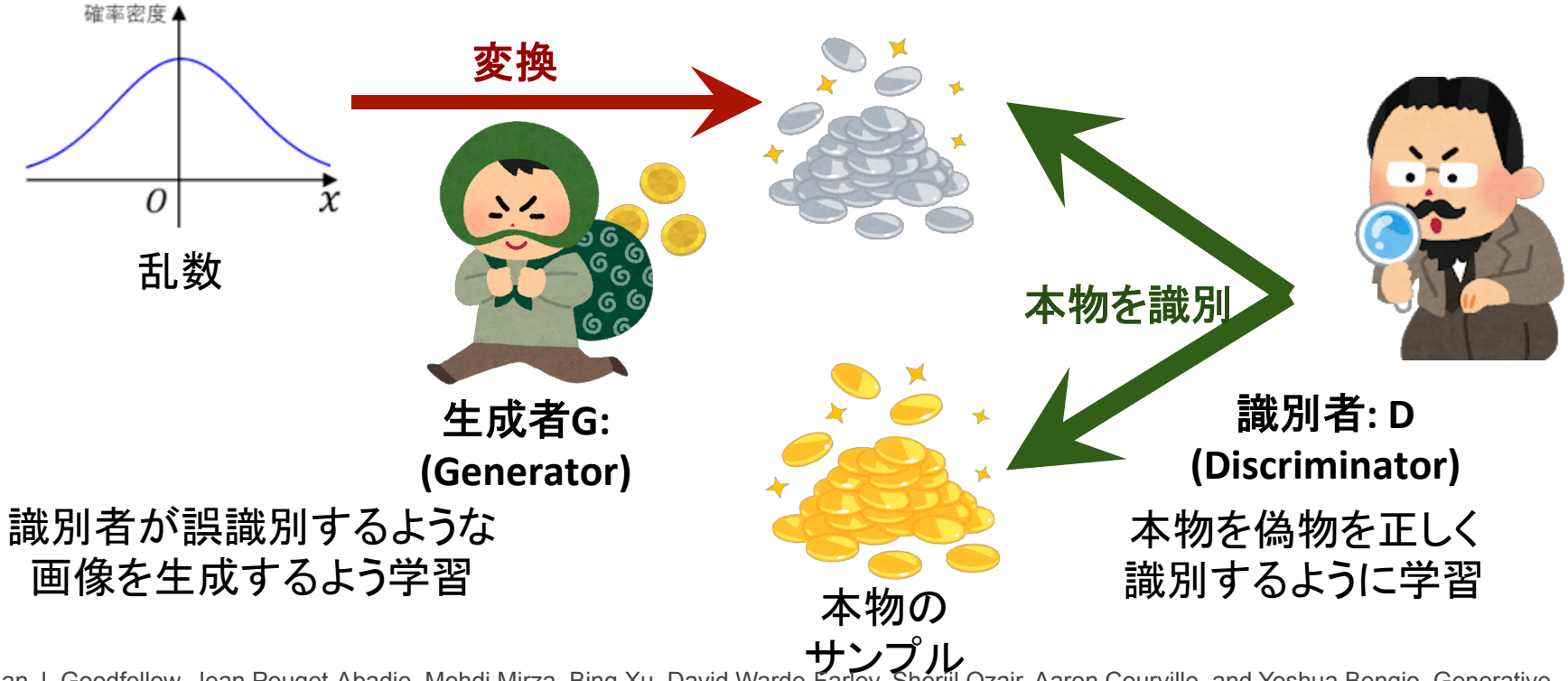
Zhang et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

From 漢字(HanZi) to 한글(Hangul)

zi2zi: Master Chinese Calligraphy with Conditional Adversarial Networks

GANsの基本コンセプト: 通貨偽造の例

- GeneratorとDiscriminatorのプレイヤーのmin-maxゲーム
  - Generatorは乱数を関数に与え偽のサンプルを作成するよう学習
  - Discriminatorは生成サンプルと本物のサンプルを見分けるよう学習
  - GeneratorとDiscriminatorを交互に更新、Generatorが本物のサンプルに近いサンプルを生成する



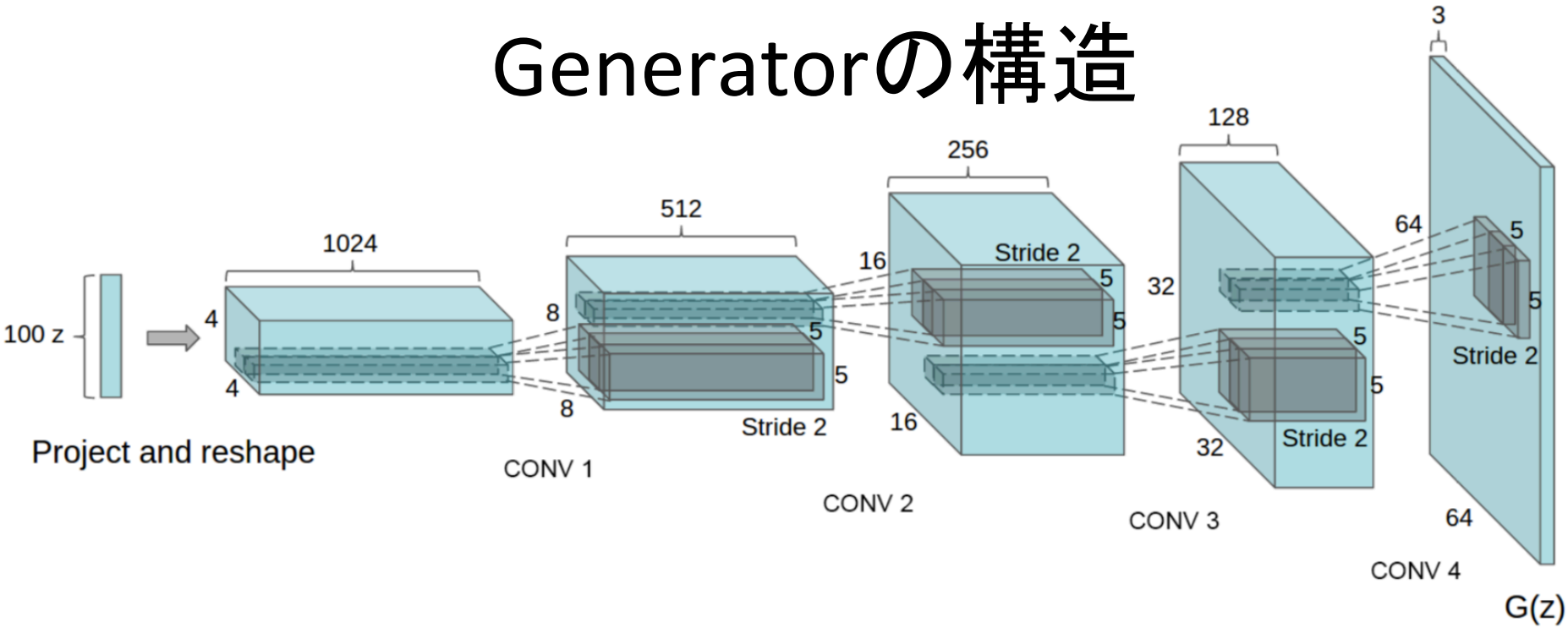
[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc., 2014a.
[絵] いらすとや

GANの目的関数

- Notation
  - P_{data}:生成の対象としたいデータの分布
  - P_G:生成モデルが生成するデータの分布
  - z:生成モデルが種として生成する乱数の確率変数
  - G:生成モデル G(z)=xは生成画像
  - D: 識別モデル D(x)は識別結果(real or fake)
- Generator: Dによる識別精度が低下するようにGを学習したい
- Discriminator: Dの識別精度が上がるようにDを学習したい

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_z}[1-\log D(G(\boldsymbol{z}))]$$

Generatorの構造



乱数生成
100次元一様乱数

転移畳み込み(transposed conv)
画像サイズを増加させ、空白を周囲の値で埋めてからフィルターを畳み込み入力サイズより大きいサイズの特徴マップを生成

生成画像
64*64, RGB画像

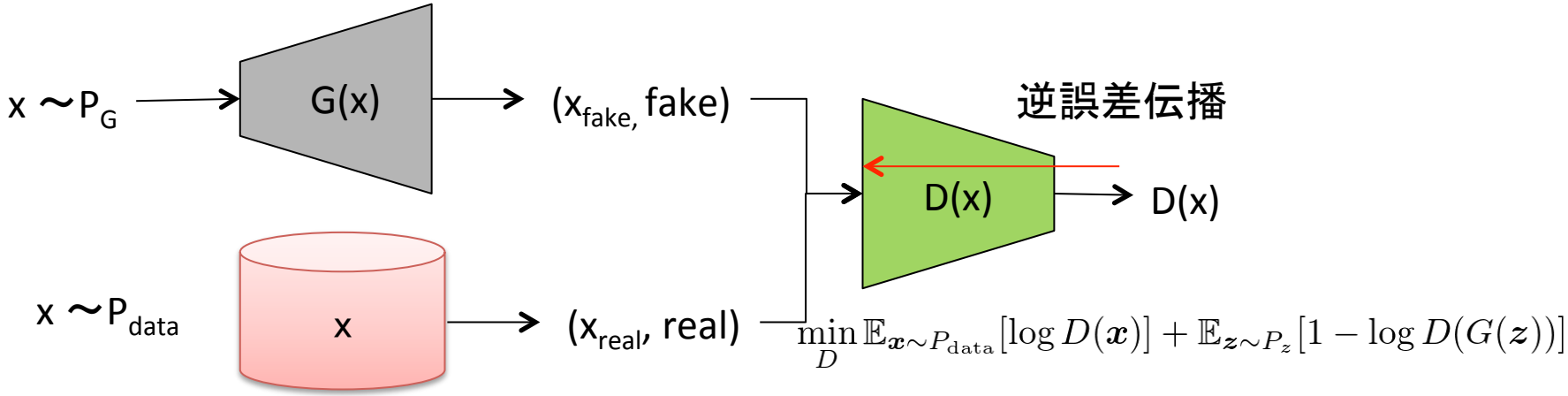
DiscriminatorはCNNによる画像分類器が利用できる

Radford et al. “Unsupervised representation learning with deep convolutional generative adversarial networks.”

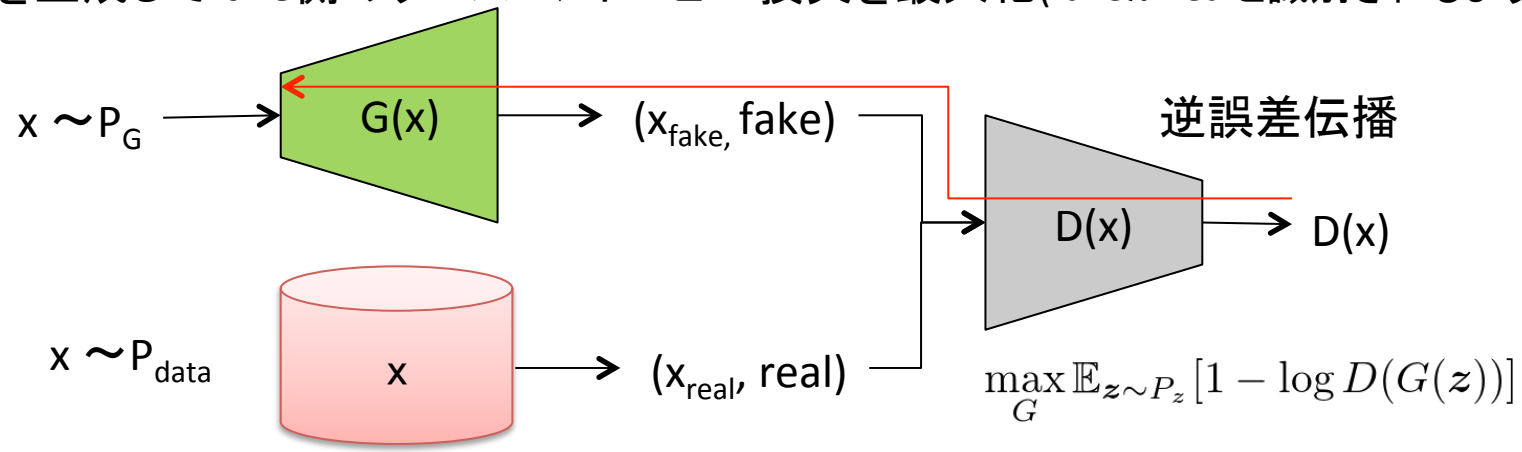
GANの最適化

識別モデルDと生成モデルGを交互に最適化する

識別モデルの最適化：
Realとfakeを同数生成してクロスエントロピー損失を最小化(realとfakeが正しく識別できるようにDを更新)



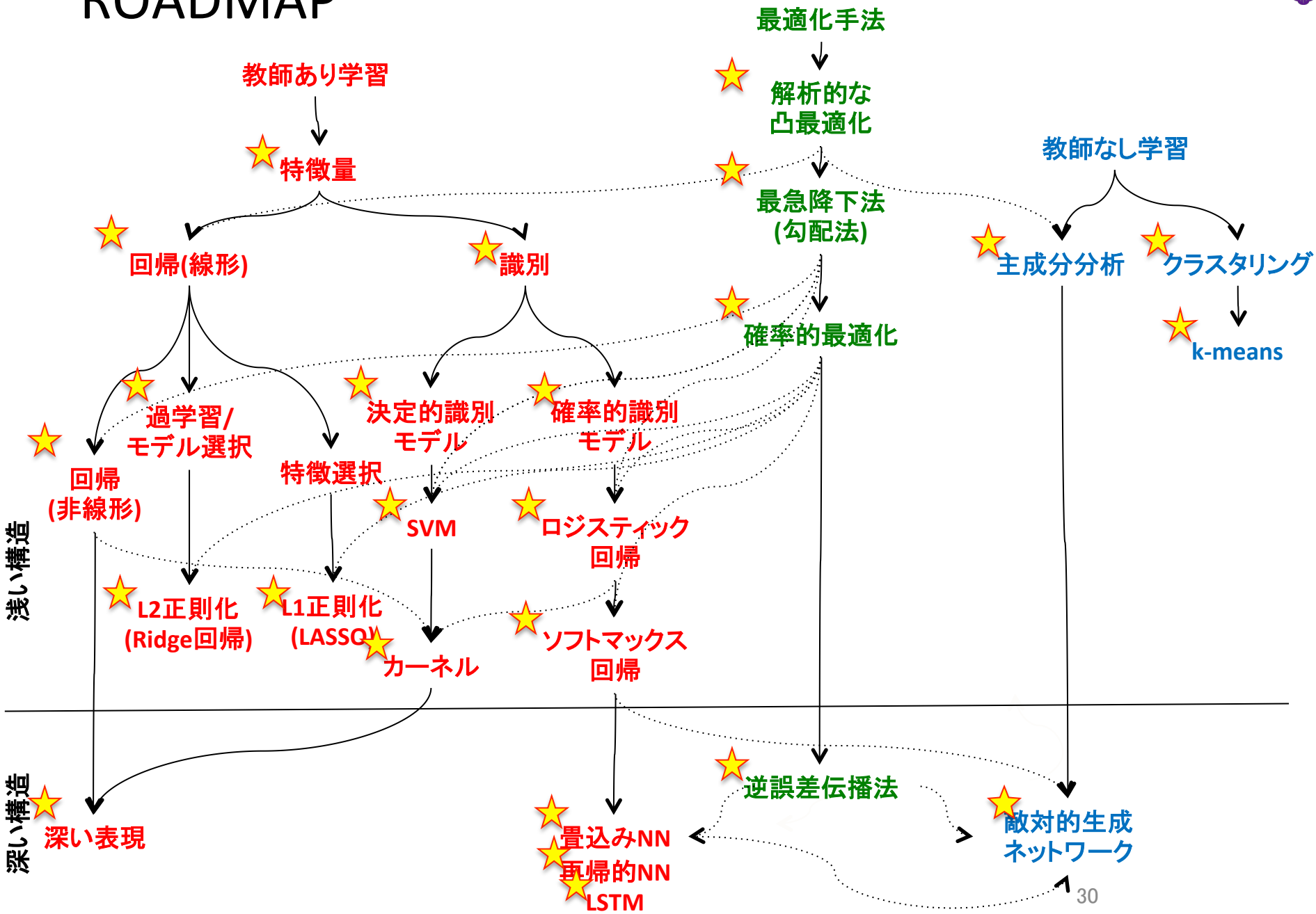
生成モデルの最適化：
Fakeを生成してfake側のクロスエントロピー損失を最大化(fakeがrealと識別されるようにGを更新)



GANまとめ

- 高次元で複雑な分布の生成モデルを実現するNN
- 生成モデル：低次元の単純な分布に従う乱数を、高次元の複雑な分布に従うサンプルに変換するNN
- 識別モデル：本物のデータ(real)と、生成モデルによる生成データ(fake)を識別するNN
- 最適化はmin-maxゲームとして定式化
  - 生成モデルは、fakeがrealと識別されるように学習
  - 識別モデルはfakeとrealが正しく識別されるように学習

ROADMAP


最適化手法
教師あり学習
解析的な
凸最適化
特徴量
最急降下法
(勾配法)
教師なし学習
回帰(線形)
識別
主成分分析
クラスタリング
確率的最適化
k-means
過学習/
モデル選択
決定的識別
モデル
確率的識別
モデル
回帰
(非線形)
特徴選択
SVM
ロジスティック
回帰
浅い構造
L2正則化
(Ridge回帰)
L1正則化
(LASSO)
カーネル
ソフトマックス
回帰
逆誤差伝播法
深い構造
深い表現
畳込みNN
再帰的NN
LSTM
敵対的生成
ネットワーク

これで機械学習の授業はおわりです

この授業の主要なテーマは

(1)データから(人間が)自然と思える予測をどのようにモデル化するか

(2)そのモデルの推定をどのように最適化問題に変換し，どのように最適化するか

でした．しかし最も重要なことは

我々は何から何を予測したいのか？

それをどう使いたいのか？

を自分で考える力を得ることです

一学期間お疲れさまでした．

テストがんばってください．