RITSUMEIKAN

機械学習 第１３回 系列データの識別

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義スケジュール

担当教員１：福森（第1回～第15回）

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	**系列データの識別**
14	半教師あり学習
15	強化学習

担当教員２：叶昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- ラベル系列に対する識別
- 系列ラベリング（Label）問題 その前後の出力や近辺の入力に依存する 場合があるので、そのような情報の活用を考える必要がある
  - 条件付き確率場（かくりつば）（Conditional Random Field）隐马尔科夫模型
- 系列識別問題 ひとまとまりの系列データを特定のクラスに識別する問題
  - 隠（かく）れマルコフモデル（Hidden Markov Model）条件概率场
- 演習問題

取り扱う問題の定義：教師あり・系列データ識別

输入特征向量（系列信息），并做一个判别器对其进行分类（输出其所属的类别）。

- 特徴ベクトル（**系列情報**）を入力して、それをクラス分けする（属するクラスを出力する）識別器を作る

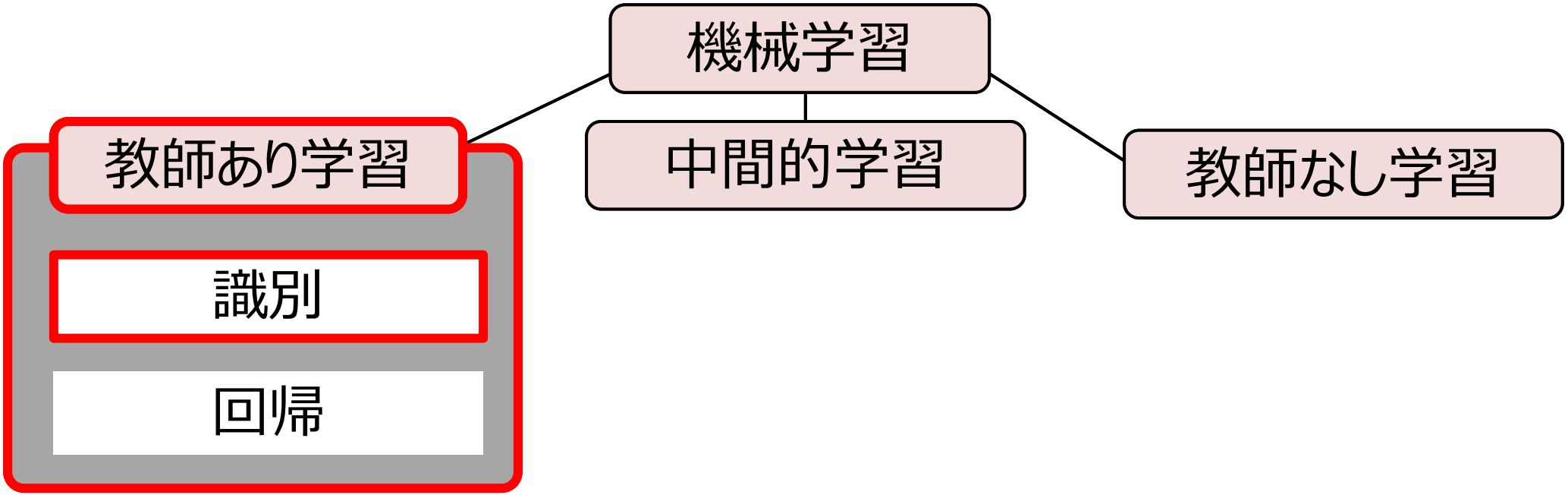
※ 教師あり学習の問題での学習データ（系列データ）は、以下のペアで構成される

$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}, \quad i = 1,2,\ldots,N$

入力データの特徴ベクトル（系列情報） ← $\boldsymbol{x}_i$

$\boldsymbol{y}_i$ → **正解情報（要素数は1個以上）**

N → 学習データの総数



ラベル系列に対する識別

- 系列データ
  - 個々の要素の間にi.i.d.の関係が成立しないデータ
    - i.i.d. (independent and identically distributed)
      - 個々の事例が独立であるということ

- 系列データの識別問題の分類
  1. 入力の系列長と出力の系列長が等しい
  2. 入力の系列長に関わらず、出力の系列長が1
  3. 入力の系列長と出力の系列長に明確（めいかく）な対応関係がない

ラベル系列に対する識別

1. 入力の系列長と出力の系列長が等しい問題

■ 例：形態素解析

- 単語列を入力して、名詞や動詞などの品詞列を出力
- １つの入力（単語）が一つの出力（品詞）が対応

■ この問題のポイント

- ある時点の出力が、その前後の出力や近辺の入力に依存する場合があるので、そのような情報の活用を考える必要がある
- これが**系列ラベリング問題** 系列标签问题

ラベル系列に対する識別

2. 入力の系列長に関わらず、出力の系列長が1

■ ひとまとまりの系列データを特定のクラスに識別する問題

- 例：動画像の分類や、音声で入力された単語の識別など

■ この問題のポイント

- 一般に長さが不定（ふてい）の時系列信号を入力系列とする
- 入力系列は共通の性質をもつ複数のセグメントに分割するが、入力系列を分割する場所（ばしょ）に関する情報は一般に得られない
- セグメントの区切りを隠れ変数として、その分布を教師なし学習するという手法と、通常の教師あり学習を組み合わせることになる
- これが**系列認識問題**

ラベル系列に対する識別

3. 入力の系列長と出力の系列長に明確な対応関係がない
   - 例：音声認識で、話速（早口か、ゆっくり話すか）によって単位時間あたりの出力単語数が異なる
   - 学習にも認識にも相当込み入った工夫が必要になるので今回は概要のみ説明

系列ラベリング問題 –CRF-

序列标注的问题

□ 系列ラベリング問題

- ラベル特徴の系列を入力として、
  それと同じ長さのラベル系列を出力する識別問題を考える

语素的解析

- 例：形態素解析

输入一连串单词并为每个单词指定一个语系的问题

  - 単語の系列を入力として、それぞれの単語に品詞を付ける問題
  - 形態素の列は、ある言語の文を構成するので、その言語の
    文法に従った並び方が要求される

需要遵从语法

    - 例えば、日本語の形態素列は「形容詞の後には名詞が来ることが多い」、
      「助詞の前には名詞が来ることが多い」などの傾向が存在

日语的语素序列有一种倾向，例如形容词后面为名词，助词前面为名词

入力	系列	で	入力	さ	れる	各	要素
出力	名詞	助詞	名詞	動詞	接尾辞	接頭辞	名詞

形態素解析

系列ラベリング問題 –CRF-

虽然训练数据是以一对输入序列和输出标签的形式给出的，但把它们当作单独的辨别问题是不合适的，因为输出序列按顺序有依赖性，就像在形态分析的例子中。如果只考虑单个的输入对应的输出的话，舍弃了序列的性质。

□ 系列ラベリング問題の問題点

- 単純に1つの入力に対して 每一个输入，判别器按顺序输出一个标签
  1つのラベルを出力する識別器を順次適用(じゅんじ)する場合
  - **問題点１：系列としての性質を捨(す)ててしまっている** 舍弃性质
    - 学習データは、入力系列と出力ラベルのペアとして与えられるが、形態素解析の例のように、出力系列には並びによる依存関係があるので、個々の識別問題として扱うのは不適当

- 出力もまとめて、出力系列を1つのクラスにする場合
  - **問題点2：膨大なクラス数を取り扱うことになる** 处理了大量的类
    - 例えば、品詞が10種類で、20単語からなる文にラベル付けする問題では10^{20}種類の出力が可能
    - これらを個別のクラスとして扱うのは、ほぼ不可能

系列ラベリング問題 –CRF-

対数線形モデル 对数线性模型

- 前後の入力や一つ前の出力などの特徴を利用し、
  かつ、系列としての確からしさを評価しながら
  探索的に出力を求める手法
  X：输入变量，需要标记的观测序列；Y：输出变量，标记序列
- 入力$\boldsymbol{x}$が与えられたときの出力$\boldsymbol{y}$の条件付き確率$P(\boldsymbol{y}|\boldsymbol{x})$

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x},\boldsymbol{w}}} \exp\{\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\}, \qquad Z_{\boldsymbol{x},\boldsymbol{w}} = \sum_{\boldsymbol{y}} \exp\{\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\}$$

每个维度是由$\boldsymbol{x}$、$\boldsymbol{y}$定义的各种特征函数。
特征函数是指当特征（输入序列的特征）的条件为1时返回1，不为1时返回0的函数。

- $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$：素性（そせい）ベクトル 特征向量
  - 各次元は$\boldsymbol{x}, \boldsymbol{y}$から定められる様々な素性関数
  - 素性関数は、ある素性（入力系列の特徴）について条件が一致すると1、一致しないと0を返す関数
- $\boldsymbol{w}$：素性関数の重みからなる重みベクトル
  由特征函数的权重组成的权重向量

系列ラベリング問題 –CRF-

- 対数線形モデル（つづき）
  - 出力は以下の最大値を解くことによって求まる

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \left(\frac{1}{Z_{\boldsymbol{x},\boldsymbol{w}}} \exp\{\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\} \right)$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \left(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \right)$$

将前面和后面的输入和输出自由组合为特征向量的元素，设定特征向量能够反映一个序列

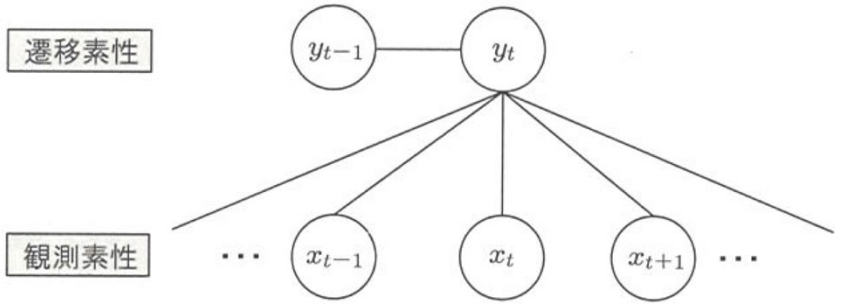
  - 素性ベクトルの要素として、前後の入力や出力を自由に組み合わせることができるので、系列としての情報を反映したものを設定できる → **問題点1** を解決
    - 全ての可能な$\boldsymbol{y}$について計算する必要があるので、まだ問題点2を解決していない

系列ラベリング問題 –CRF-

组合的限制

- 利用する素性の組み合わせを以下のように限定
  - 出力系列で参照できる情報は一つ前のみ、入力系列は自由な範囲で参照できるとする
    - この限定によって、**問題点2**を解決

输出序列的参照信息只有前面一个序列，输入序列可以在自由范围进行参考



  - 出力系列を参照する素性を**遷移素性**、入力と対応させる素性を**観測素性**と呼ぶ
    - 遷移素性の例：名詞の次に助詞が出現
    - 観測素性の例：助詞で「が」が観測される

系列ラベリング問題 –CRF-

- 対数線形モデルの出力は下式に書き換えられる

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}}{\operatorname{argmax}} \left(\sum_t \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, y_t, y_{t-1}) \right)$$

  - 右辺（うへん）の最大値を求めるには、先頭$t = 1$からスタートして、tを増やしながら、その時点での最大値を足し続ける
    - この手順を**ビタビアルゴリズム**と呼ぶ

维特比算法：从时刻t=1开始，递推计算在时刻t状态为i的各条部分路径的最大概率，直到得到t=T 的最大概率为最优路径的概率P

- **条件付き確率場**（Conditional Random Field; CRF）
  - 前スライドのような制限を設けて、対数線形モデルを系列識別問題に適用したもの
    - CRFの学習は、対数線形モデルほど簡単ではないが、素性関数の値が1つ前の出力にしか影響されないという条件のもとで、重複（ちょうふく）する計算をまとめることができるという性質を利用

系列識別問題 –HMM-

系列識別問題

- 入力の系列長に関わらず出力の系列長が1である問題を考える

系列識別問題の例

- ユーザのキー入力やマウス操作をシンボルで表して、その系列で初心者と熟練者を識別する問題を考える
- 入力は、キー入力・マウス操作を抽象化したものとする
  - k：10回以上連続で通常キーを入力
  - e：エラーキー（DeleteキーやBack spaceキー）の入力
  - g：ファイル保存や文字修飾などのGUI操作

系列識別問題 –HMM-

- 系列問題の例（つづき）
  - 以下のような初心者と熟練者の傾向があったとする
    - 初心者 初心者Bさんの記録：k e k g k e k g g k g k k e g e e k e e e g e
      - kとgを頻繁（ひんぱん）に繰り返し、時間が経過（けいか）するにつれてeが増える
    - 熟練者 熟練者Sさんの記録：k k e k g k k k e k g k g g g e g k g
      - 最初にキー入力を重点的（じゅうてんてき）に、あとからGUI入力をまとめて行う
  - 「k g e k g k k g e k g e k e e k e g e k」が観測されたとき、この人は初心者か、熟練者か？
  - 与えられた系列を$\boldsymbol{x}$として、クラスyの事後確率$P(y|\boldsymbol{x})$を何かしらのモデルを使って計算することになる
    - 今回の例の場合
      - $y = \mathrm{B}$：初心者、$y = \mathrm{S}$：熟練者

系列識別問題 –HMM-

□ 隠れマルコフモデル（Hidden Markov Model; HMM）

■ 下式の$\boldsymbol{x}$、yの同時確率を考える生成モデルアプローチ

$$\underset{y}{\operatorname{argmax}} P(y|\boldsymbol{x}) = \underset{y}{\operatorname{argmax}} \frac{P(\boldsymbol{x}, y)}{P(\boldsymbol{x})} = \underset{y}{\operatorname{argmax}} \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})} = \underset{y}{\operatorname{argmax}} P(\boldsymbol{x}|y)P(y)$$

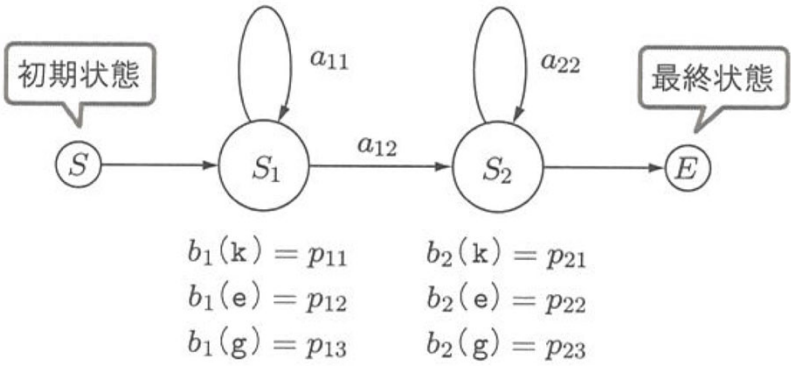
- HMMは$P(\boldsymbol{x}|y)$を計算する確率的非決定性オートマトンの一種（いっしゅ）
- 各状態であるシンボルをある確率で出力し、
  ある確率で他の状態（あるいは自分自身）に遷移する

■ HMMを構成する要素と確率

- 状態の集合：$\{S_i\}\ (1 \leq i \leq n)$
- 初期状態、最終状態の集合
- 遷移確率：状態 i から状態 j への遷移確率 a_{ij}
- 出力確率：状態 i で記号 o を出力する確率 $b_i(o)$

系列識別問題 –HMM-

- 初心者・熟練者の識別問題の例について
  左図のHMMの構造を仮定する
  - 状態S_1から状態S_2に移(うつ)る系列内の位置を隠れ変数と見て
    EMアルゴリズムで各状態の確率を推定できる

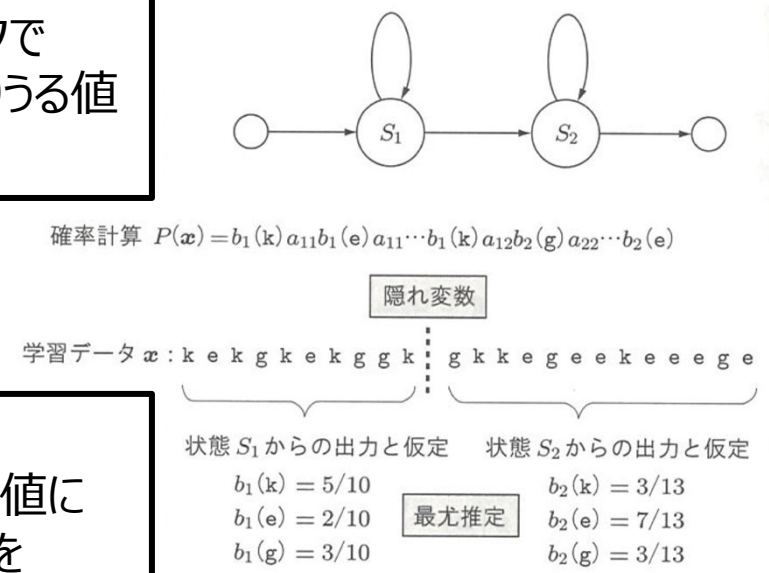


HMMの構成

Eステップ
現在のHMMのパラメータで
隠れ変数のすべての取りうる値
について$P(\boldsymbol{x})$を計算

繰り返す

Mステップ
隠れ変数の全て取りうる値に
ついてHMMのパラメータを
最尤推定し、$P(\boldsymbol{x})$を重みとして
足し合わせる



EMアルゴリズムによる
HMMの学習

系列識別問題 –HMM-

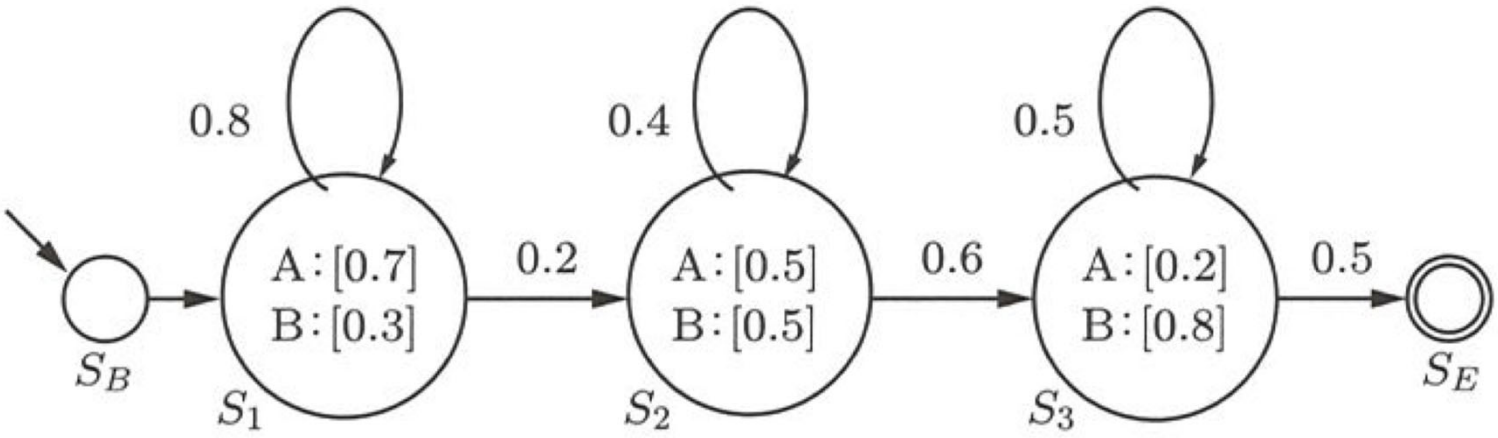
- HMMは「入力系列長と出力系列長に明確な対応関係がない問題」にも適用できる
  - 任意長の出力系列の表現
    - クラスごとに作成した全てのHMMの初期状態と最終状態をそれぞれ1つにまとめ、最終状態から初期状態へ戻る遷移を加える
    - この連結されたHMMを用いて、入力系列に対して最も確率が高くなる遷移系列をビタビアルゴリズムによって求める

最大概率的迁移序列被确定，全部隐藏变量的位置被确定，每个输入的子序列的输出是确定的

  - 最も確率が高くなる遷移系列が定まるということは、全ての隠れ変数の位置が定まるということに等しくなるので、入力の各部分系列に対して出力が定まる

演習問題13-1（10分間）

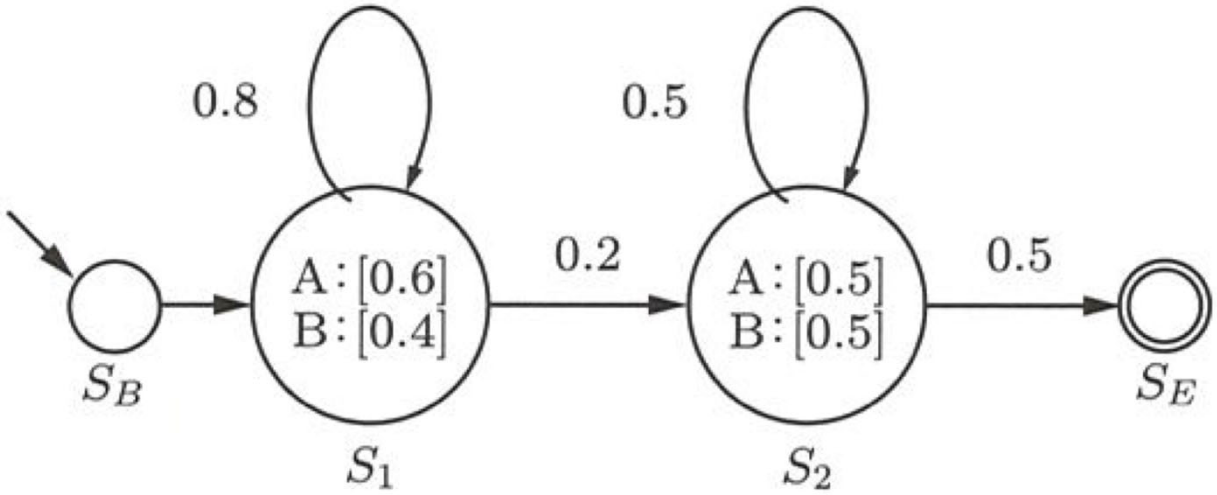
- 下図に示すHMMが与えられているとき、
  特徴ベクトル系列 “AAABB” が出力されるような
  状態遷移系列を全て求めよ



HMM

演習問題13-2（15分間）

- 下図に示すHMMが与えられているとき、
  特徴ベクトル系列 “AAB” の出力確率を求めよ



HMM