導入

通信路モデル

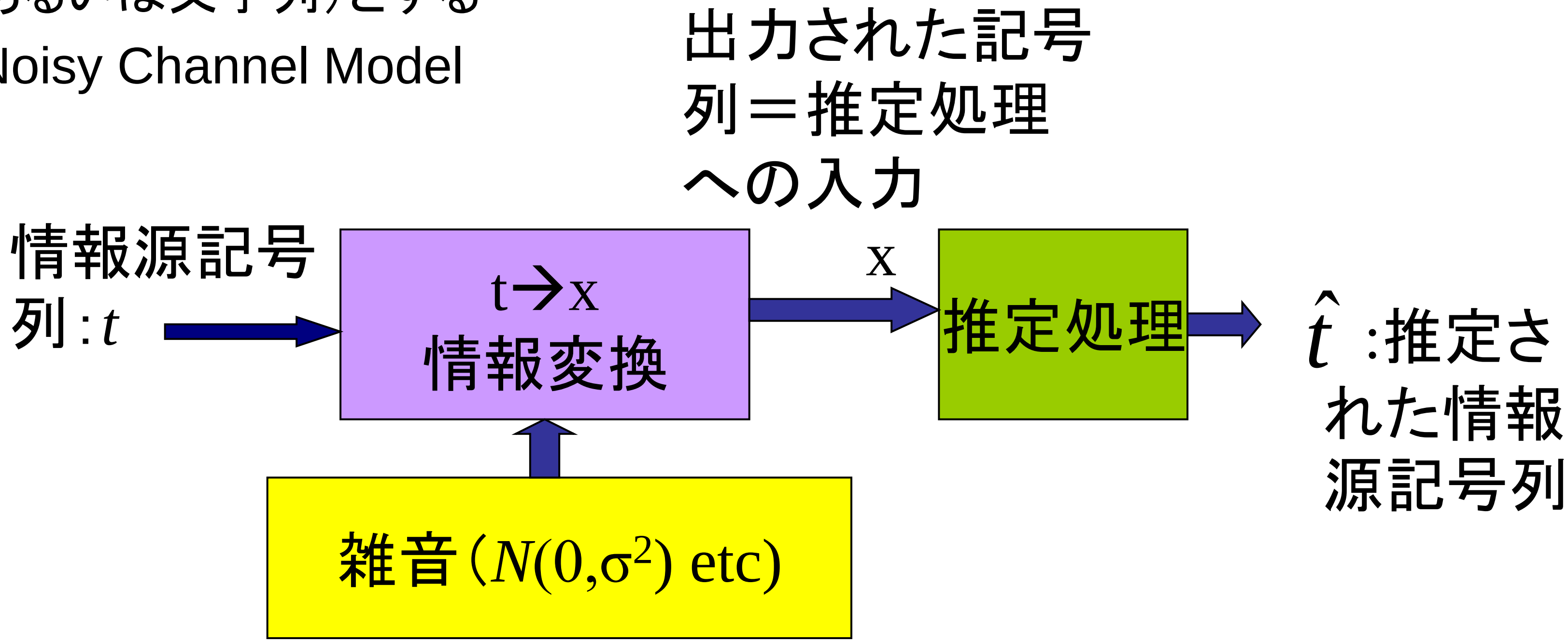
Bayes統計

最尤推定とMAP推定

データの性質

機械学習の先史時代
--情報の変換過程のモデル化--

- 情報源を記号列（例えば単語列あるいは文字列）とする
- Noisy Channel Model



出力された記号列＝推定処理への入力データxから情報源記号列tを推定し $\hat{t}$ を計算する

Bayes統計の意義

- Bayesの定理 $$P(t \mid x) = \frac{P(x \mid t)P(t)}{P(x)}$$
- *P(t|x)*は、新たな出力記号列*x*が得られたときの情報源から出力された記号列 t を推定する式で、これを最大化する t すなわち $\hat{t} = \arg\max_{t} P(t \mid x)$ を求めるのが目標。
- ところが、このままでは、既に得られている情報を使えないので、Bayesの定理で変換する。
- すると、既知の情報源状態と出力記号列のペアに関する条件付き確率*P(x|t)*（＝教師データ）
- 情報源についての事前知識*P(t)*が使える形になる。

Bayes統計とは

1. 常にBayesの定理を用いる
2. 用いられる確率は主観確率（＝確信度）
3. 事前情報を利用する
4. 未知量(確率分布のパラメター）は確率的に変動
5. 観測されたデータは絶対的
6. 推測は常に条件付
7. アドホックな手続きを認めない

Bayes統計を用いた情報変換過程のモデルにおける出力データからの情報源の推定方法

- 通信路を条件付確率でモデル化：$P(x|t)$
- 目的は x が観測されたときの t の確率すなわち事後確率 $P(t|x)$を最大化する情報源の確率。

$$\hat{t} = \arg\max_t P(t \mid x)$$

ここでベイズの定理により

$$= \arg\max_t P(x \mid t) P(t)$$

- $P(t)$は情報源記号列の既知の統計的性質が利用できる
- $P(x|t)$ は情報源記号列tが情報変換およびnoisy channelの雑音によってx毎に変化する確率。
- この確率は多数の＜t,x＞対の観測データにより計算する

情報変換過程モデルの適用例

- 例:機械翻訳

元言語
x:私がリンゴを食べる

機械
翻訳

翻訳先言語
t: I eat an apple

- $P(t|x)$　は元言語のテキスト x（既知）が翻訳先言語のテキスト t に翻訳される確率
- $P(x|t)$　はtという翻訳結果に対する元言語のテキストがxである確率
- $P(t)$　翻訳先言語におけるテキストtの自然さ。例えば、N単語列のコーパスにおける　単語3-gram確率
- 以上の設定で下の式　　　は機械翻訳の出力

$$\hat{t} = \arg\max_t P(t \mid x) = \arg\max_t P(x \mid t)P(t)$$

- この考え方を元にしたのが現在主流となってきている統計的機械翻訳（IBMで1993年に開発された）

日英機械翻訳の例

- Ｐ(リンゴを食べる|eats an apple)=0.3
- P(リンゴを食べる|eats apples)=0.2
- P(彼は｜He)=1.0
- P(He eats apples)=0.2、P(He eats an apple)=0.5
- P(He eats an apple|彼はリンゴを食べる)
- ＝1.0x0.3x0.5=0.15
- P(He eats apples|彼はリンゴを食べる)
- ＝1.0x0.2x0.2=0.04

→"He eats an apple"のほうが良い英訳

- 事前知識として Ｐ(He eats apple)=0.0 があれば

→非文P(He eats apple)=0にできるのがベイズの強み

- 例：文書分類
  - $P(t|x)$ においてxが与えられた文書、tがカテゴリ

$$\text{推定されたカテゴリ} : \hat{t} = \arg\max_{t} P(x \mid t) P(t)$$

  - $P(t)$ はカテゴリtの文書の出現確率
  - $P(x|t)$はカテゴリtにおいて文書xが出現する確率
  - このモデル化にはいろいろな方法があるが、簡単なのは、出現する単語$w_1, \ldots w_N$
  - $P(x|t) = P(w_1, \ldots w_N|t)$だが、このままでは計算しにくいので$w_1, \ldots w_N$が独立だとすると

$$P(w_1, \ldots, w_N \mid t) = \prod_{n=1}^{N} P(w_n \mid t)$$

Why?

  - これを naïve Bayse 分類とよぶ。

文書分類の例：長澤まさみ vs 上野樹里

- 「長澤まさみ」関連の文書に高い確率で出現する単語
  - 主演、映画、東宝、吉田礼、薬師丸ひろ子、サッカー、
- 「上野樹里」関連の文書に高い確率で出現する単語
  - 主演、のだめ、カンタービレ、ドラマ、ラスト、フジテレビ、
- 分類したい文書：Dの含む単語は
  - 主演、ラスト、フレンズ
  - P(主演|長澤）＝０．１、P（主演｜上野）＝０．１
  - P（ラスト｜長澤）＝０．２、P（ラスト｜上野）＝０．２
  - P（フレンズ｜長澤）＝０．２、P（フレンズ｜上野）＝０．２

Googleのヒット数から推定したところ、P(長澤)＝0．6、　P(上野)＝0．4

- P(長澤｜D)＝P(D|長澤)P(長澤)

＝P(主演｜長澤)P(ラスト｜長澤)P(フレンズ｜長澤)P(長澤)

＝0.1x0.2x0.2x0.6=0.0024

- P(上野｜D)＝P(D|上野)P(上野)

＝P(主演｜上野)P(ラスト｜上野)P(フレンズ｜上野)P(上野)

＝0.1x0.2x0.2x0.4=0.0016

よって、文書Dは長澤に分類

しかし、Dに「カンタービレ」という単語も含まれ、

P(カン..|長澤)＝0．1、P(カン..|上野)＝0．8だと

P(長澤｜D)＝0．00024、P(上野｜D)=0．00128

で文書Dは上野に分類。　直感にあっているようだ！

教師あり学習

- 上記の例では、情報源のモデルである$P(\mathrm{t})$や$P(\mathrm{x}|\mathrm{t})$は単に出現確率だったが、ここで適切な確率分布を考えることが可能
- すると、その分布を決めるパラメターを推定する必要が出てくる。
- そのために＜t,x＞という情報源の状態と出力データの対データが多数入手できれば利用する。
- この＜t,x＞を教師データ（あるいは観測データ）と呼ぶ。
- すると、機械学習の中心となる教師あり学習は、

確率分布$P(t)$、$P(x|t)$のパラメターを教師データ＜t,x＞を利用して求める

という問題になる。

教師なし学習

- 教師あり学習では教師データ＜t,x＞の集合が与えられた状態で、P(t)やP(x|t)のパラメターを求めた。
- しかし、データ＜x＞の集合だけが与えられていて（tは与えられていない）ときはどうする？

- データ＜x＞の集合から、P(x)のパラメターだけを求めることになる。

  - 直観的には、データ＜x＞を類似したものにグループ化する
  - →クラスタリングと言い、グループのことをクラスタと呼ぶ。

- これを教師なし学習と呼ぶ。

識別モデルと生成モデル

- 入力データxに対応する予測値tを求める
- 識別モデル(discrimiative model) : p(t|x)を直接モデル化する。このp(t|x)によって、未知のxに対するtを予測（あるいは推定）する方法
  - t=f(x)となる関数を直接求めるものもあり。
- 生成モデル(generative model) : ベイズの定理でp(t|x)をp(x|t)p(t)/p(x)に変換。p(x|t)を学習。p(t)を事前データから求める。これと既知の<x,t>のペアのデータからp(x|t)のパラメターを更新。これによって、未知のxに対するtを求めるp(t|x)の確率分布をモデル化する。

事前分布

観測データが知られて後のp(x|t)の事後分布

最尤推定とMAP推定

- 最尤推定
  - 分布 $P(X\mid\theta)$ のパラメターθの推定値 $\hat{\theta}$ を以下の式で求める

$$\hat{\theta} = \arg\max_{\theta} P(X_1,\ldots,X_N \mid \theta)$$

  - あるいは対数をとり推定:対数尤度の最大化

$$\hat{\theta} = \arg\max_{\theta}\ \log P(X_1,\ldots,X_N \mid \theta)$$

- MAP推定(事後確率の最大化)
  - 事前確率$P(\theta)$ が与えられていたときには、次式のように事後分布の確率を最大化するパラメターを求める

$$\hat{\theta} = \arg\max_{\theta}\ \log P(X_1,\ldots,X_N \mid \theta)P(\theta)$$

ただし、$X_1,..,X_N$はN個の観測データ

◆問題1 $P(X)=\theta^X(1-\theta)^{1-X}$ (X_iは、0か1)で定義されるベルヌーイ試行を独立にN回繰り返したとき、0がm回、1がN-m回観測されたとする。最尤推定して θを求めよ

また、事前分布として、 $P(\theta)=b\theta$ただし、$0\le\theta\le1$
のときのMAP推定した θ を求めよ。
この場合の結果の意味を考察せよ。

◆問題2 次式の多項分布 $P(X)=\dfrac{N!}{X_1!\cdots X_K!}\theta_1^{X_1}\cdots\theta_K^{X_K}$ $\sum_{k=1}^{K}\theta_k=1$

において最尤推定して θ_i を求めよ。

事前分布が、 $P(\theta)=\dfrac{\Gamma(\alpha_1+\cdots+\alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1}$

の場合の、MAP推定した θ_i を求めよ

データの性質

- 今までは、情報源の記号tと出力記号列(＝直接に観測されたデータ)xは、機械学習において直接に計算の対象としていた。
- この仮定が成立する場合も多い。
  - 身長、体重、薬の濃度、価格、などの(連続)数値データ
  - 人数、個数、などの整数をとる数値データ
  - 割合、%など
  - 男女、国籍など属性が記号の場合(整数に変換すれば数値として処理可能)
- しかし、必ずしも直接に観測されたデータだけを使える場合ばかりではない。

観測データを表す情報の次元

- 観測データ点が人間の場合の例
  - x=(身長、体重、血圧、収入金額)^T→数値だけなので簡単。単位は外部知識とする。

    Ex (170, 50, 120, 10,000,000）
    - 確率分布としては正規分布など。
  - x=(職業、発熱)^T　→記号。2つの方法
    - 記号に番号を与える。Ex 無職=0, 学生＝1,..、発熱無＝0、有＝1
      - 確率分布としては離散数値をとる分布など。数値の意味付けが難しい。
    - 記号の種別ごとに1次元を与える（次のページ参照）

記号の種類ごとに次元を割り当てる方法

- x=(訪問国1、,…,訪問国N)^T　　ex(USA,UK,Italy)^T
  - 対策：国を番号つける。

  (USA=1, UK=2, Japan=3, China=4, Italy=5,…)
  - この番号がベクトルの何番目の要素かを示すとして、数値のベクトルとして表現：Bernoulli分布：$Bern(x|\mu) = \mu^x (1-\mu)^{1-x}$
    - 上のexは　(1,1,0,0,1,….)^T
  - このベクトルの次元は世界中の国の数だけあるため、かなり大きい。しかし、観測データには0が多く、スパースなデータ
- 記号の出現回数のある場合

  x=((訪問国1、滞在日数1),…,(訪問国N、滞在日数))^T
  - ex((USA,15),(UK,5),(Italy,3))^T → (15,5,0,0,3,….)^T
  - 多項分布：$Mult(15,5,0,0,3,\ldots|\mu_{USA}, \mu_{UK}, \mu_{Japan}, \mu_{China}, \mu_{Italy}, \ldots)$

$$\propto \mu_{USA}{}^{15} \mu_{UK}{}^{5} \mu_{Japan}{}^{0} \mu_{China}{}^{0} \mu_{Italy}{}^{3} \cdots$$

次元の大きさ

- 国と滞在日数の例と同じタイプの問題を、テキストデータで考えてみよう。
- あるテキストを表現するには、そのテキストに出現した各単語の個数で表現する。
  - 次元は語彙数　日本語の新聞では約40万語。固有名詞や複合語まで入れると、100万以上。→　100万次元のベクトルを扱う必要あり！
  - 個々の単語だけを対象にすれば済むのか？
    - ABC証券、ABC証券株式会社、… 、総理が失言、総理が訂正、…、というような単語の連鎖で見ないと分からない場合は？
    - N単語の連鎖（＝N-gram)の種類数は、100万のN乗！！！
    - **しかし、このような多次元がすべて重要な情報だとも思えない**
    - 次元圧縮の技術が有望 i.e. Singular Value Decomposition (SVD)とかLatent Semantic Indexing(LSI)

特殊性を表すデータ　１

- これまでに示したデータ点の数値は、観測された数値（出現回数など）を直接使っていた。
- 観測データ全体の構造を利用した**tf*idf**と呼ばれる数値も有力
- データ点頻度　Data point Frequency　:　DF
- ただし、DF(j)はj番目の次元のデータが0でないデータ点の数
- また、観測データ点の総数をNとする。

特殊性を表すデータ ２

- データ点頻度　Data point frequency:DF
- ただし、DF(j)はj番目の次元のデータが0でないデータ点の数
- また、観測データ点の総数をNとする。
- IDF(j)=1/DF(j)
- TF(i,j)=観測データ点iで第j次元のデータの出現回数
- TF*IDFの定義:

$$w_{i,j} = TF * IDF(i,j) = TF(i,j) \cdot \log \frac{N}{DF(j)}$$

例

- データ例

  旅行者a:(USA=10, UK=2, Japan=3, China=0, Italy=0）

  旅行者b:(USA=0, UK=2, Japan=0, China=4, Italy=0）

  旅行者c:(USA=5, UK=0, Japan=2, China=0, Italy=0）

  旅行者d:(USA=2, UK=0, Japan=1, China=2, Italy=1）

- DF(USA)=3, DF(UK)=2, DF(JP)=3. DF(CH)=2, DF(IT)=1
- N/DF(..)は USA=4/3, UK=4/2, JP=4/3, CH=4/2, IT=4/1

- TF*IDF(USA,a)=10*log(4/3)=4.114, TF*IDF(USA,b)=0

  TF*IDF(UK,a)=2*log(4/2)=2

  TF*IDF(IT,d)=1*log(4/1)=2

特殊性を表すデータ ３

- TF*IDFの定義：$w_{i,j} = \mathrm{TF*IDF}(i,j) = \mathrm{TF}(i,j) \cdot \log \frac{N}{\mathrm{DF}(j)}$

- TF*IDF(i,j)は、データ点: iだけで特別に多く現れる次元: jの数値を表す。
  - 例えば、新聞の１記事を観測データ点とし、次元を単語とすると、TF*IDF(i,j)の大きな単語iは、偏りのある特殊ないし専門の単語、小さな単語は一般的な単語といえる。

- TF*DIFを用いて観測データ点を表現しなおすと、いろいろなことが見えてくることがある。

距離の定義

- 観測データ点を多次元空間中の点と定義
  - そこで2つの問題
  - 各次元は観測データ点からどのように定義するか
    - 次元のことを**feature**あるいは**素性**（そせい）と呼ぶ
    - この問題をfeature design：素性設計と呼ぶ。例えば、
      - 2つの素性の比を新たな素性とする ex 身長/体重
      - 2つの素性の連結したもの　ex　日本・銀行、　日本・沈没、
    - しかし、これは個別適用分野に応じて工夫すべし。
  - 多次元空間における2点間の距離の定義
    - ユークリッド距離ばかりではないのだ！