

情報意味論(10)

EMと事例ベースアプローチ

櫻井彰人

慶應義塾大学理工学部

目次

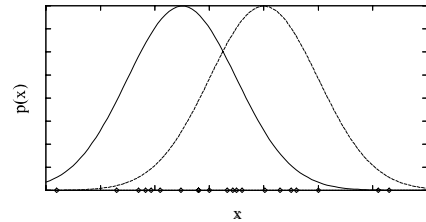
- EM-アルゴリズム
- 事例ベースアプローチ
 - k-最近傍法
 - Radial basis function network

EM: Expectation Maximization

- 使う時:
 - データに観測可能でないものがあるとき
 - 教師なしクラスタリング(目標値が観測不能)
 - 教師付きクラスタリング(属性に観測不能なものがある)
- 利用場所:
 - ベイジアンネットワークの学習
 - 教師なしクラスタリング(例: AUTOCLASS)
 - P.Cheeseman <http://ic.arc.nasa.gov/ic/projects/bayes-group/people/cheeseman/>
 - HMM(Hidden Markov Model)の学習

k個の正規分布の混合

- 事例 x は次のように生成される
 - k 個の正規分布から一個を一様ランダムに選択
 - その正規分布に従い一事例をランダムに生成



k個の平均を推定するEM

- 所与:
 - k 正規混合分布 X からの事例
 - その k 個の正規分布の未知の平均 $\langle \mu_1, \dots, \mu_k \rangle$
 - どの事例 x_i がどの正規分布から生成されたかは不明
- 決定すべき:
 - $\langle \mu_1, \dots, \mu_k \rangle$ の最尤推定値
- 個々の事例の完全な記述: $y_i = \langle x_i, z_{i1}, \dots, z_{ik} \rangle$
 - z_{ij} は x_i が j 番目の正規分布から生成されたとき1
 - x_i は観測可能
 - z_{ij} は観測不能

k個の平均を推定するEM

- EMアルゴリズム: ランダムな初期値 $\langle \mu_1, \mu_2 \rangle$ を選び、次を繰り返す
 - Eステップ: 現在の仮説 $h = \langle \mu_1, \dots, \mu_k \rangle$ のもと、隠れ変数 z_{ij} の期待値 $E[z_{ij}]$ を求める(実は分布 $P(z_{ij}=1)$ を求めている)

$$P(z_{ij}=1) = E[z_{ij}] = \frac{p(x=x_i | \mu = \mu_j)}{\sum_{n=1}^k p(x=x_i | \mu = \mu_n)} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^k e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}$$

- Mステップ: 隠れ変数 z_{ij} の値は今求めた $E[z_{ij}]$ に等しいと仮定して、新しい最尤推定値 $h' = \langle \mu'_1, \dots, \mu'_k \rangle$ を求める(実は、 $P(z_{ij}=1)$ を用いて、 h' の最尤推定をしている)。

$$\mu_j \leftarrow \frac{\sum_{i=1}^m E[z_{ij}] x_i}{\sum_{i=1}^m E[z_{ij}]}$$

EMアルゴリズム

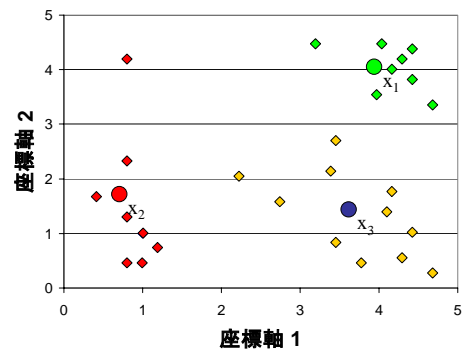
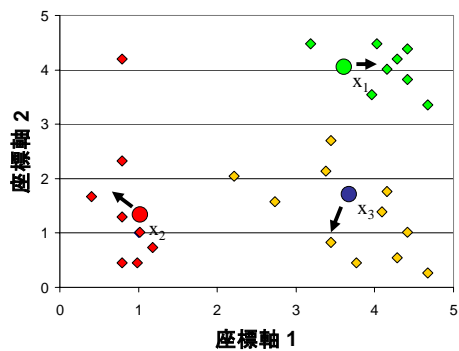
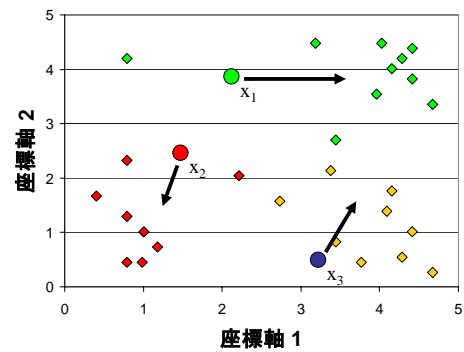
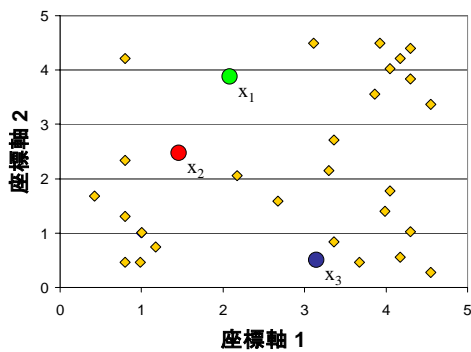
- 本アルゴリズムは局所的に最尤な h に収束し、隠れ変数 z_{ij} の推定値を与える
- 事実、 $E[\ln P(Y|h)]$ の極大値を与える。但し
 - Y は完全なデータ(観測済みと未観測なデータの組合せ)
 - 期待値は、未観測データがとりうる全ての値の上で計算される

K-Means クラスタリング: Lloyd アルゴリズム

1. Lloyd アルゴリズム
2. 任意に k 個のクラスタ中心を選ぶ
3. **while** クラスタ中心が変更された
4. 各データ点をクラスタ C_i に割り当てる
割り当てるクラスタは、最も近いクラスタ中心のクラスタ ($1 \leq i \leq k$)
5. すべての点を割り当て終わったら、各クラスタの重心を新たなクラスタ中心とする
すなわち、新たなクラスタ中心は、クラスタ C それぞれに

$$\frac{\sum_{v \in C} |v|}{\text{全 } v \in C}$$

*このアルゴリズムは局所解に陥ることがある。



一般のEM問題

- 所与:
 - 観測データ $X = \{x_1, \dots, x_m\}$
 - 観測しないデータ $Z = \{z_1, \dots, z_m\}$
 - パラメータ化分布 $P(Y|h)$. 但し
 - $Y = \{y_1, \dots, y_m\}$ は完全なデータ $y_i = x_i \cup z_i$
 - h はパラメータ集合
- 求めるもの: h の最尤推定量
 - 観測データの尤度の最大化によって得られる
 - $P(X|h) = \int P(Y|h) dZ$ を最大化する h

一般のEM法

- 完全データ $Y = X \cup Z$ の対数尤度の (X とパラメータ h に対する条件付) 期待値 $Q(h'|h)$ を考える
 $Q(h'|h) = E[\log P(Y|h') | h, X] = \int \log P(Y|h') P(Z|X, h) dZ$
- EMアルゴリズム
 - Estimationステップ: 現在の仮説 h と観測データ X を用いて、 Q を計算する
 - 計算は、現在の h と X のもと $P(Z|X, h)$ を求めることに集約される
 - Maximizationステップ: Q を最大化する h' を求めそれを h とする
- このとき、 $\log P(X|h)$ が単調に非減少である
 - $P(X|h) = \int P(Y|h) dZ$ の極大化が図られる

事例ベース学習

- キーアイデア: 訓練データ $\langle x_i, f(x_i) \rangle$ を全て憶えていよう
- 最近傍法 (Nearest neighbor)
- k -Nearest neighbor
- Locally weighted regression
- Radial basis functions
- Lazy 対 eager

最近傍法

- 最近傍法 (Nearest neighbor)
 - 問合せ x_q に対し、最近接の x_n を見つけ、 $f(x_q) \leftarrow f(x_n)$ とする
- k -Nearest neighbor
 - k 個の最近接データの間で、多数決
 - k 個の最近接データの間で、平均値

最近傍法の特徴

- いつ使うか
 - 属性が R^n の点とみなせる
 - 大体20個以下の属性
 - 大量の訓練データ
- 長所
 - 学習が速い
 - 複雑な目標関数も可能
 - (訓練データがもつ) 情報を失うことがない
- 短所
 - 問合せ時、遅い
 - 無関係な属性で、簡単にごまかされる

極限における振舞い

- $p(x)$: 事例 x がラベル1 (正) をもつ確率
- Nearest neighbor:
 - 事例数 $\rightarrow \infty$ のとき、Gibbsアルゴリズムに漸近
 - Gibbs: 確率 $p(x)$ で1を予測
- k -Nearest neighbor
 - 事例数 $\rightarrow \infty$ かつ k が大きくなると、Bayes最適
 - Bayes最適: $p(x) > 0.5$ なら1、それ以外0

注: Gibbs の期待誤差はBayesの倍以下

距離荷重つき k-NN

- 近い事例の判断を重視したい

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k w_i f(x_i)}{\sum_{i=1}^k w_i}, \quad w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

但し、 $d(x_q, x_i)$ は、 x_q と x_i 間の距離

- これにより、 k 個のみならず全データを使うことに意味がでてくる⇒Shepardの方法

次元の呪い

- 20個の属性で記述されるが、その内、たった2属性のみが意味ある場合を考える
- 次元の呪い:
 - k -NNなら、他の18属性の値でどんな結論も出うる
- 解決方法
 - j 番目の属性に z_j の荷重を。 z_j は予測誤差最小となるように選択
 - cross-validationを用いて自動的に z_j を決定

Locally weighted regression

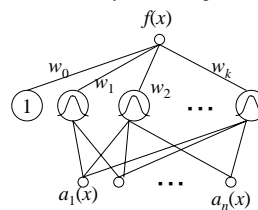
- k -NN は各問合せ x_q で f の局所近似を構成していた
- x_q の周囲で $f(x)$ の近似関数を明示的に構成したらどうだろうか?
 - k -NNに線型回帰したら?
 - 2次回帰では?
 - 区分回帰したら?
- 最小化すべき誤差にもいくつかの候補が

$$E_1(x_q) \equiv \frac{1}{2} \sum_{x \in x_q \text{ の } k\text{-NN}} (f(x) - \hat{f}(x_q))^2$$

$$E_2(x_q) \equiv \frac{1}{2} \sum_{x \in D} (f(x) - \hat{f}(x_q))^2 K(d(x_q, x))$$

Radial Basis Function Network

- 局所近似の線型結合による大域近似
- 神経回路網の一種
- distance-weighted regression に類似
 - lazy ではなく eager であるが



$$f(x) = w_0 + \sum_{u=1}^k w_u K_u(d(x_u, x))$$

$K_u(d(x_u, x))$ の一例

$$K_u(d(x_u, x)) \equiv e^{-\frac{1}{2\sigma^2} d(x_u, x)^2}$$

RBFの学習

- $K_u(d(x_u, x))$ の x_u の定め方
 - 事例空間に一樣にばら撒く
 - 事例を使用(事例の分布が反映)
- 荷重の学習(K_u は正規分布とする)
 - 各 K_u の分散(と平均)を定める
 - 例えば、EMを使用
 - K_u を固定したまま、線型出力部分を学習
 - 線型回帰で高速に

Lazy 対 eager

- Lazy: 事例からの一般化をしないている。問合せがあったときに考える
 - k-Nearest Neighbor
- Eager: 問合せ前に予め一般化しておく
 - 「学習」アルゴリズム、ID3、回帰、RBF,,
- 違いはあるか?
 - Eager学習は全域的な近似を作成
 - Lazy学習は局所近似を大量に作成
 - 同じ仮説空間を使うなら、lazyの方が複雑な関数を作成
 - over-fittingの可能性
 - 柔軟(複雑なところと単純なところの組合せ)



まとめ

- 事例ベースアプローチ
 - 大域的な構造を仮定しない
 - どんな場合にも使える
 - 雑音に弱い(大域構造を用いた平滑化ができない)
 - 次元の呪い