統計的機械学習による音/画像/購買・販売/WEBユーザビリティ情報解析に関する研究

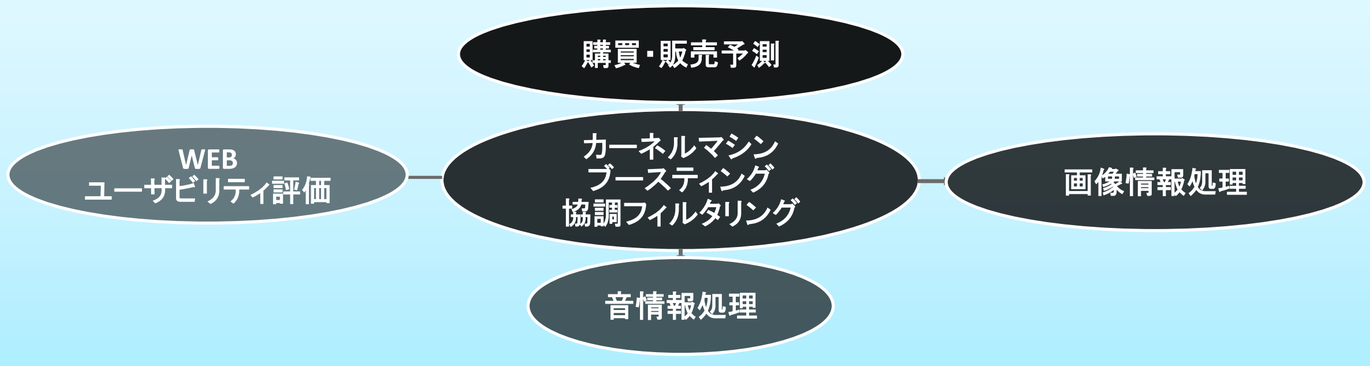
松井　知子　　モデリング研究系　教授

【概要】

本研究室では統計的学習機械を用いて、音声/音楽/画像/WEBユーザビリティデータを処理する方法について研究しています。具体的にはカーネルマシン、ブースティング、協調フィルタリングの手法を用いて、

1. 音声・話者認識
2. 画像識別
3. 購買・販売予測
4. トピック分類
5. 音楽ジャンル分類
6. WEBユーザビリティ評価　など

の研究課題に取り組んでいます。



【統計的機械学習】

- 統計科学を用いて、
  - データから、内在する数学的な構造を発見する。
  - その数学的な構造に基づいて、予測や判別などの情報処理を行う。
- 帰納的アプローチ

  v.s.
- 自然科学でよく見られる演繹的アプローチ
  - 仮説をたて、推論し、実験的または理論的に検証する。

- カーネルマシン
  - 自動的な特徴(/モデル)選択機構を含む。
  - 非線形の扱いに優れている。
  - サポートベクターマシン(SVM)、罰金付ロジスティック回帰マシン
- いろいろな確率モデルによる方法
  - 混合ガウス分布モデル
  - 隠れマルコフモデル
- 協調フィルタリング　など

【協調フィルタリングによる商品販売量の予測】

概要:

- 協調フィルタリングを用いて、日本のあるスーパーマーケットでの商品販売量を予測する方法について検討
  - ✓ 協調フィルタリングは個人や集団の嗜好をモデル化して、そのモデルを他人の行動予測に用いる技術
  - ✓ 個人の各商品の購買数はその商品に対する嗜好を表していると考えて、協調フィルタリングを適用

利用データ:

- POSデータ『CCL-CAFÉデータ』
  - ✓ カスタマー・コミュニケーションズ(株)がスーパーマーケットで収集
  - ✓ 収集期間:2007～2009年(3年間)
  - ✓ 内容:
    - ユーザ情報:ユーザID, 年齢, 性別
    - 商品情報:商品ID, 価格, 購買数, 商品カテゴリー
    - 日付情報:購買日時
  - ✓ サイズ:
    - ユーザ数:145M(女性93M、男性30M、その他22M)
    - 商品数:138M

Matrix Factorization法:

- ユーザID: p_u　商品ID: q_i　購買数: r_{ui}

$$\hat{r}_{ui} = \hat{q}_i^T \hat{p}_u$$

$$O(Q,P) = \tfrac{1}{2} \sum_{(u,i)\in\Omega} \left(r_{ui} - q_i^T p_u\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2\right)$$ [Koren et al., 2009]

繰り返し推定:

$$\sum_{(u',i)\in\Omega} q_i \left(r_{u'i} - q_i^T p_{u'}\right) + \lambda p_{u'} = 0$$

$$\sum_{(u,i')\in\Omega} p_u \left(r_{ui'} - q_{i'}^T p_u\right) + \lambda q_{i'} = 0$$

修正Matrix Factorization法:

- ユーザID: p_u　商品ID: q_i　日付(年、月): s_y , o_m　購買数: r_{ui}

$$\hat{r}_{uimy} = \hat{q}_i^T \hat{p}_u + \hat{q}_i^T \hat{o}_m + \hat{q}_i^T \hat{s}_y + \hat{p}_u^T \hat{o}_m + \hat{p}_u^T \hat{s}_y + \hat{o}_m^T \hat{s}_y$$

$$O(Q,P) = \tfrac{1}{2} \sum_{(u,i,m,y)\in\Omega} \left|r_{ui} - q_i^T p_u - q_i^T o_m - q_i^T s_y - p_u^T o_m - p_u^T s_y - o_m^T s_y\right|$$

$$+ \lambda\left(\|q_i\|^2 + \|p_u\|^2 + \|o_m\|^2 + \|s_y\|^2\right)$$

滑降シンプレックス法[Nelder and Mead, 1965]で推定

実験:

- データ
  - ✓ 学習:97M サンプル
  - ✓ テスト:0.4M サンプル(3セット、ランダムに選択)
  - ✓ 購買数が少ない商品を排除(しきい値:20、もしくは100)
- モデルパラメータ
  - ✓ $\lambda = 20$
  - ✓ 潜在ベクトルの次元 $d \in \{2, 5, 10\}$
- ベースライン

$$\hat{r}_{ui} = \frac{1}{N_i} \sum_{u,m,y} r_{uimy} \qquad \hat{r}_{uim} = \frac{1}{N_{im}} \sum_{u,y} r_{uimy} \qquad \hat{r}_{uimy} = \frac{1}{N_{imy}} \sum_{u} r_{uimy}$$

表 1　ベースラインの推定結果:テストセットに対する誤り $\|r - \hat{r}\|$ の平均.

Estimator	Mean test error
$\hat{r}_{ui}$	1.21598
$\hat{r}_{uim}$	1.15113
$\hat{r}_{uimy}$	1.13248

※購買数が20以下の商品を排除

表 2　λ と n を変えた時の協調フィルタリングの結果. 学習とテストセットに対する誤り $\|r - \hat{r}\|$ の平均.

Latent vectors	Dimension n	Train error	Test error
u,i	2	0.5583	0.9661
u,i	5	0.5255	0.9612
u,i	10	0.5346	0.9678
u,i,m	2	0.5312	0.9505
u,i,m	5	0.5175	0.9547
u,i,m	10	0.5138	0.9477
u,i,m,y	2	0.5727	1.0200
u,i,m,y	5	0.5751	1.0354
u,i,m,y	10	0.5343	0.9922

※購買数が100以下の商品を排除

【アンサンブルカルマンフィルタを用いた購買行動の予測】

概要:

- データ同化の一手法であるアンサンブルカルマンフィルタを用いて、小売店における消費者の購買行動を予測する 。

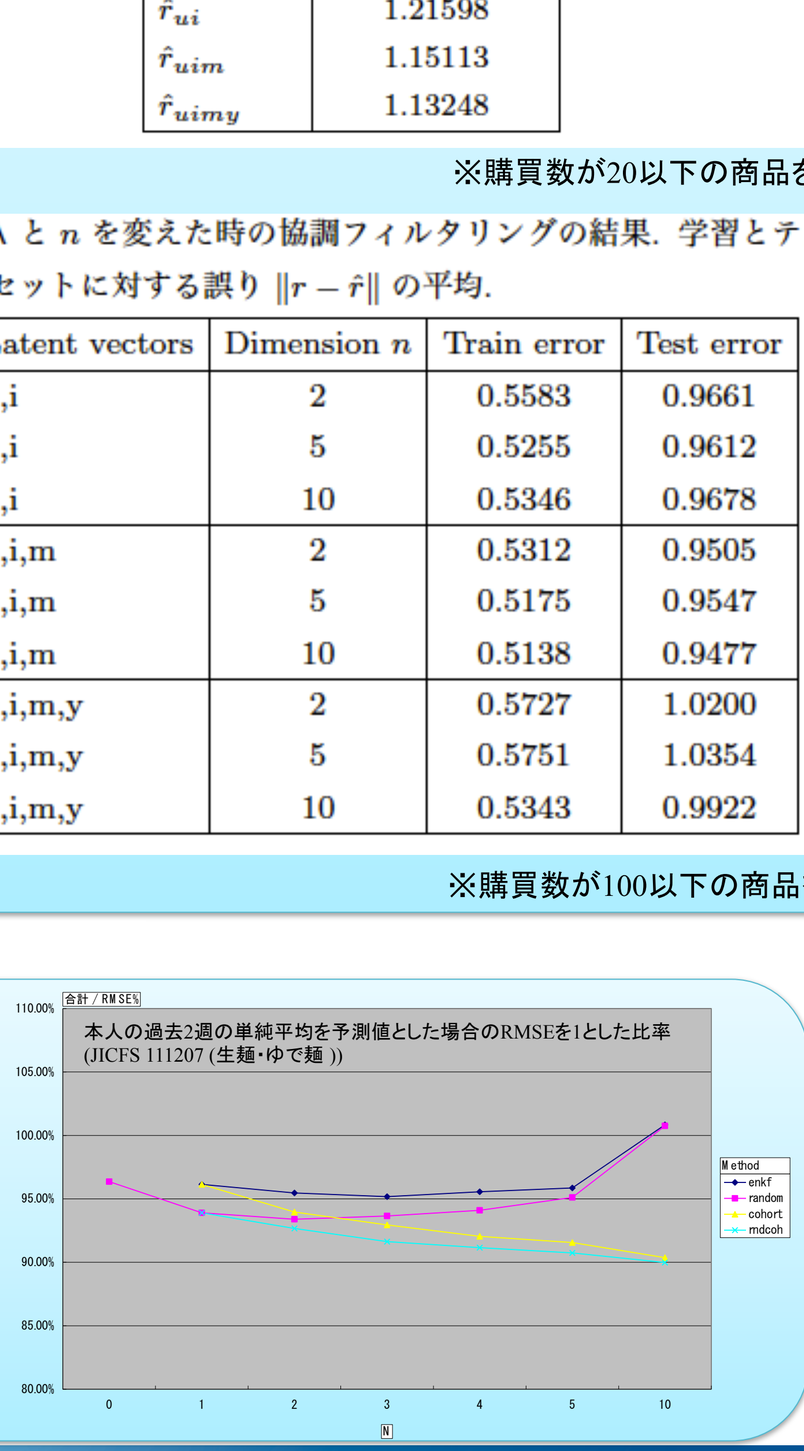
販売行動モデル:

- 本人とそのコホートユーザによる自己回帰移動平均モデル
  - ✓ 本人sとそのコホートユーザ$\{c_1, c_2, \ldots\}$の第t週における購入金額$y_t^s, y_t^{c1}, y_t^{c2}, \ldots$は、過去$p$週の購入金額によって決定

$$\mathbf{y}_t = \left(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{t-p}\right)\mathbf{x}_t + w_t \qquad \mathbf{y}_t = \left(y_t^s, y_t^{c1}, y_t^{c2}, \ldots\right)^T$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} = \left(a_1, a_2, \ldots, a_p\right)^T + v_t$$

- コホートユーザの選択
  - ① 予備実験からn人を選択 (EnKF)
  - ② ランダムにn人を選択 (random)
  - ③ 上記①のn人の平均 (cohort)
  - ④ 上記②のn人の平均 (rndcoh)



本研究室では統計的機械学習とその応用研究に興味のある学生さんを募集しています！