Feature Selection - ICML 07

Presenter: Zheng (Alan) Zhao
Feature Selection

1. Feature Selection in Kernel Space
2. Minimum Reference Set Based Feature Selection for Small Sample Classifications
3. Supervised Feature Selection via Dependence Estimation
4. Spectral Feature Selection for Supervised and Unsupervised Learning
5. Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks
In Bayesian model, the high level idea of feature relevance can be transferred among tasks through Meta-Level Prior.

\[ y = w^T x, \quad P(w \mid D) \propto P(D \mid w)P(w) \]

The work uses hierarchical Bayesian model.

- Specify distributions on different level
  - Observation–Response, Parameter, Hyper–Parameter
- Transfer knowledge on Hyper–Parameter Level.
  - Using Meta–features to generate priors
Problem Setup

Meta-features used in our collaborative filtering domain

1. **Genre**: Whether the movies share a particular genre.
2. **Decade**: Whether both movies were released in 2000s, 1990s, 1980s, 1970s or a previous decade.
3. **Actors/Directors**: How many and which actors/directors the movies have in common.
4. **Keywords**: How many and which keywords the movies have in common.
Technical Details (1)

- Traditional Learning Model:
  \[
  P(y_r = 1 \mid x_r, w_r) = \frac{1}{1 + \exp(-w_r^T x_r)}
  \]
  \[
  \gamma: P(w \mid \gamma) = \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{w^2}{2\gamma}\right)
  \]

- In the paper, authors assume different features have different variance.
  \[
  P(w_{rk} \mid \beta, f_{rk}) = \frac{1}{\sqrt{2\pi\beta^T f_{rk}}} \exp\left(-\frac{w_{rk}^2}{\beta^T f_{rk}}\right)
  \]
  \[
  P(\gamma_{rk}) \propto \gamma_{rk}^{D-1} \exp(-C\gamma_{rk}) \quad \gamma_{rk} \geq 0
  \]
  \[
  \gamma_{rk} = \beta^T f_{rk}
  \]
Joint Conditional Distribution

\[
P(W, \beta | X, F, Y) = \prod_{r=1}^{R} \prod_{m=1}^{M} P(y_r[m] | x_r[m], w_r) \cdot \prod_{r=1}^{R} \prod_{k=1}^{K_r} P(w_{rk} | \beta, f_{rk})P(\gamma_{rk}),
\]

Joint maximum a posteriori estimation

\[
P(W, \beta | X, F, Y) = \sum_{r=1}^{R} \sum_{m=1}^{M} \log P(y_r[m] | x_r[m], w_r) - \sum_{r=1}^{R} \sum_{k=1}^{K_r} \left( \frac{w_{rk}^2}{\beta^\top f_{rk}} + C\beta^\top f_{rk} + \text{Const} \right)
\]
The proposed problem is a convex problem. The authors use a coordinate ascent procedure over the two sets of parameters $W$ and $\beta$.

\[
\arg\min_{W} \sum_{r=1}^{R} \sum_{m=1}^{M} - \log P(y_r[m] \mid \mathbf{x}_r[m], \mathbf{w}_r) + \sum_{r=1}^{R} \sum_{k=1}^{K_r} \frac{w_{rk}^2}{\beta^\top \mathbf{f}_{rk}}.
\]

\[
\arg\min_{\beta} \sum_{r=1}^{R} \sum_{k=1}^{K} \left( \frac{w_{rk}^{(t)}^2}{\beta^\top \mathbf{f}_{rk}} + C\beta^\top \mathbf{f}_{rk} \right)
\]

subject to $\beta^\top \mathbf{f}_{rk} > 0$. 

**Technical Details (3)**
Results (1)
Results (2)

![Graph showing RMSE vs. Training set Size]

- **Baseline**
- **Meta-level Prior**
Results (3)

The graph shows the relationship between the Training set Size and the RMSE for two different methods: Baseline and Meta-level Prior. As the Training set Size increases, the RMSE decreases for both methods. The Baseline method (solid line) consistently has a lower RMSE compared to the Meta-level Prior method (dashed line).
Thanks. Any Question?