Multi-Task SVM
Feature Selection

... or Convex Meta Learning for Discriminatively Finding Features and Kernels

Tony Jebara (Columbia) & Tommi Jaakkola (MIT)
**Multi Task Representations**

*Meta Learning:* (Caruana, Thrun, Baxter)
Use multiple related tasks to improve learning
Typically implemented in Neural Nets (local minima) with a shared representation layer and input layer

*SVMs:* Solve for a single classification/regression model
Can we combine multi SVMs for different tasks yet with a shared input space and learn a common representation

*MED:* A probabilistic approach to SVMs permitting many extensions
Solves multiple SVM models sharing by selecting a representation via a convex program with unique solution

*Selection:* The representations we consider here with MED:
Linear Feature Selection (Jebara, Weston)
Nonlinear Kernel Selection (Lanckriet, Cristianini)
Given training examples: binary (+/- 1) labels: discriminant function:

Minimize penalty function: with classification constraints:

Solve QP
Get widest margin model \( \Theta^* \)
BUT: not probabilistic, no priors, no flexible models
Maximum Entropy Discrimination Approach

Many solutions may be valid.

Solve for distribution $P(\Theta)$ over all good $\Theta$ (instead of $\Theta^*$).

Find $t \int \int \int$ that mins $t \int t \int \int \int$ subject to constraints:

\[
\int t \int \int t \int t \int \int \int \int t \int \int \int \int t
\]

For non-separable, integrate over distribution over models & margins (favoring large margins) $t \int \int \int$.
Maximum Entropy Discrimination Solution

Analytic and Unique:

\[
\begin{align*}
\text{partition function:} & \quad t \int \int \int \int \frac{1}{t} \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \int \\
\text{dual objective to max:} & \quad t \int \int \int \int \int \int \int \\
\text{+ve Lagrange multipliers:} & \quad \int \int \int \int \left\{ \int \int \int \int \right\} \\
\end{align*}
\]

Gaussian mean prior and linear \( L(X; \Theta) \) gives back SVM

\[
\begin{align*}
\text{MED Generalization Guarantees: Sparsity, VC-Dimension, PAC-Bayes}
\end{align*}
\]
Feature Selection

Purpose: pick 100 of 10000 features to get largest margin classifier (NP)

Turn features on/off via binary switches $t_t \{ \Sigma \}

Switch Prior: Bernoulli distribution

$\rho$ parameter varies pruning level $t_t \{ \Sigma \}

MED uniquely & efficiently finds discriminative feature subset, analytic partition fn:

$t \{ \Sigma \}

The Admissible Set
Feature Selection

Objective Function for Classification

\[ t \sum_{i=1}^{t} \left( \sum_{j=1}^{t} \left( \sum_{k=1}^{t} \left( \sum_{l=1}^{t} \left( \sum_{m=1}^{t} \left( \sum_{n=1}^{t} \left( t_{i,j,k,l,m,n} \right) \right) \right) \right) \right) \right) \]

Epsilon-Tube Regression also straightforward

Example: Intron-Exon Protein Classification:
UCI: 240 dims; 200 train, 1300 test
Meta Feature Selection

**Given** series of tasks: map inputs to binary:

using M discriminants with 1 feature selection vector:

Subject to MED classification constraints:

Solve by optimizing joint objective function for all Lagrange
Meta Feature Selection

Example: To ensure coupled tasks, turn multi-class data set into multiple 1 versus many tasks

UCI Dermatology Dataset: 200 trains, 166 tests, 33 features, 6 classes

Variable Feature Selection & Regularization Levels

Cross-validating over Regularization Levels
Meta Feature Selection for Regression

D. Ross Cancer Data: 67 expression level feats.
Use subset of 800 genes to predict all others

Compared with random feature selection
Kernel Selection

Purpose: pick mixture of subset of Kernel matrices to get largest margin classifier, (learn the Gram matrix)

Turn kernels on/off via binary switches $t_i \int \{1, \int \}

Switch Prior: Bernoulli distribution $t_i, t \int \int \int t_i \int 1 \int 1 t_i

Discriminant uses N models with multiple nonlinear mappings of datum

MED solution has analytic concave objective fn:
Meta Kernel Selection

**Given** series of tasks with common (unknown) kernel matrix:
using M discriminants with 1 feature selection vector:

Subject to MED classification constraints:

Solve by optimizing joint objective function for all Lagrange
Meta Kernel Selection

UCI Isolet data set (letter recognition from audio)
26 Classes used as 1 to Many Binary Classification

200 training
600 testing

SVM (X’s)
Kernel Selection (red)
Meta Kernel Selection (blue)

Used rho = 0.1
Used rho = 0.01
Used rho = 0.001
Meta Feature Segmentation (Current Work)

*Given* single task, single model, but each of the T points has its own feature selection configuration.

use 1 discriminant with T feature selection vectors:

\[ t \int \int t \int t \int t \int t \int t \int t , \]

*Subject to* MED classification constraints:

\[ \int t \int \int t \int t \int t \int t \int t \int t , \]

*Solve* transductively by computing distribution over unlabeled
Conclusions and Ongoing Work

Meta Learning and Representation Learning can be applied to SVMs for both classification & regression

MED permits unique SVM solution with
- Learning feature selection
- Learning kernel selection

MED permits straightforward Meta learning extensions
- Meta Learning feature selection
- Meta Learning kernel selection

Feature selection helps performance
Metalearning can help performance if tasks are coupled

Segmentation inverts multiplicity: single model, multiple selections