Multi-Task SVM
Feature Selection

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

Tony Jebara (Columbia) & Tommi Jaakkola (MIT)
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)
Classification - Discriminative SVMs

**Given** training examples:

\[
\{X_1, \ldots, X_T\}
\]

binary (+/- 1) labels:

\[
\{y_1, \ldots, y_T\}
\]

discriminant function:

\[
L(X; \Theta) = \Theta^T X + b
\]

**Minimize** penalty function:

\[
\|\Theta\|^2
\]

with classification constraints:

\[
y_t L(X_t; \Theta) - \gamma_t \geq 0, \forall t
\]

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 $P(\Theta)$ that mins $KL(P \| P_0)$ subject to constraints:

$$\int P(\Theta) \left[ y_t L(X_t; \Theta) - \gamma \right] d\Theta \geq 0, \quad \forall t$$

$$\hat{y} = \text{sign} \int P(\Theta) L(X; \Theta) \, d\Theta$$

For non-separable, integrate over distribution over models & margins (favoring large margins) $P(\Theta, \gamma)$

Information Projection

The Admissible Set
Maximum Entropy Discrimination Solution

**Analytic and Unique:**

\[ P (\Theta, \gamma) = \frac{1}{Z(\lambda)} P_0 (\Theta, \gamma) \exp \left( \sum_t \lambda_t [y_t L(X_t; \Theta) - \gamma_t] \right) \]

partition function:

\[ Z(\lambda) = \int P(\Theta, \gamma) \]

dual objective to max:

\[ J(\lambda) = -\log Z(\lambda) \]

+ve Lagrange multipliers:

\[ \lambda = \{ \lambda_1, \ldots, \lambda_T \} \]

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

\[ J(\lambda) = \sum_t \left[ \lambda_t + \log \left( 1 - \frac{\lambda_t}{C} \right) \right] - \frac{1}{2} \sum_{t, t'} \lambda_t \lambda_{t'} y_t y_{t'} (X_t^T X_{t'}) \]

MED Generalization Guarantees: Sparsity, VC-Dimension, PAC-Bayes
Feature Selection

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

Turn features on/off via binary switches \( s_i \in \{0, 1\} \)

**Switch Prior:** Bernoulli distribution

\[ P_{s,0}(s_i) = \rho^s_i (1 - \rho)^{1-s_i} \]

\[ L(X; \Theta) = \sum_i s_i \theta_i X_i + b \]

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

\[ Z(\lambda) = \int P(\Theta, \gamma, s) \]

*The Admissible Set*
Feature Selection

Objective Function for Classification

\[ J(\lambda) = \sum_t \left[ \lambda_t + \log \left( 1 - \frac{\lambda_t}{c} \right) \right] - \sum_{i=1}^n \log \left[ 1 - \rho + \rho e^{\frac{1}{2} \left( \sum_t \lambda_t y_t X_{t,i} \right)^2} \right] \]

Epsilon-Tube Regression also straightforward

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

Given a series of tasks: \( m \in [1..M] \)

map inputs to binary: \( X_{tm} \rightarrow y_{tm} \quad \forall t \in [1..T_m] \)

using \( M \) discriminants with 1 feature selection vector:

\[
L(X;s, \theta_m, b_m) = \sum_i s_i \theta_{m,i} X_i + b_m
\]

Subject to MED classification constraints:

\[
\int P(s, \theta_1, \ldots, \theta_M, b_1, \ldots, b_M) \left[ y_{tm} \left( X_{tm}; s, \theta_m, b_m \right) - \gamma \right] d\Theta \geq 0, \quad \forall t \forall m
\]

Solve by optimizing joint objective function for all Lagrange multipliers:

\[
J(\lambda) = \sum_{t,m} \left[ \lambda_{tm} + \log \left( 1 - \frac{\lambda_{tm}}{c} \right) \right] - \sum_{i=1}^n \log \left( 1 - \rho + \rho \exp \left( \frac{1}{2} \sum_{m=1}^M \sum_{t=1}^{T_m} \lambda_{tm} y_{tm} X_{tm,i} \right) \right)
\]
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 \( s_i \in \{0, 1\} \)

**Switch Prior:** Bernoulli distribution \( P_{s_i,0}(s_i) = \rho^{s_i} (1 - \rho)^{1-s_i} \)

Discriminant uses N models with multiple nonlinear mappings of datum

\[
L(X; \Theta) = \sum_i s_i \theta_i^T \Phi_i(X) + b
\]

MED solution has analytic concave objective fn:

\[
J(\lambda) = \sum_t \left[ \lambda_t + \log \left( 1 - \frac{\lambda_t}{c} \right) \right] - \sum_{i=1}^n \log \left[ 1 - \rho + \rho \exp \left( \frac{1}{2} \sum_{t=1}^T \sum_{t'=1}^T \lambda_t \lambda_{t'} y_t y_{t'} K(X_t, X_{t'}) \right) \right]
\]
Meta Kernel Selection

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

\[ L \left( X; s, \Theta_m, b_m \right) = \sum_i s_i \theta^T_{m,i} \Phi_i \left( X \right) + b_m \]

**Subject to** MED classification constraints:

\[ \int P \left( s, \Theta_1, \ldots, \Theta_M, b_1, \ldots, b_M \right) \left[ y_{tm} L \left( X_{tm}; s, \Theta_m, b_m \right) - \gamma \right] dsdb \Theta \geq 0, \ \forall t \forall m \]

**Solve by** optimizing joint objective function for all Lagrange

\[ J \left( \lambda \right) = \sum_{t,m} \left[ \lambda_{tm} + \log \left( 1 - \lambda_{tm} / c \right) \right] - \sum_{i=1}^n \log \left( 1 - \rho + \rho \exp \left( \frac{1}{2} \sum_{m=1}^M \sum_{t=1}^T \sum_{t'=1}^T \lambda_{tm} \lambda_{tm'} \gamma_{tm} y_{tm} y_{tm'} K \left( X_{tm}, X_{tm'} \right) \right) \right] \]
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:

\[ L(X_t; s_t, \theta, b) = \sum_i s_{t,i} \theta_i X_{t,i} + b \]

**Subject to** MED classification constraints:

\[ \int P(s_1, \ldots, s_T, \theta, b) \left[ y_t L(X_t; s_t, \theta, b) - \gamma \right] d\Theta \geq 0, \forall t \]

**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*