Advanced Methodologies for Predictive Learning

Vladimir Cherkassky
University of Minnesota
cherk001@umn.edu

Presented at MinneAnalytics Forum, March 31, 2011

OUTLINE

- **Motivation and Background**
  - standard inductive learning
  - assumptions
- **Advanced Methodologies**
  - Transduction and semi-supervised learning
  - Universum learning
  - Multi-Task Learning
  - Learning Using Privileged Information
- **Practical Issues and Concerns**

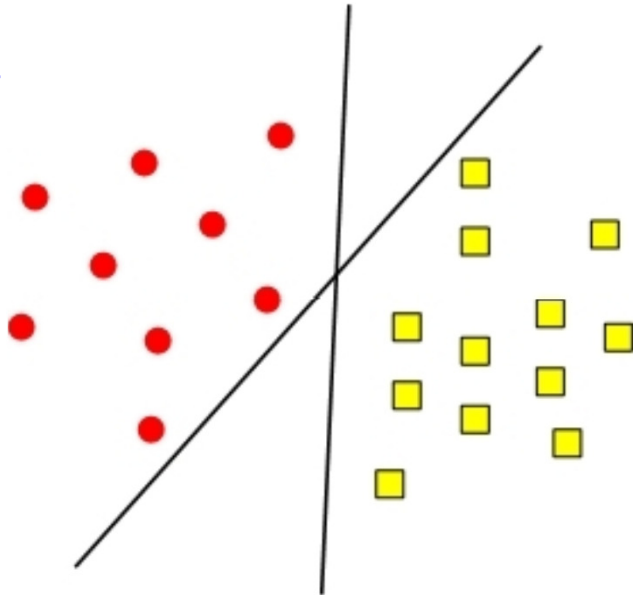
Standard Inductive Learning - 1

Given: data samples (~ *training data*)

Estimate: a model (function) that can

- explain this data
- predict future *test data*

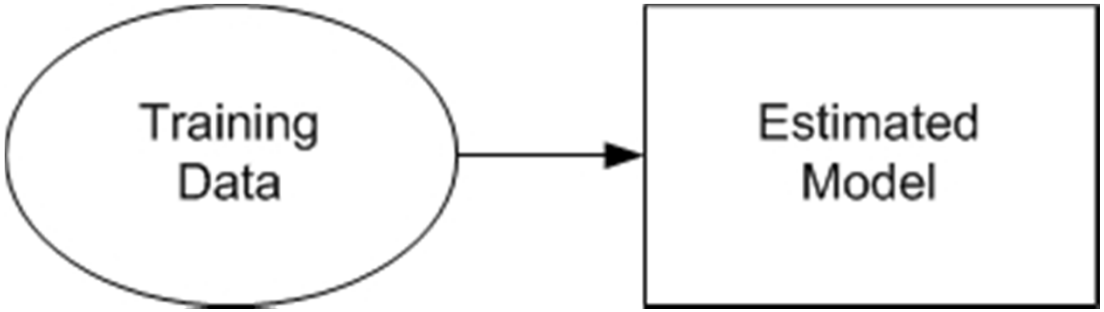
Classification problem:



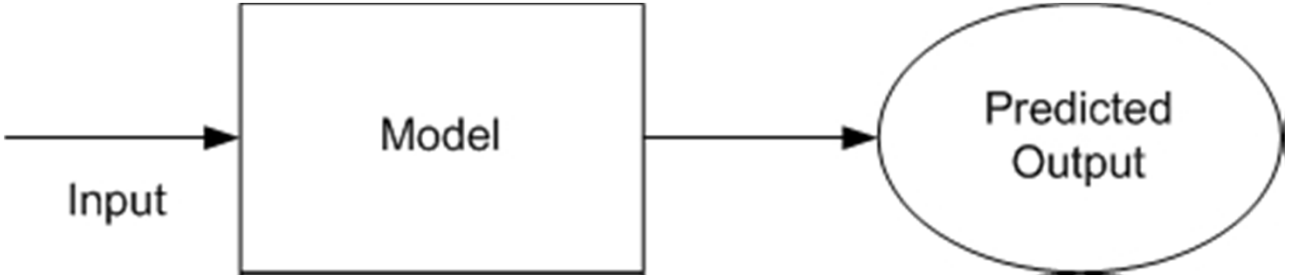
Note: training data is similar to test data

Standard Inductive Learning -2

Induction ~ function estimation from data:



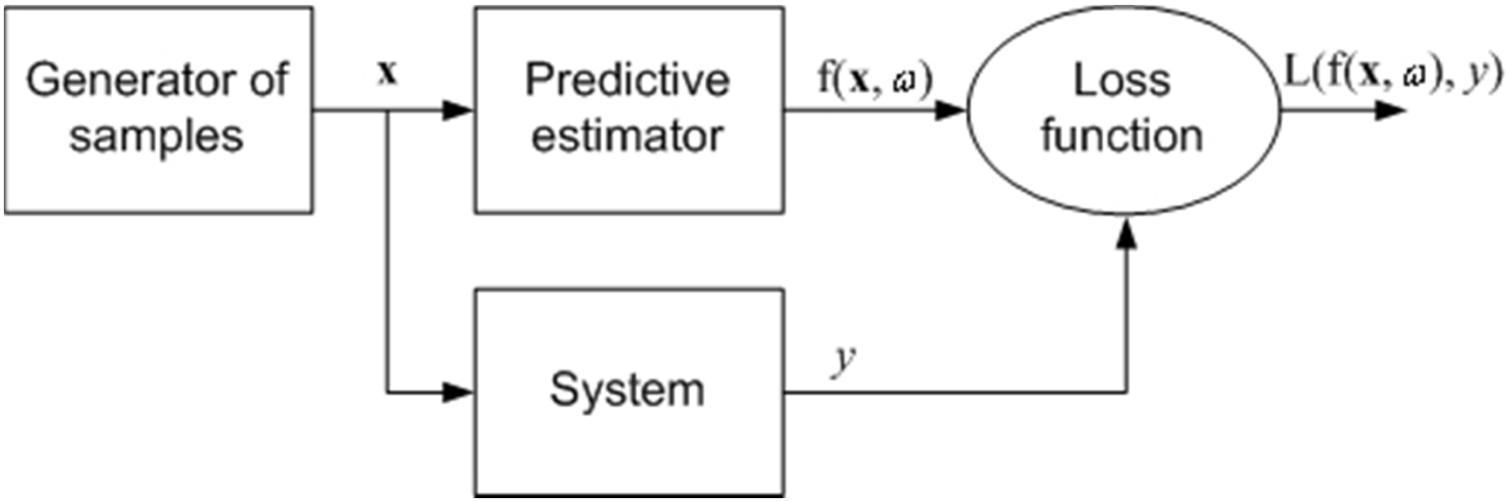
Deduction ~ prediction for new inputs:



Note: All practical methods follow this paradigm

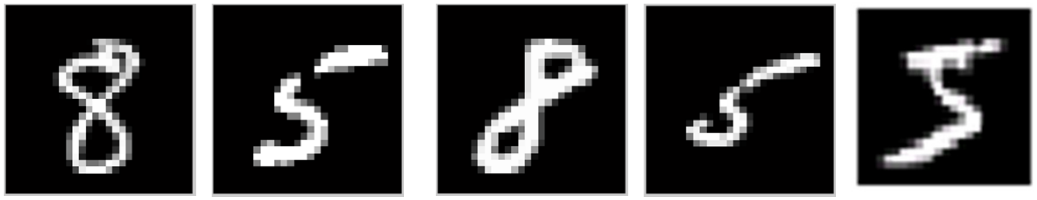
Assumptions for Inductive Learning

- Finite training set (**x** ,*y*), large test set, i.i.d. sampling
- Test inputs unknown during learning
- Training and future (test) data from the same *P(**x**,y)*
- Given loss function – e.g. MSE, error rate, …
- Performance index is minimization of test error



Non-standard Learning Settings

- Example: Handwritten digit recognition 5 vs 8



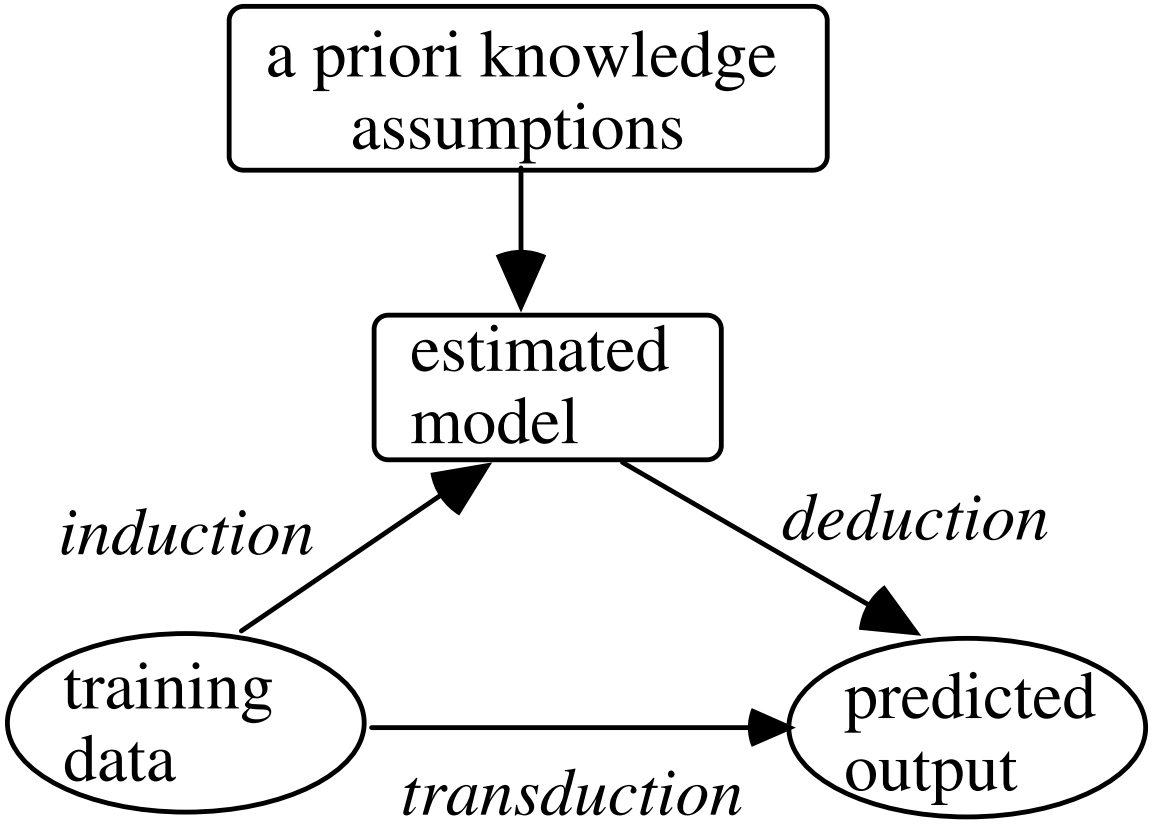
each example ~ 28x28 pixel image → 784-dimensional input **x**

- **Inductive learning**: estimate a model from training data
- **Transduction and SSL**: **x**-values of test inputs also known
- **Universum learning**: additional samples ('other digits')
- **Multi-Task Learning**: training and test data is generated by t different persons. *Need to estimate t related classifiers*
- **Learning Using Privileged Information**: training data is generated by t persons. *Need to estimate a single classifier*

Transduction and SSL

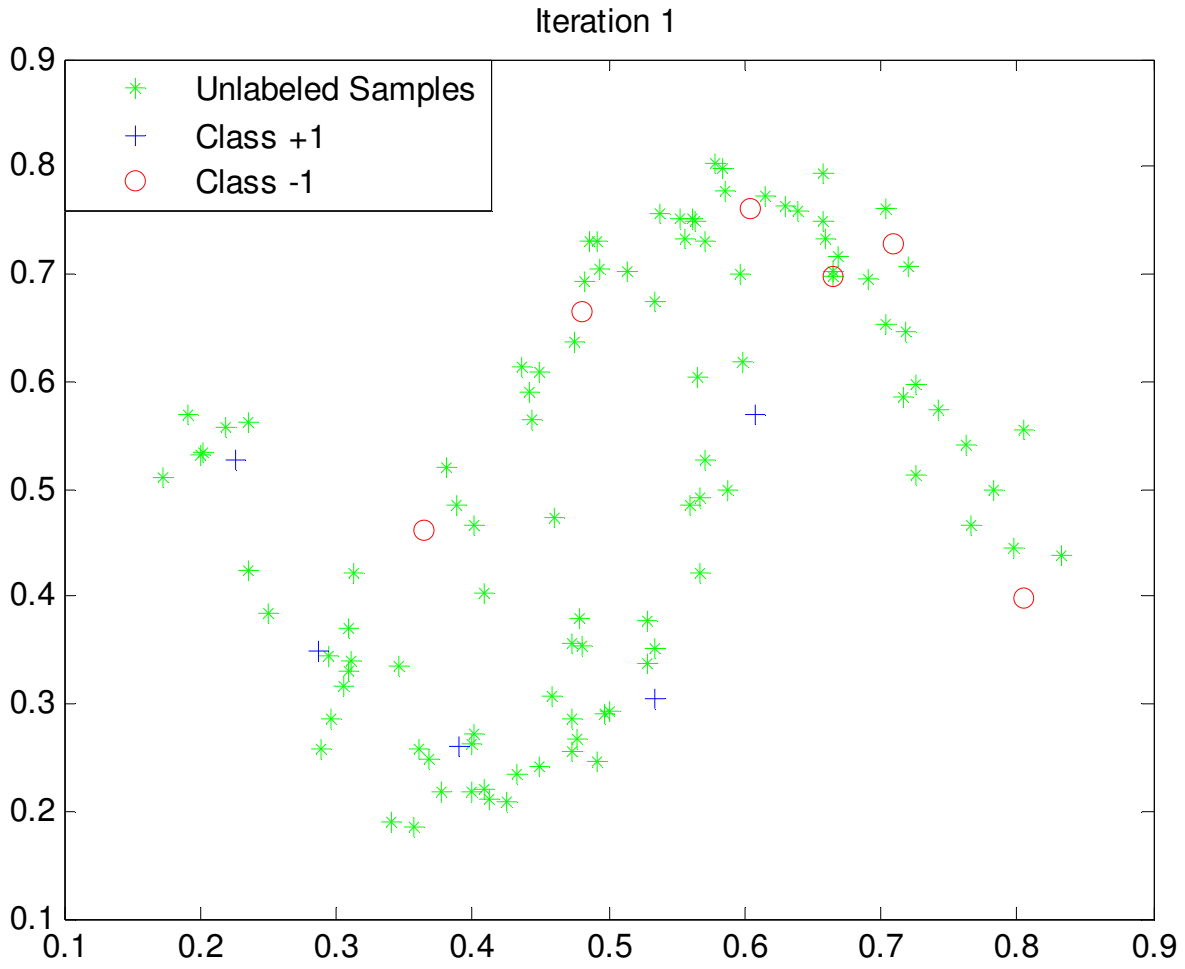
Given: labeled data L + unlabeled data U
(from the same distribution)

Estimate: predictive model



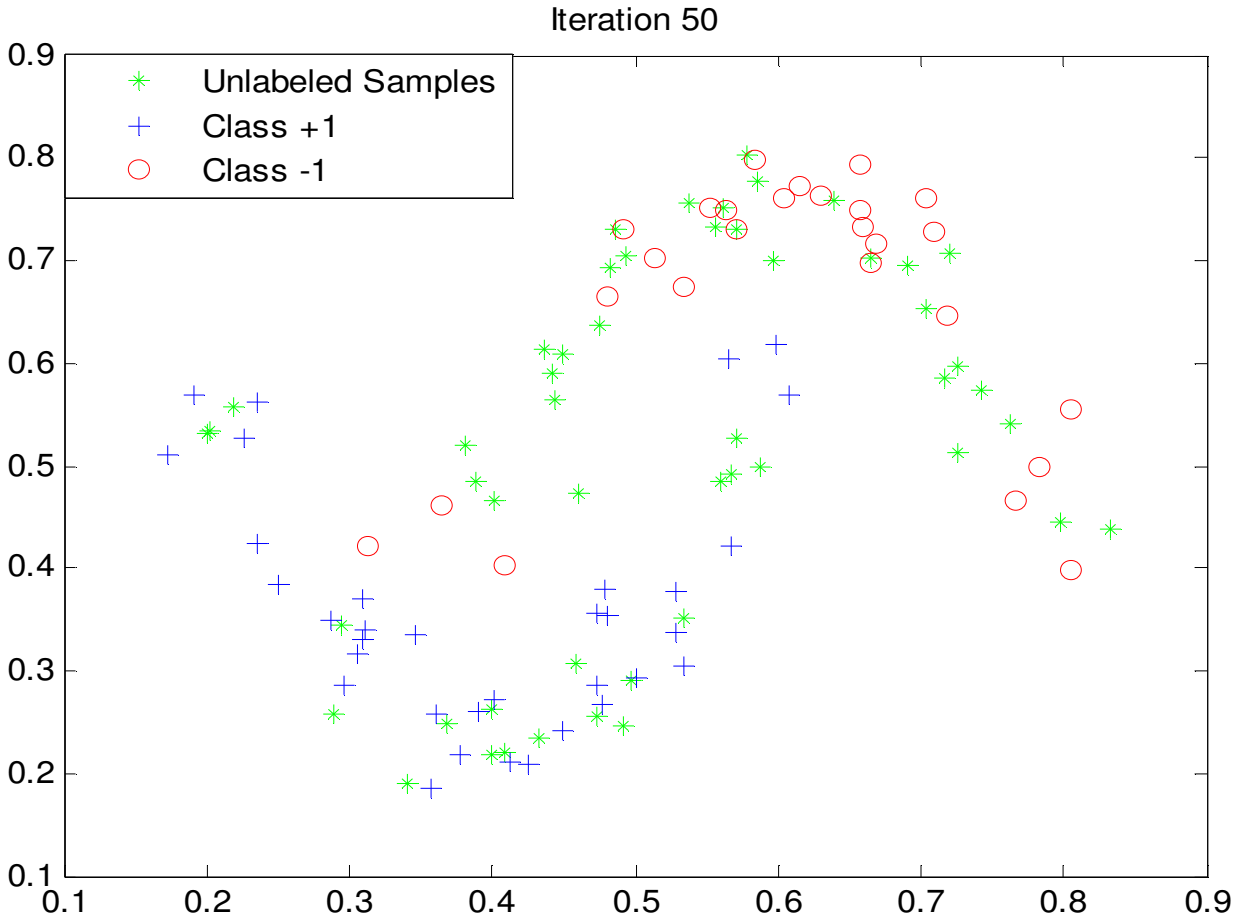
Self-Learning Algorithm for SSL (1)

Given: labeled data L and unlabeled U: U >> L



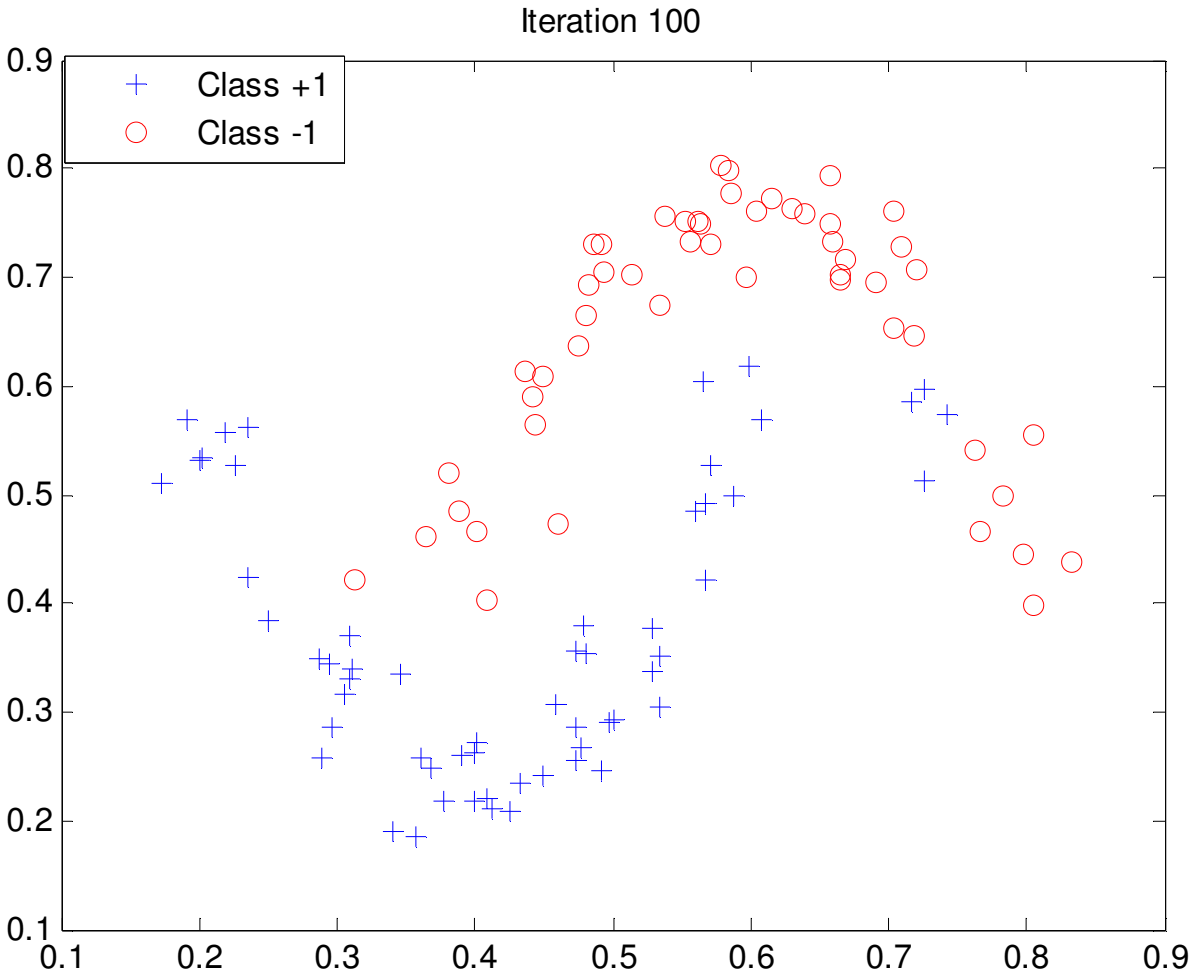
Self-Learning Algorithm for SSL (2)

Classification of test data: after 50 iterations



Self-Learning Algorithm for SSL (3)

Classification of test data: after 100 iterations



Many applications for transduction

- Text categorization: classify word documents into a number of predetermined categories
- Email classification: Spam vs non-spam
- Web page classification
- Image database classification
- All these applications:
  - high-dimensional data
  - small labeled training set (human-labeled)
  - large unlabeled test set

Universum Learning

Given: labeled data L + unlabeled data U
(known *NOT to belong* to either class)

Estimate: predictive model

- **Universum samples:** reflect a priori knowledge about application domain
- **For digit recognition problem 5 vs 8:**
  - other digits 0, 1, 2, 3, 4, 6, 7, 9
  - handwritten letters
  - averaging of randomly selected 5 and 8
- Aka Learning through Contradiction

Random Averaging for digits 5 and 8

- Two randomly selected examples


- Universum sample (via Random Averaging):

Cultural Interpretation of the Universum

- Absurd examples, jokes, some art forms:

Dadaism – M. Duchamp

Multi-Task Learning

- **Application: Handwritten digit recognition**

  Labeled training data provided by *t* persons (t >1)

  Goal: to estimate *t* related classifiers (for each person)

- **Application: Medical diagnosis**

  Labeled training data provided by *t* groups of patients (t >1), say men and women (t = 2)

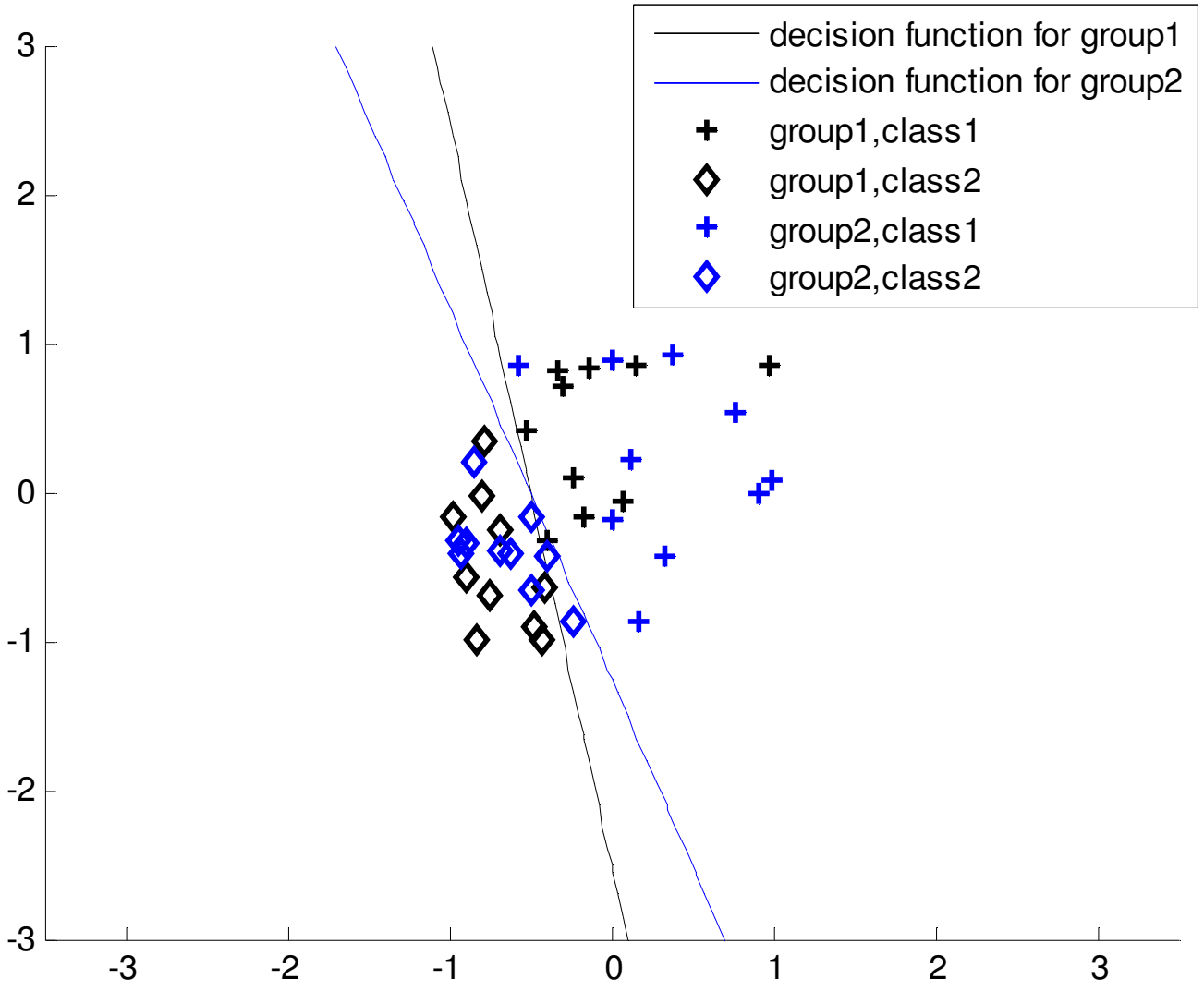
  Goal: to estimate *t* classifiers specialized for each group of patients ~ MTL

- **Practical issues**

  - how to model task-relatedness?

  - with small samples, it may be better to ignore the group (task) label

Multi-Task Learning: Simple Example



Learning Using Privileged Info (LUPI)

Given: training data ($\mathbf{x}$, $\mathbf{x}^*$, y)

where $\mathbf{x}^*$ is privileged info *only for training data*

Estimate: predictive model y = f($\mathbf{x}$)

Common in many applications:

- **Medical diagnosis:** privileged info ~ results of expensive tests, patient history *after* initial diagnosis, genetic data etc.
- **Time series prediction:** privileged info ~ future values of the training time series

LUPI and SVM+

SVM+ is a new learning technology for LUPI

- **Improved prediction accuracy** (proven both theoretically and empirically)
- **Similar to learning with a human teacher**: 'privileged info' ~ feedback from an experienced teacher(can be very beneficial)
- **SVM+ is perhaps the most promising new method for 'difficult' problems**
- **Computationally complex:** no public domain software currently available

Practical Concerns

- Availability of public-domain software
  - **YES** transduction, SSL, Universum, MTL
  - **NO** for SVM+ (LUPI)
- Methodological:
  new technologies require ***understanding*** of predictive methodology +application domain
- Model selection/ parameter tuning
  - very hard (typically 3-5 parameters)
- Practical advantages of these new methodologies ***cannot be guaranteed***

References and Resources

- Vapnik, V. *Estimation of Dependencies Based on Empirical Data*. *Empirical Inference Science: Afterword of 2006*, Springer, 2006
- Cherkassky, V. and F. Mulier, *Learning from Data*, second edition, Wiley, 2007
- Cherkassky, V. and Y. Ma, *Introduction to Predictive learning*, Springer, 2011 (to appear)

Public-domain SVM software

- Main web page link http://www.kernel-machines.org
- LIBSVM software library http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SVM-Light software library http://svmlight.joachims.org/
- Non-standard SVM-based methodologies: Universum, SVM+, MTL http://www.ece.umn.edu/users/cherkass/predictive_learning/

Short Course on Predictive Modeling

- Offered at ECE Dept, U of M
- Half-a-day course on June 1, 2011
  See course description at
  http://www.ece.umn.edu/ECETECHTUNEUP.html
- Course Contents:
  - Predictive data modeling methodology
  - Support Vector Machines (SVM)
  - Practical issues and application studies
  - Advanced SVM-based technologies