Non-Inductive Approaches for Learning with Sparse Data

Vladimir Cherkassky
University of Minnesota
cherk001@umn.edu
Presented at IJCNN-07

Acknowledgements

- Joint work with (former) graduate students at UMN
  Y. Ma (Honeywell Labs), T. Xiong (eBay), L. Liang

- Special thanks to V. Vapnik (NEC Labs) for discussions on non-inductive learning methods and VC-falsifiability

OUTLINE

- **Background**

  uncertainty and risk-taking

  predictive learning

  philosophy and induction
- Inductive learning and VC-theory
- Motivation for non-inductive approaches
- Non-inductive learning formulations
- Summary

Handling Uncertainty and Risk(1)

- Probability for quantifying uncertainty
  - degree-of-belief
  - frequentist (Pascale, Fermat)
  - probability theory and statistics (20^{th} century)
- Modern science: causal determinism
  (A. Einstein)

→Goal of science: estimating a true model or system identification

~ estimation of statistical distribution

Handling Uncertainty and Risk(2)

- Making decisions under uncertainty
  = risk management
- Probabilistic approach:
  - estimate probabilities (of future events)
  - assign costs and minimize expected risk
- Risk minimization approach:
  - apply decisions to known past events
  - select one minimizing expected risk
- Common in all living things: learning, generalization

Predictive learning: why and how

- The problem of predictive learning

  Given past data + reasonable assumptions

  Estimate unknown dependency for future predictions
- Driven by applications (not theory)
- 3 parts of Predictive Learning:

  - conceptual/ philosophical

  - mathematical (technical)

  - practical (implementations, applications)

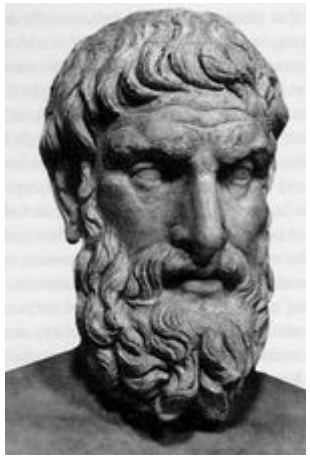
Philosophy of science and induction

- Oxford English dictionary:

  Induction is the process of inferring a general law or principle from the observations of particular instances.
- Clearly related to Predictive Learning.
- All science and (most of) human knowledge involves induction
- How to form ‘good’ inductive theories?

Background: philosophy

William of Ockham: entities should not be multiplied beyond necessity

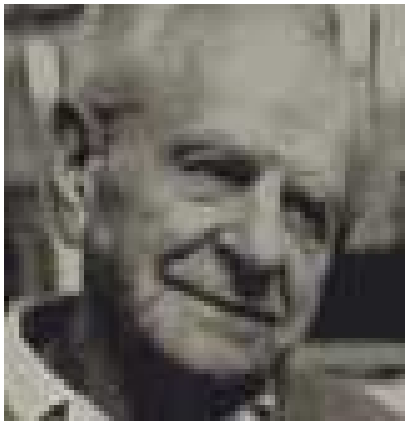


Epicurus of Samos: If more than one theory is consistent with the observations, keep all theories

Background: philosophy

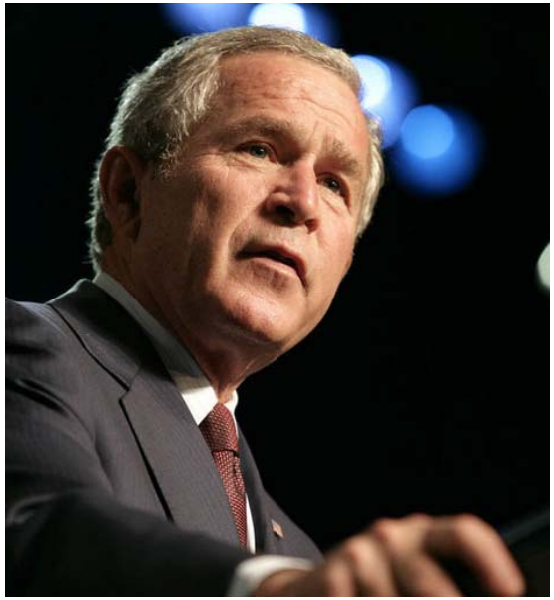
Thomas Bayes:

How to update/ revise beliefs in light of new evidence



Karl Popper: Every true (inductive) theory prohibits certain events or occurences, i.e. it should be **falsifiable**

Background: philosophy



George W. Bush:

I am The Decider

Observations, Reality and Mind

Philosophy is concerned with the relationship btwn

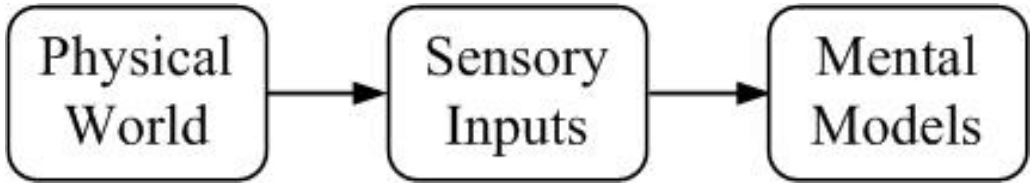
- Reality (Nature)
- Sensory Perceptions
- Mental Constructs (interpretations of reality)

Three Philosophical Schools

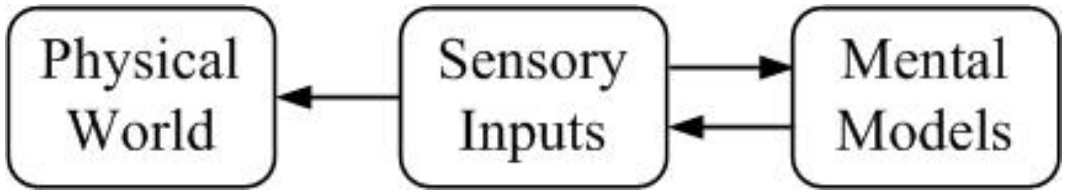
- *REALISM:*
  - objective physical reality perceived via senses
  - mental constructs reflect objective reality
- *IDEALISM:*
  - primary role belongs to ideas (mental constructs)
  - physical reality is a by-product of Mind
- *INSTRUMENTALISM:*
  - the goal of science is to produce useful theories

3 Philosophical Schools

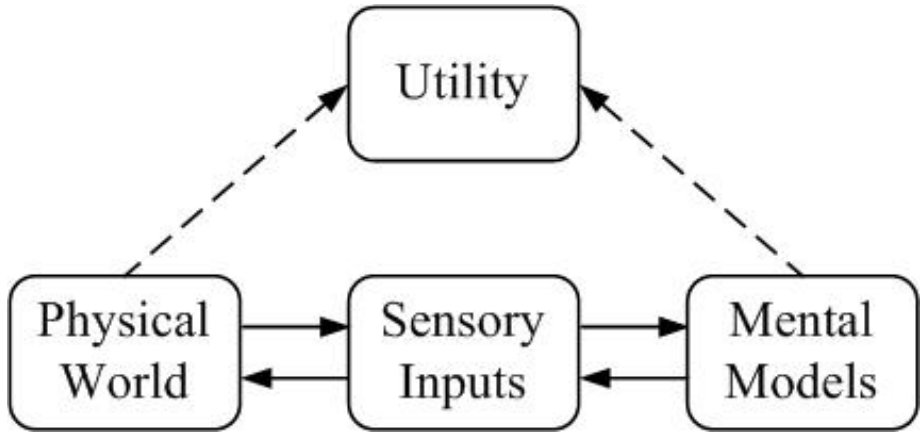
- Realism (materialism)



- Idealism



- Instrumentalism



- Realism is essential to common sense, but *can not be proven* by logic arguments
- Idealism states that only mental constructs exist:

  R. Descartes: Cogito ergo sum

  I. Kant: … if I remove the thinking subject, the whole material world must at once vanish because it is nothing but a phenomenal appearance in the sensibility of ourselves as a subject, and a manner or species of representation.
- Hegel (1770-1831): Reality and Mind are parts of a system

  Whatever exists (is real) is rational, and

  whatever is rational is real
- Instrumentalist view:

  Whatever is useful is rational and (maybe) real

Inductive Inference and Generalization

- Any scientific theory ~ **generalization** over finite number of observations (i.e., experiments used to confirm it)
- Impossible to logically justify a new theory by experimental data alone

  → need a philosophical principle known as
- **inductive inference** = generalization from repeatedly observed instances (observations) to some as yet unobserved instances (Popper, 1953).

  Similar to psychological induction (or learning by association – Pavlov's conditional reflex etc.)
- **Philosophy of Science** and **Predictive Learning** both are concerned with general strategies for obtaining good (valid) models from data

  - known as **inductive principles** in learning theory

OUTLINE

- Background
- **Inductive learning and VC-theory**
  - **Empirical Risk Minimization**
  - **Inductive learning setting**
  - **System identification vs imitation**
- Motivation for non-inductive approaches
- Non-inductive learning formulations
- Summary

Empirical Risk Minimization

- Two goals of inductive inference:

  1. Explanation (of observed data)

  2. Generalization for new data

- ERM is only concerned with 1st goal
- ERM ~ biological/ psychological induction (learning by association/ correlation)
- Example: continue given sequence

  6, 10, 14, 18, ….

Empirical Risk Minimization (ERM)

Given: training data $(\mathbf{x}_i, y_i), i = 1,2,...n$

and model estimates $\hat{y} = f(\mathbf{x}, w)$

Find a function $f(\mathbf{x}, w^*)$ that explains data best, i.e. minimizes total error $\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_k, w)) \rightarrow min$

$L(y, f(\mathbf{x}, w))$ is a non-negative loss function given a priori (reflects application requirements)

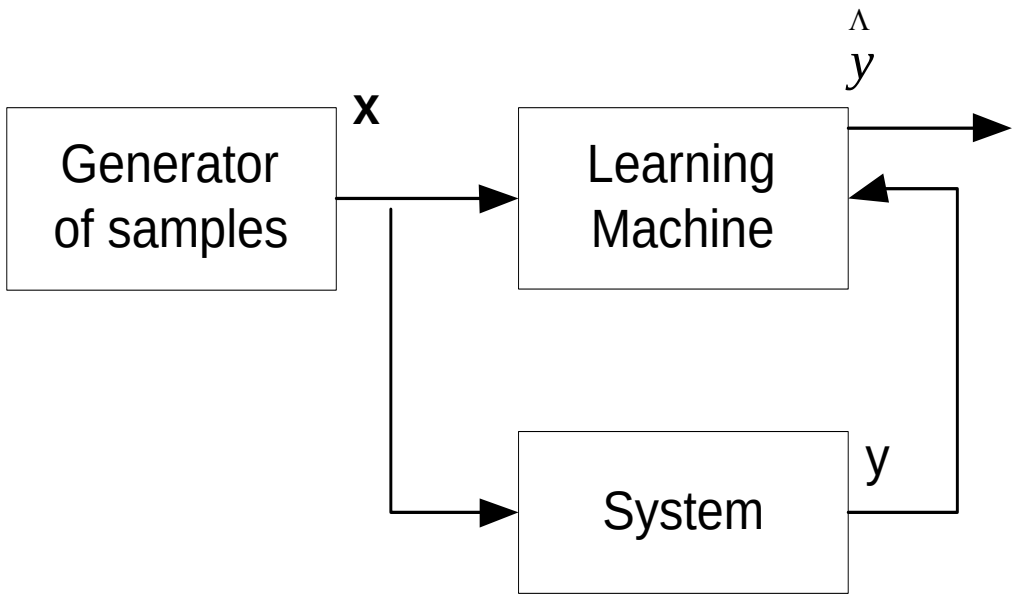
- Why/ when inductive models estimated via ERM can generalize???

Statistical Learning Theory

- Also known as VC-theory
- Theory for estimating dependencies from finite samples (predictive learning setting)
- Based on the *risk minimization* approach
- All main concepts and results developed in 1970's but remained largely unknown
- Recent popularity due to success of Support Vector Machines

Inductive Learning Setting

- The *learning machine* observes samples (**x** ,*y*), and returns an estimated response $\hat{y} = f(\mathbf{x}, w)$
- Goal of Learning: find a function (model) $f(\mathbf{x}, w^*)$
  minimizing Prediction Risk: $\int Loss(y, f(\mathbf{x},w))\, dP(\mathbf{x},y) \rightarrow min$



Two Common Learning Problems

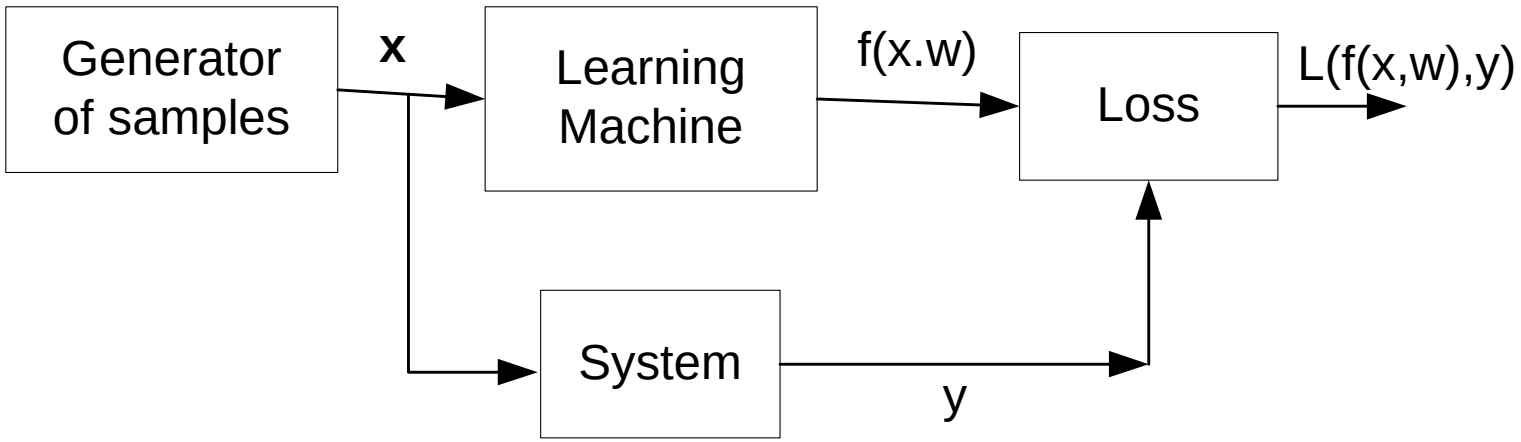
- Learning ~ estimating mapping $\mathbf{x} \rightarrow y$
  (in the sense of risk minimization)

- Binary Classification: estimating an indicator function (with 0/1 loss)
- Regression: estimating a real-valued function (with squared loss)

- Assumptions: iid, training/test, loss function

System Identification vs Imitation

- The goal of learning ~ system imitation
  rather than system identification



- Implications: curse-of-dimensionality
  philosophical

Keep-It-Direct Principle

- The goal of learning is generalization rather than estimation of true function (system identification)

$$\int Loss(y, f(\mathbf{x},w))\, dP(\mathbf{x},y) \rightarrow min$$

- **Keep-It-Direct Principle** (Vapnik, 1995)

  Do not solve an estimation problem of interest by solving a more general (harder) problem as an intermediate step
- A good predictive model reflects some properties of unknown distribution $P(\mathbf{x},y)$
- Since model estimation with finite data is ill-posed, one should never try to solve a more general problem than required by given application

  → Importance of formalizing application requirements as learning problem formulation.

Contributions of VC-theory

- The Goal of Learning (inductive learning)
  system imitation vs system identification

- Two factors responsible for generalization
  VC-dimension and empirical risk

- Keep-It-Direct Principle (Vapnik, 1995)

- Clear distinction between
  - problem setting
  - solution approach (inductive principle)
  - learning algorithm

Learning vs System Identification

- Consider regression problem
  where unknown target function $y = g(\mathbf{x}) + \delta$, $g(\mathbf{x}) = E(y/\mathbf{x})$

- Goal 1: Prediction (system imitation)

$$R(\mathbf{w}) = \int (y - f(\mathbf{x}, \mathbf{w}))^2 \, dP(\mathbf{x}, y) \rightarrow \min$$

- Goal 2: Function Approximation (system identification)

or

$$R(\mathbf{w}) = \int (f(\mathbf{x}, \mathbf{w}) - g(\mathbf{x}))^2 \, d\mathbf{x} \rightarrow \min$$

$$\| f(\mathbf{x}, \mathbf{w}) - E(y/\mathbf{x}) \| \rightarrow \min$$

- Goal 2: → curse-of-dimensionality
- Goal 1: good generalization still possible

OUTLINE

- Background
- Inductive learning and VC-theory
- Motivation for non-inductive approaches
  - Induction with sparse high-dimensional data
  - Formalizing application requirements
  - Philosophical motivation
- Non-inductive learning formulations
- Summary

Induction with sparse high-dimensional data

- Inductive learning with high-dimensional, low sample size (HDLSS) data: $n << d$
  - Gene microarray analysis
  - Medical imaging (i.e., sMRI, fMRI)
  - Object and face recognition
  - Text categorization and retrieval
  - Web search
- Sample size is much smaller than dimensionality of the input space, d ~ 10K–100K, n ~ 100's
- Inductive learning methods usually fail for such HDLSS data.

Insights provided by SVM(VC-theory)

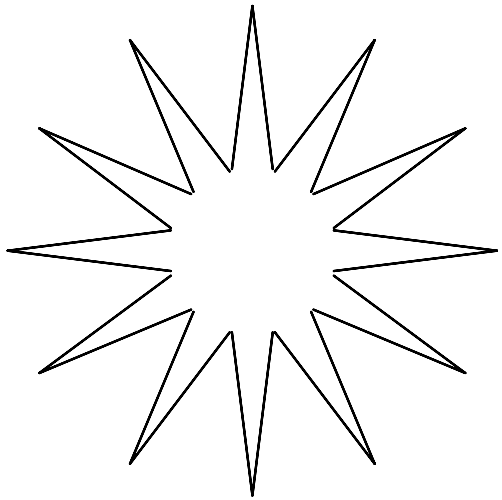
- Why linear classifiers can *generalize*?

  (1) Margin is *large* (relative to *R*)

  (2) % of SV's is *small*

  (3) ratio *d/n* is *small*
- SVM offers an effective way to control complexity (via margin + kernel selection) i.e. implementing (1) or (2) or both
- What happens when *d*>>*n* ?

Sparse High-Dimensional Data



- HDLSS data looks like a **porcupine**: the volume of a sphere inscribed in a *d*-dimensional cube gets *smaller* as the volume of *d*-cube gets *larger* !
- A point is *closer to an edge* than to another point
- Pairwise distances between points are **the same**

Classification with HDLSS data

- Data is *linearly separable* (since $d >> n$)
- Empirical studies:

  for HDLSS data, many *reasonable* methods give **similar performance** (LDA, SVM, boosting,….)
- Most data samples are SV's

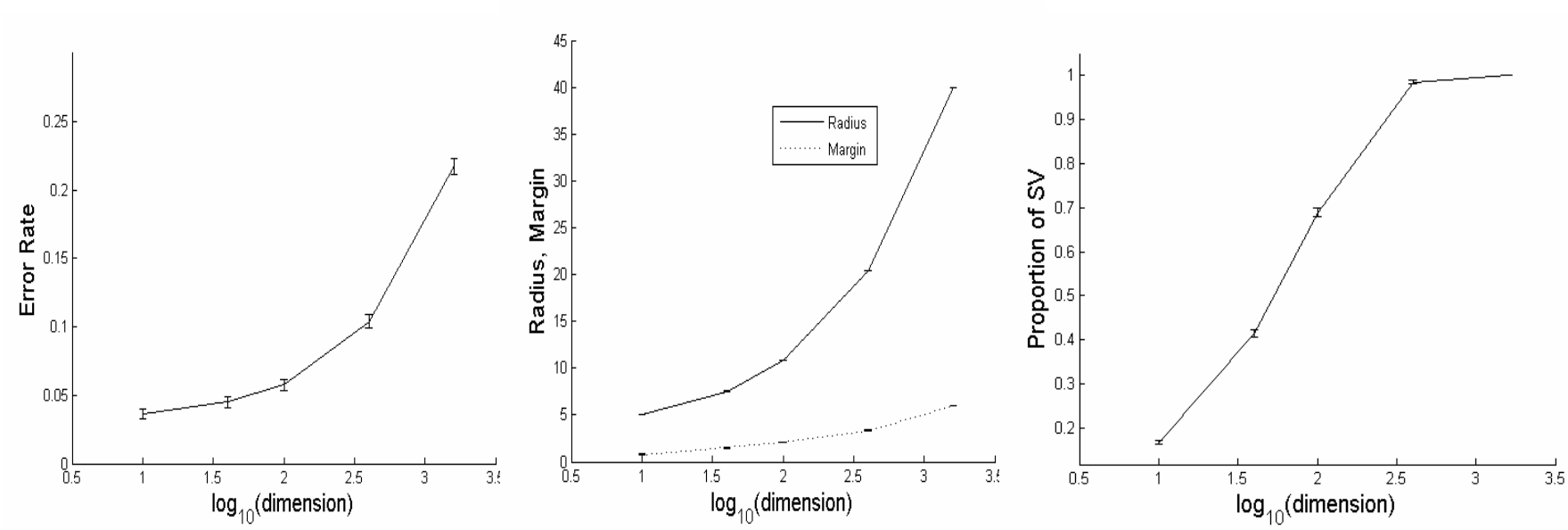
  → generalization controlled by **margin size**

  (under standard classification formulation)

Degradation of SVM performance with *d*

- Synthetic Data for binary classifier:

  30 samples per each class, with **x**-values from a spherical (*zero mean, unit variance*) Gaussian in a *d*-dimensional space, except that the first coordinate has mean +3.2 for Class +1, −3.2 for Class −1.



- Conclusion: Good generalization for HDLSS is difficult

  (using plain-vanilla SVM or any other classifier)

How to improve generalization for HDLSS?

Conventional approaches: use a priori knowledge

- *Preprocessing and feature selection* (prior to learning)
- *Model parameterization* (~ selection of good kernels)
- *Generate artificial training examples* (Virtual SV method)

  The idea is to apply the desired invariance transformations to SV's (Schoelkopf and Smola, 2001):

  (1) Apply SVM classifier to training data

  (2) Generate Virtual SVs by applying invariance transformations to support vectors obtained in (1)

  (3) Train another SV classifier using Virtual SV's.

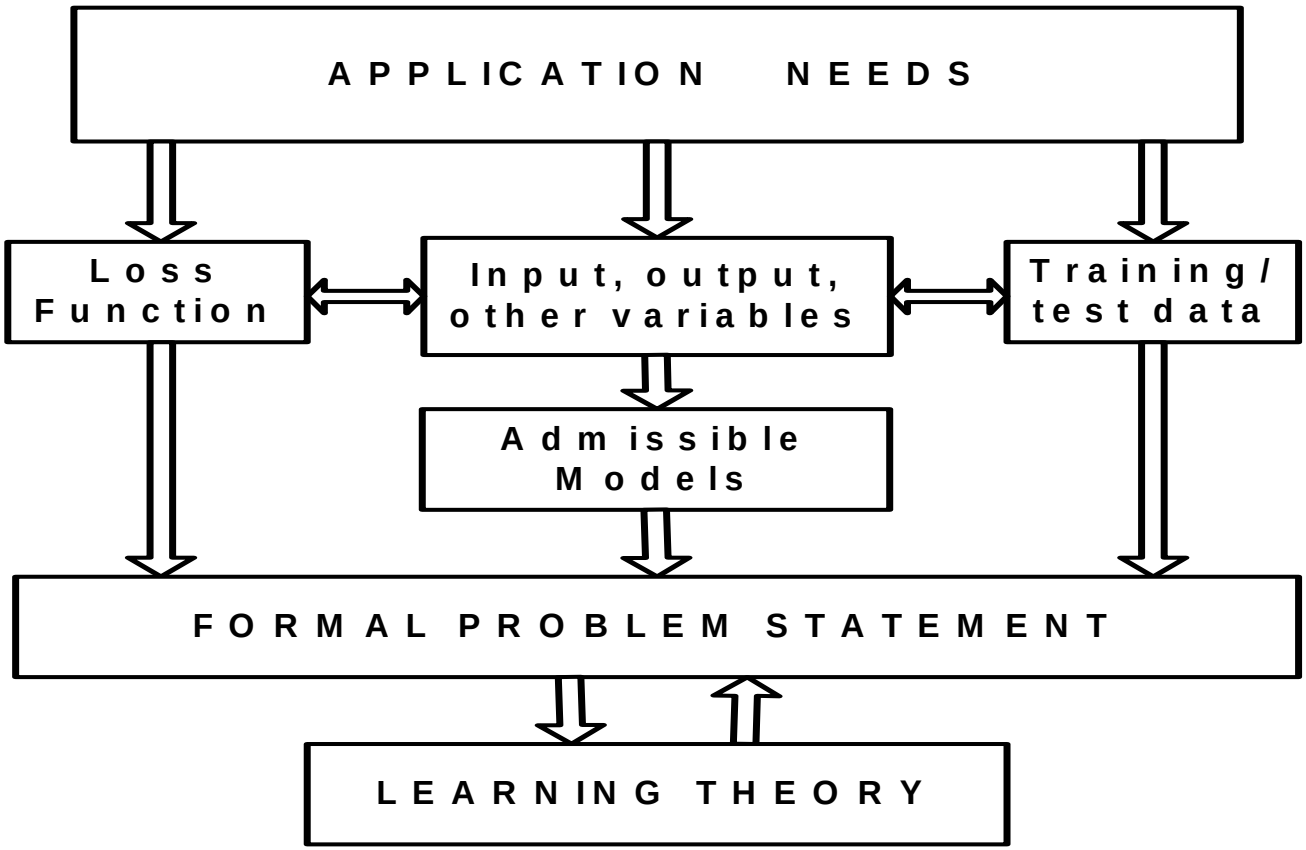
Non-inductive learning formulations

- Seek new generic formulations *(not methods!)* that better reflect application requirements

Formalizing Application Requirements

- Classical statistics: parametric model is given (by experts)
- Modern applications: complex iterative process

  → Non-inductive formulation may be better than inductive

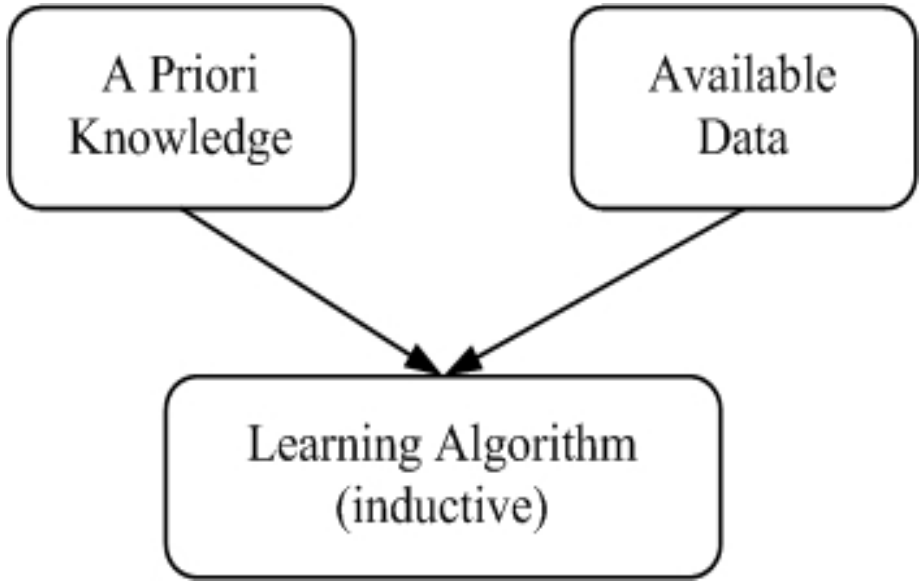


Philosophic Motivation

- Philosophical view 1 (Realism):
  Learning ~ search for the truth (estimation of true dependency from available data)

→ System identification

~ Inductive Learning

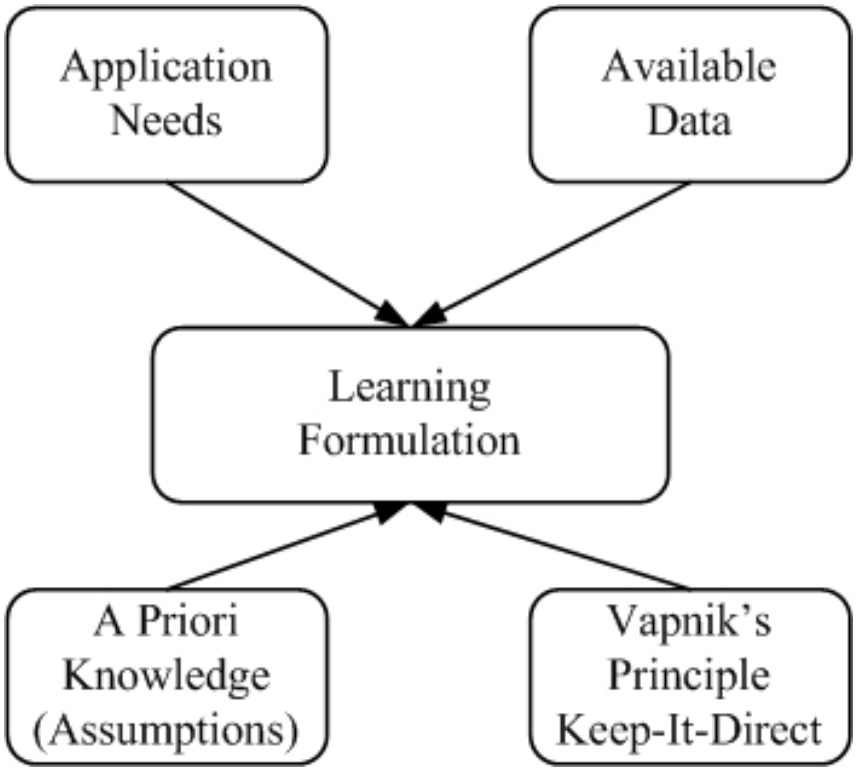


where a priori knowledge is about the true model

Philosophic Motivation (cont'd)

- Philosophical view (Instrumentalism):
  Learning ~ search for the instrumental knowledge
  (estimation of useful dependency from available data)

→ VC-theoretical approach ~ focus on learning formulation



VC-theoretical approach

- Focus on the learning setting (formulation), ***not*** on the learning method
- Learning formulation depends on:

  (1) available data

  (2) application needs

  (3) a priori knowledge (assumptions)
- Factors (1)-(3) combined using Vapnik's Keep-It-Direct (KID) Principle yield learning formulation

Contrast these two approaches

- **Conventional (statistics, data mining):**

  a priori knowledge typically reflects properties of a true (good) model, i.e.

  a priori knowledge ~ parameterization $f(\mathbf{x}, w)$

- Why a priori knowledge is about the true model?
- **VC-theoretic approach:**

  a priori knowledge ~ how to use/ incorporate available data into the problem formulation

  often a priori knowledge ~ available data samples of different type → *new learning settings*

OUTLINE

- Background
- Inductive learning and VC-theory
- Motivation for non-inductive approaches
- Non-inductive learning formulations

  Transduction

  Inference through contradictions

  Learning with structured data

  Multi-task learning
- Summary

Modifications of inductive setting

- **Inductive learning** assumes

  Finite training set $(\mathbf{x}_i, y_i)$

  Predictive model derived using only training data

  Prediction for all possible test inputs

- Possible modifications

  1. Predict only for given test points → transduction

  2. A priori knowledge in the form of additional 'typical' samples → learning through contradiction

  3. Additional (group) info about training data → Learning with structured data

  4. Additional (group) info about training + test data → Multi-task learning

Examples of non-inductive settings

- **Application domain:** hand-written digit recognition
- **Standard inductive setting**
- **Transduction:** labeled training + unlabeled data
- **Learning through contradictions:**
  labeled training data ~ examples of digits 5 and 8
  unlabeled examples (Universum) ~ all other (eight) digits
- **Learning with structured data:**
  Training data ~ *t* groups (i.e., from *t* different persons)
  Test data ~ group label not known
- **Multi-task learning:**
  Training data ~ *t* groups (from different persons)
  Test data ~ *t* groups (group label is known)

Transduction (Vapnik, 1982, 1995)

- How to incorporate unlabeled test data into the learning process? Assume binary classification

- Estimating function at given points

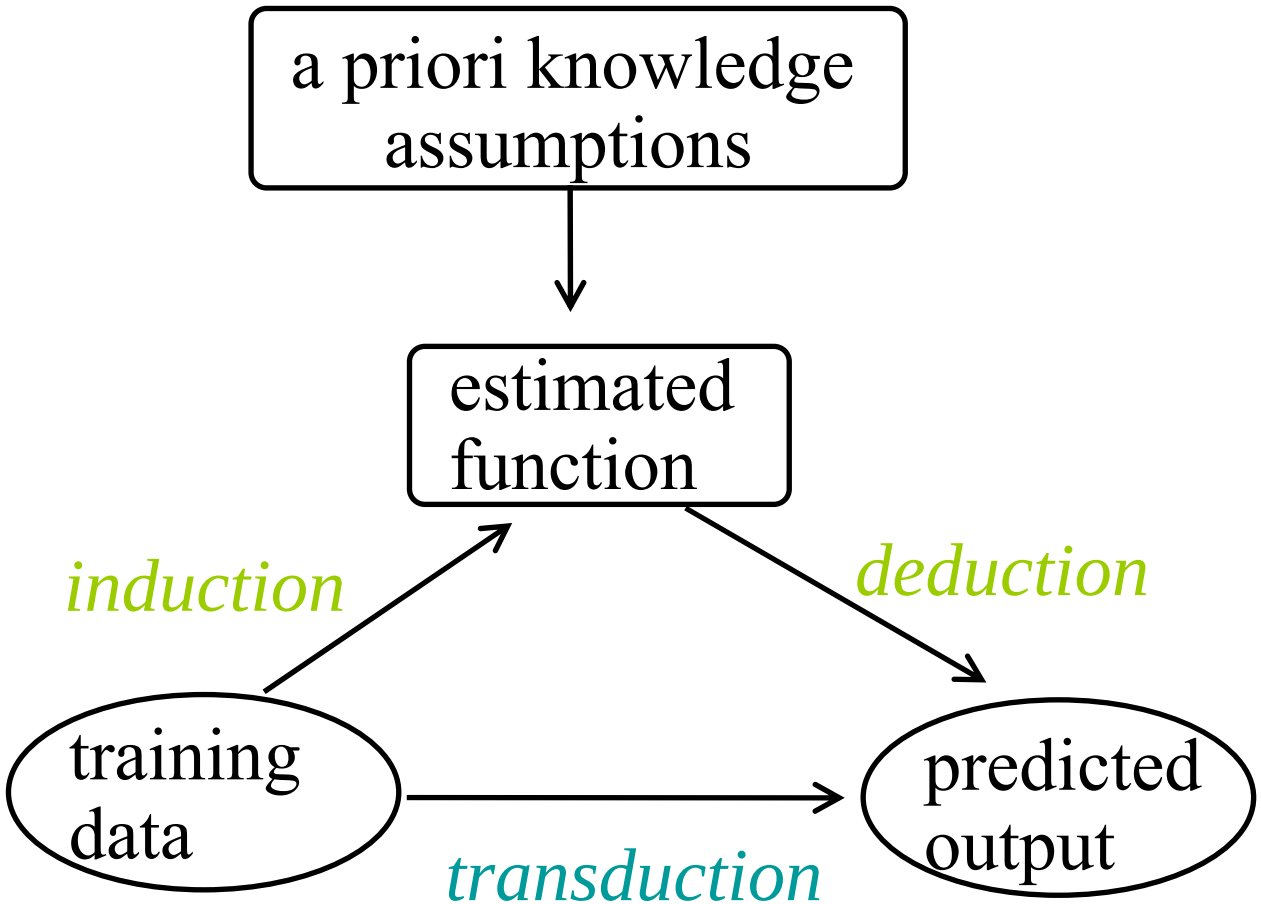
  Given: labeled training data $(\mathbf{x}_i, y_i) \quad i = 1,..., n$
  and unlabeled test points $(\mathbf{x}_j^*) \quad j = 1,..., m$

  Estimate: class labels $\mathbf{y}^* = (y_1^*,.... y_m^*)$ at these test points

Goal of learning: minimization of risk on the test set:

$$R(\mathbf{y}^*) = \frac{1}{m}\sum_{j=1}^{m} \int_y L(y, y_j^*) dP(y / \mathbf{x}_j^*) \text{ where } \mathbf{y}^* = \left(f(\mathbf{x}_1^*, \omega),.... f(\mathbf{x}_m^*, \omega)\right)$$

Transduction vs Induction



Transduction based on size of margin

- Binary classification, linear parameterization, joint set of (training + working) samples
- Equivalence classes (on a joint set) $F_1, F_2, ..., F_N$ –see next page

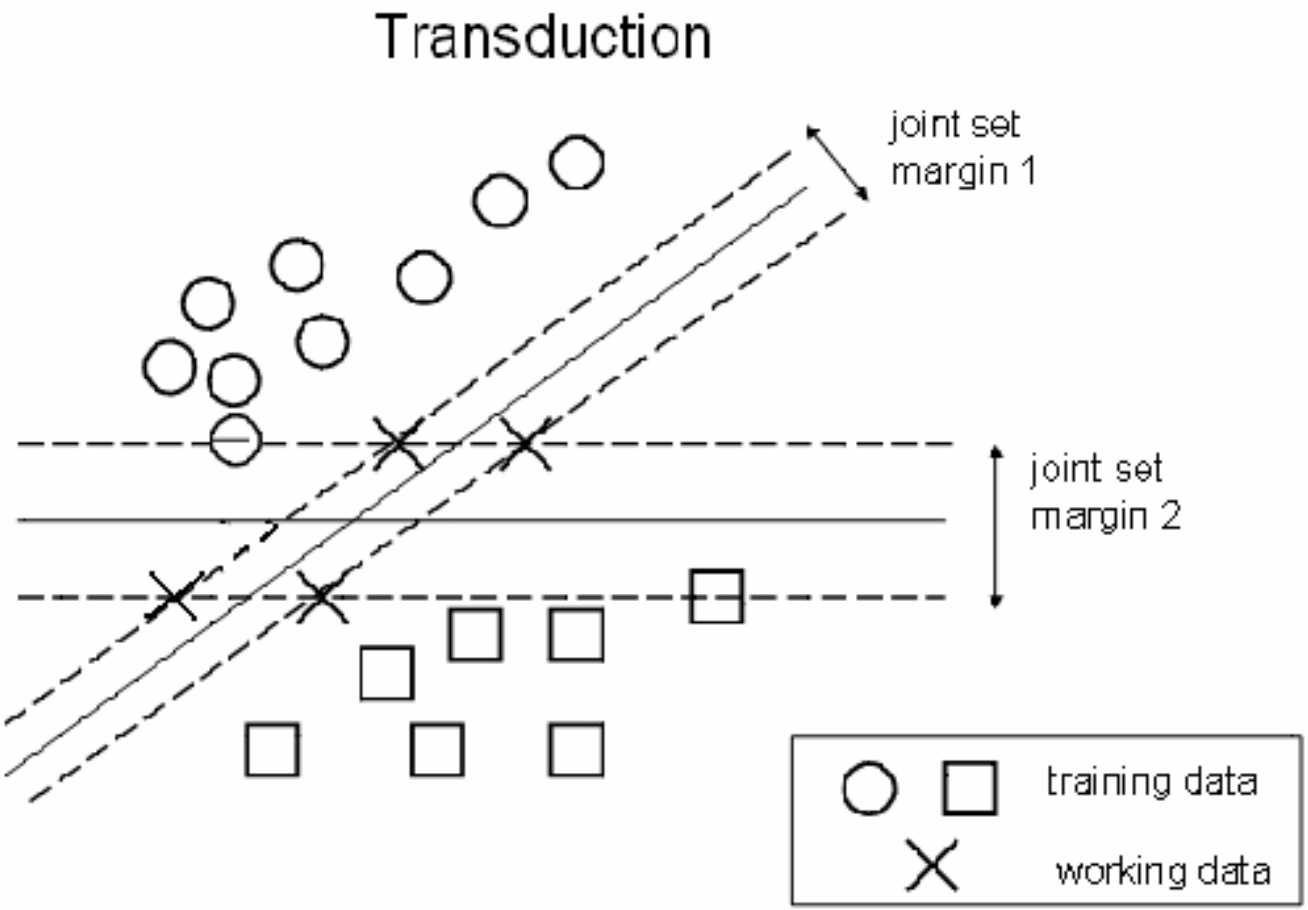
- Goal of learning

  (1) explain well available data (~ joint set)

  (2) achieve max falsifiability (~ margin)

~ Classify test (working) samples by the equivalence class that explains well available data and has large margin

- → Optimization formulation (see later)



- Example of two equivalence classes
- The size of an equivalence class is indexed by the largest value of its margin on a joint set

Optimization formulation for SVM transduction

- Given: joint set of (training + working) samples
- Denote slack variables ξ_i for training, ξ_j^* for working
- Minimize $R(\mathbf{w},b) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}) + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{m}\xi_j^*$

  subject to $\begin{cases} y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b] \geq 1-\xi_i \\ y_j^*[(\mathbf{w}\cdot\mathbf{x}_i)+b] \geq 1-\xi_j^* \\ \xi_i\ ,\xi_j^* \geq 0, i=1,...,n,\ j=1,...,m \end{cases}$

  where $y_j^* = sign(\mathbf{w}\cdot\mathbf{x}_j + b),\ j=1,...,m$

  → Solution (~ decision boundary) $D(\mathbf{x}) = (\mathbf{w}^*\cdot\mathbf{x}) + b^*$
- Unbalanced situation (small training/ large test)

  → all unlabeled samples assigned to one class
- Additional constraint: $\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{m}\sum_{j=1}^{m}[(\mathbf{w}\cdot\mathbf{x}_i)+b]$

Optimization formulation (cont'd)

- Hyperparameters $C \; and \; C^*$ control the trade-off between explanation and falsifiability
- Soft-margin inductive SVM is a special case of soft-margin transduction with zero slacks $\xi_j^* = 0$
- Dual + kernel version of SVM transduction
- Transductive SVM optimization is not convex

  (~ non-convexity of the loss for unlabeled data) – ***elaborate/explain***

  → different opt. heuristics ~ different solutions
- Exact solution (via exhaustive search) possible for small number of test samples (*m*) – but this solution is *NOT* very useful (~ inductive SVM).

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

Example application

- Prediction of molecular bioactivity for drug discovery
- Training data~1,909; test~634 samples
- Input space ~ 139,351-dimensional
- Prediction accuracy:

SVM induction ~74.5%; transduction ~ 82.3%

Ref: J. Weston et al, KDD cup 2001 data analysis: prediction of molecular bioactivity for drug design – binding to thrombin, *Bioinformatics 2003*

Inference through contradiction (Vapnik 2006)

- *Motivation:* what is a priori knowledge?

  - info about the **space of admissible models**

  - info about **admissible data samples**

- *Labeled* training samples + *unlabeled* samples from the **Universum**
- Universum samples encode info about the region of input space *(where application data lives):*

  - Usually from a different distribution than training/test data

- Examples of the Universum data
- Large improvement for small sample size *n*

Main Idea

- Handwritten digit recognition: digit 5 vs 8

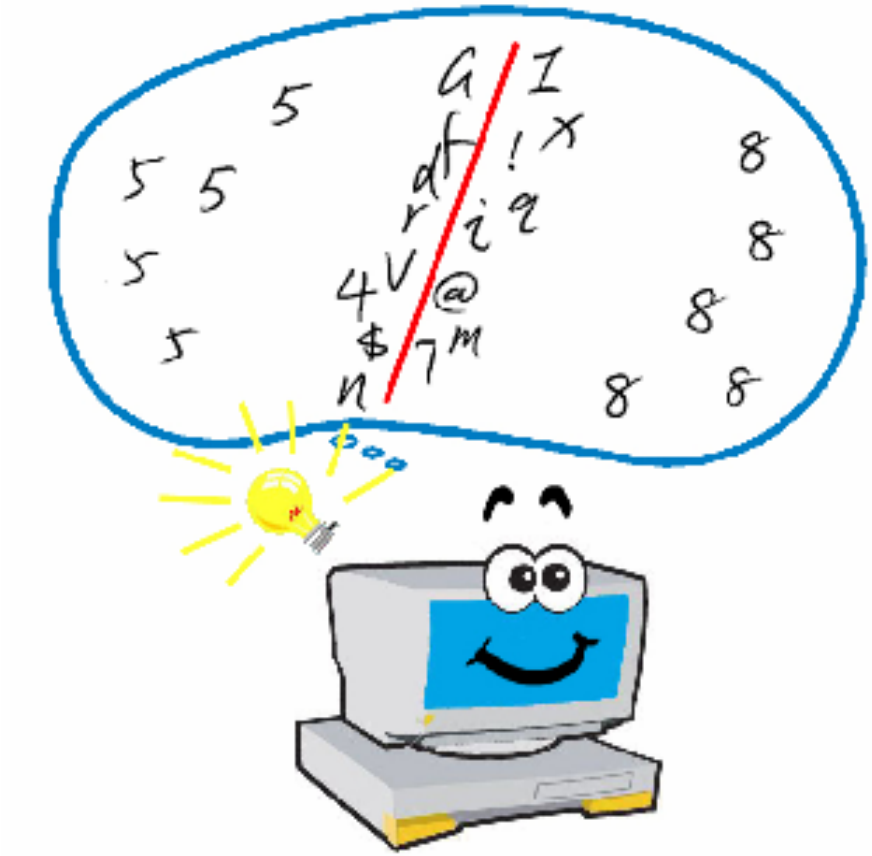


Fig. courtesy of J. Weston (NEC Labs)

Learning with the Universum

- Inductive setting for binary classification

  Given: labeled training data $(\mathbf{x}_i, y_i) \quad i = 1,...,n$

  and unlabeled Universum samples $(\mathbf{x}_j^*) \quad j = 1,...,m$

  Goal of learning: minimization of prediction risk (as in standard inductive setting)

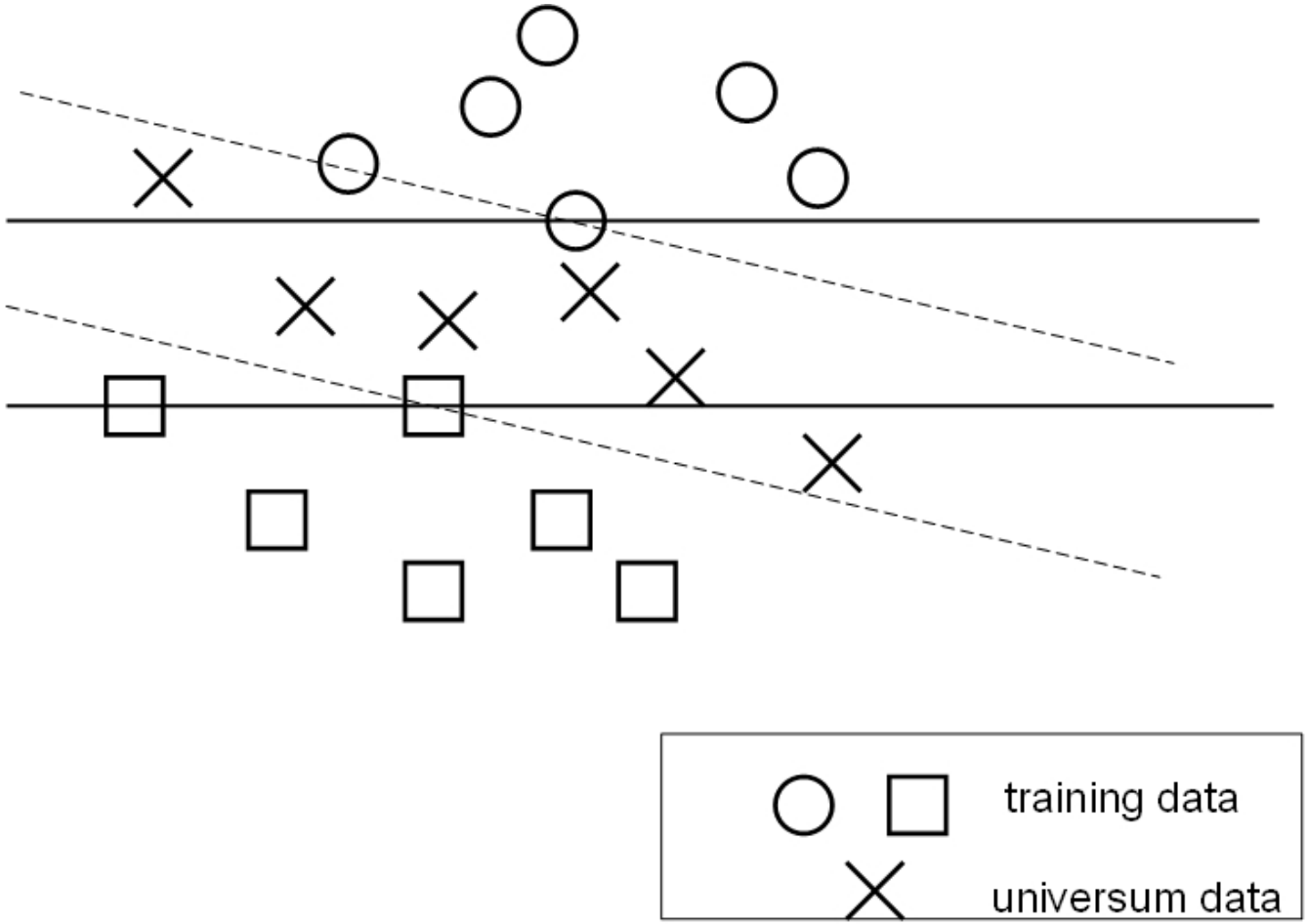
- Balance between two goals:

  - explain labeled training data using large-margin hyperplane

  - achieve maximum falsifiability ~ max # contradictions on the Universum

~ For separating hyperplanes, a set of equivalence classes is ordered by the number of contradictions on the Universum

Inference through contradictions



ε

SVM inference through contradictions

- Given: labeled training + unlabeled Universum samples
- Denote slack variables ξ_i for training, ξ_j^* for Universum
- Minimize $R(\mathbf{w},b)=\frac{1}{2}(\mathbf{w}\cdot\mathbf{w})+C\sum_{i=1}^{n}\xi_i+C^*\sum_{j=1}^{m}\xi_j^*$ where $C,C^*\geq 0$

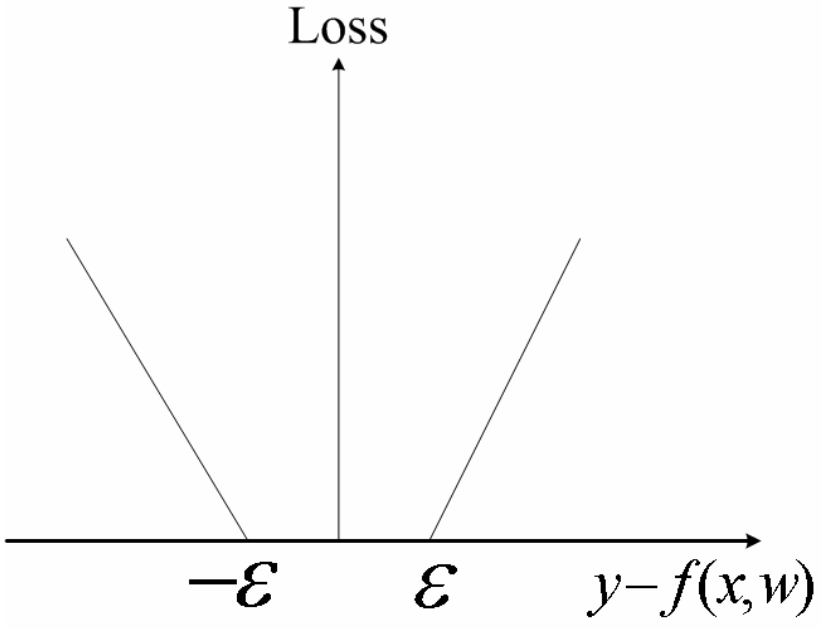
subject to $y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b]\geq 1-\xi_i \quad \xi_i,\geq 0, i=1,\ldots,n$ for labeled data

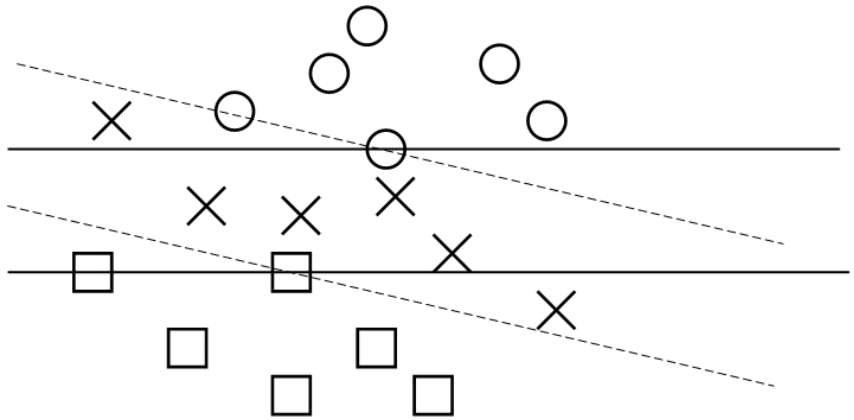
$$\left|(\mathbf{w}\cdot\mathbf{x}_i)+b\right|\leq\varepsilon+\xi_i^* \quad \xi_j^*\geq 0,\ j=1,\ldots,m$$ for the Universum

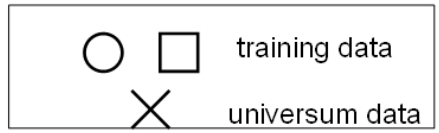
where the Universum samples use ε-insensitive loss

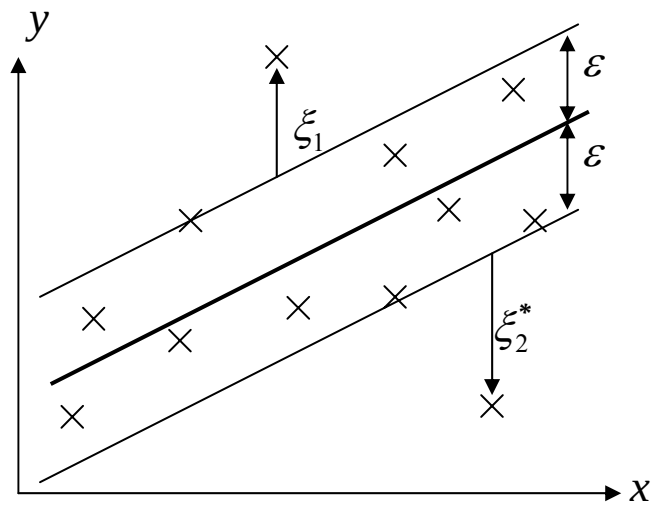
- Convex optimization
- Hyper-parameters $C,C^*\geq 0$ control the trade-off btwn minimization of errors and maximizing the # contradictions
- When C^*=0, → standard soft-margin SVM

ε-insensitive loss for Universum samples









Application Study (Vapnik, 2006)

- Binary classification of handwritten digits 5 and 8
- For this binary classification problem, the following Universum sets had been used:

  U1: randomly selected digits (0,1,2,3,4,6,7,9)

  U2: randomly mixing pixels from images 5 and 8

  U3: average of randomly selected examples of 5 and 8

  Training set size tried: 250, 500, … 3,000 samples

  Universum set size: 5,000 samples
- Prediction error: improved over standard SVM, i.e. for 500 training samples: 1.4% vs 2% (SVM)

Another application example

- **Gender classification** of human faces (male/ female)
- Data: 260 pictures of 52 individuals (20 females and 32 males, 5 pictures for each individual) from Univ. of Essex
- Data representation and pre-processing: image size 46x59 – converted into gray-scale image, following standard image processing (histogram equalization)

- Training data: 5 female and 8 male photos
- Test data: remaining 39 photos (of other people)
- Universum set generation:

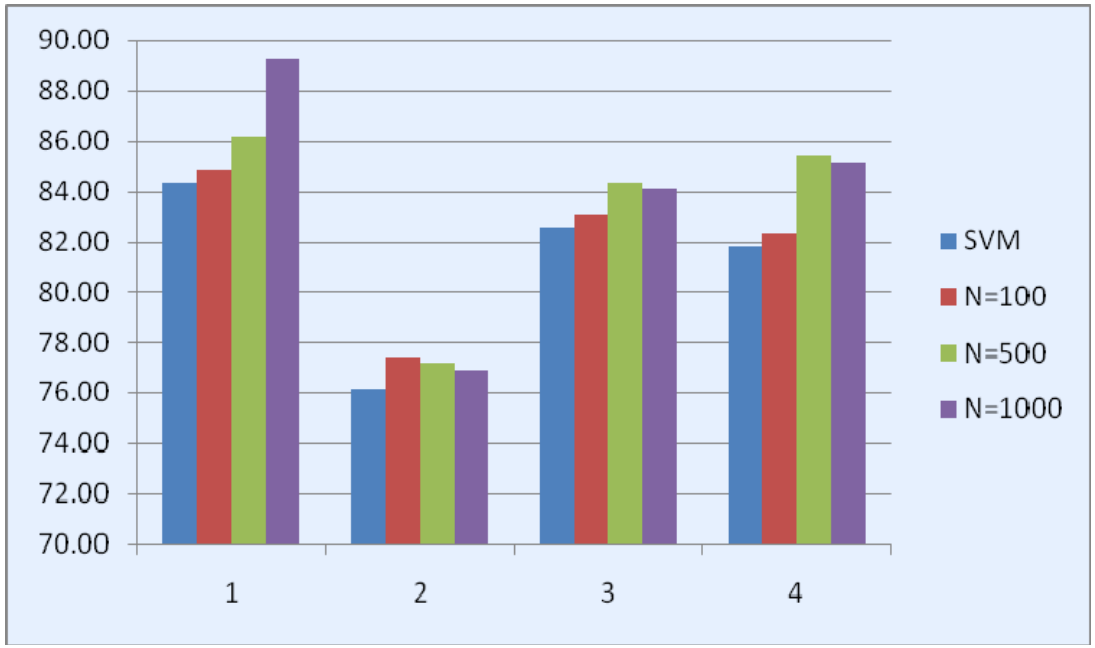
  U1 Average: randomly get one male and one female fro the training set, and compute their average vector

  U2 Empirical distribution: estimate pixel-wise distribution of training data. Generate a new picture from this distribution. randomly mixing pixels from images 5 and 8
- Experimental procedure: randomly select 13 training samples (and 39 test samples). Estimate and compare inductive SVM classifier with SVM classifier using *N* Universum samples (where *N*=100, 500, 1000).

  - Report results for 4 partitions (of training and test data)

Results of gender classification (X. Bai, 2006)

- Classification accuracy: improves vs standard SVM by ~ 2% with U1, and ~ 1% with U2 Universum



- Universum generated by averaging gives better results for this problem, when number of Universum samples *N* = 500 or 1,000 (see above)

Discussion

- Maximum margin principle (SVM) performs complexity control independent of the data distribution
- Inference by contradiction principle controls complexity depending on the properties of data distribution
- In both instances, complexity control is achieved independent of dimensionality.

Discussion (cont'd)

- Technical aspects: why the Universum data improves generalization when *d*>>*n*?

  -SVM is solved in a *low-dimensional subspace implicitly defined* by the Universum data
- How to specify the Universum data?

  - no formal procedure exists
- Philosophical aspects: relation to human learning (cultural Universum)? New type of inference? Impact on psychology?

Learning with Structured Data(Vapnik, 2006)

- **Application: Handwritten digit recognition**

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

  Goal 1: find a classifier that will generalize well for future samples generated by these persons ~ Learning with Structured Data (LWSD)

  Goal 2: find *t* classifiers with generalization (for each person) ~ Multi-Task Learning (MTL)

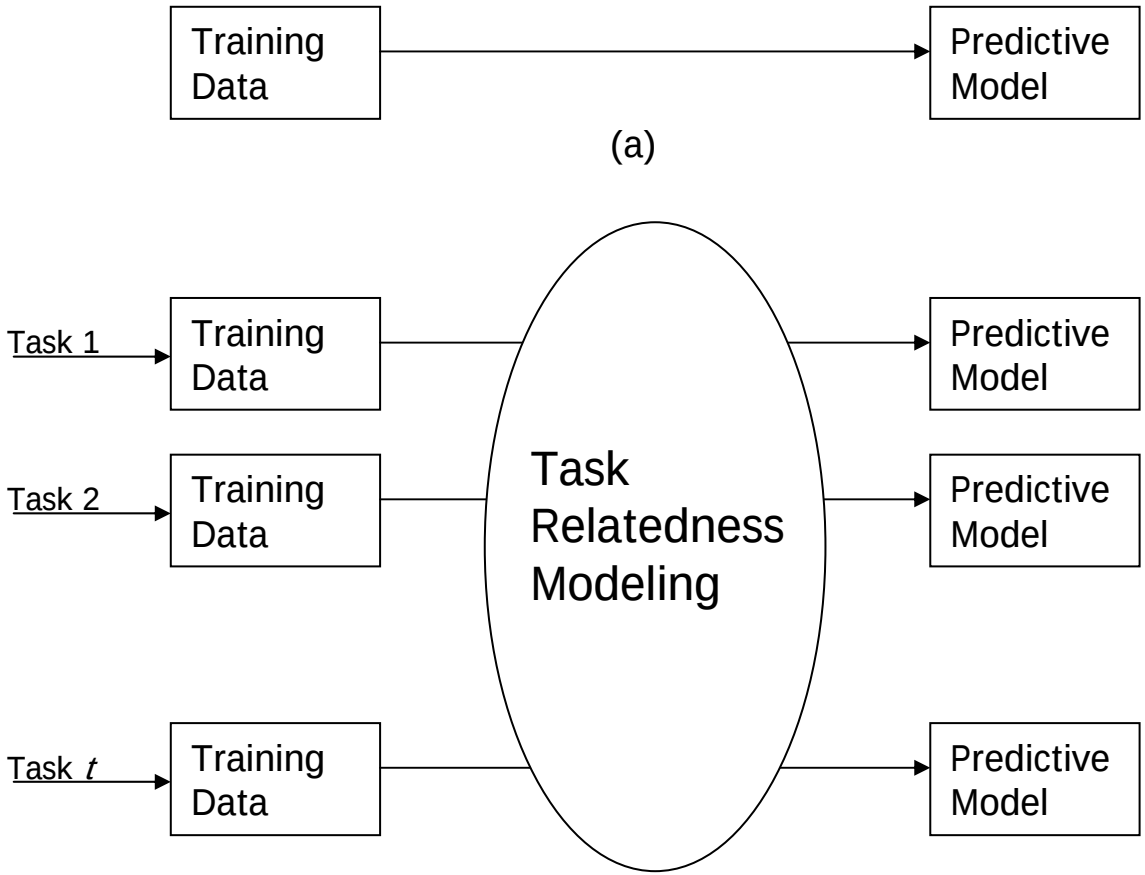
- **Application: Medical diagnosis**

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

  Goal 1: estimate a classifier to predict/diagnose a disease using training data from *t* groups of patients ~ LWSD

  Goal 2: find *t* classifiers specialized for each group of patients ~ MTL

Multi-task Learning



(a) Single task learning

(b) Multi-task learning (MTL)

Problem setting for LWSD

- Assume binary classification problem
- Training data ~ a union of *t* related groups

  Each group *r* has n_r *i.i.d.* samples

  $$(\mathbf{x}_i^r, y_i^r) \qquad i = 1,2,...,n_r \qquad r = 1,2,...,t$$

  generated from a joint distribution $P_r(\mathbf{x}, y)$

- Training and test data are *i.i.d.* samples from the same distribution $P(\mathbf{x}, y) = \cup_{r=1,..t} P_r$
- Goal of learning: estimate a single model $f(\mathbf{x}, w^*)$ minimizing prediction risk (similar to inductive learning)

$$R(w) = \int L\big(y, f(\mathbf{x}, w)\big) dP(\mathbf{x}, y)$$

Problem setting for MTL

- Assume binary classification problem
- Training data ~ a union of *t* related groups(tasks)

  Each group *r* has n_r *i.i.d.* samples

  $$(\mathbf{x}_i^r, y_i^r) \qquad i = 1,2,...,n_r \qquad r = 1,2,...,t$$

  generated from a joint distribution $P_r(\mathbf{x}, y)$

- Test data: task label for test samples is known

- Goal of learning: estimate *t* models $\{f_1, f_2, ..., f_t\}$ such that the sum of expected losses for all tasks is minimized:

  $$R(w) = \sum_{r=1}^{t} (\int L(y, f_r(\mathbf{x}, w)) dP_r(\mathbf{x}, y)$$

SVM+ technology (Vapnik, 2006)

- Map the input vectors simultaneously into:

  - Decision space (standard SVM classifier)

  - Correcting space (where correcting functions model slack variables for different groups)

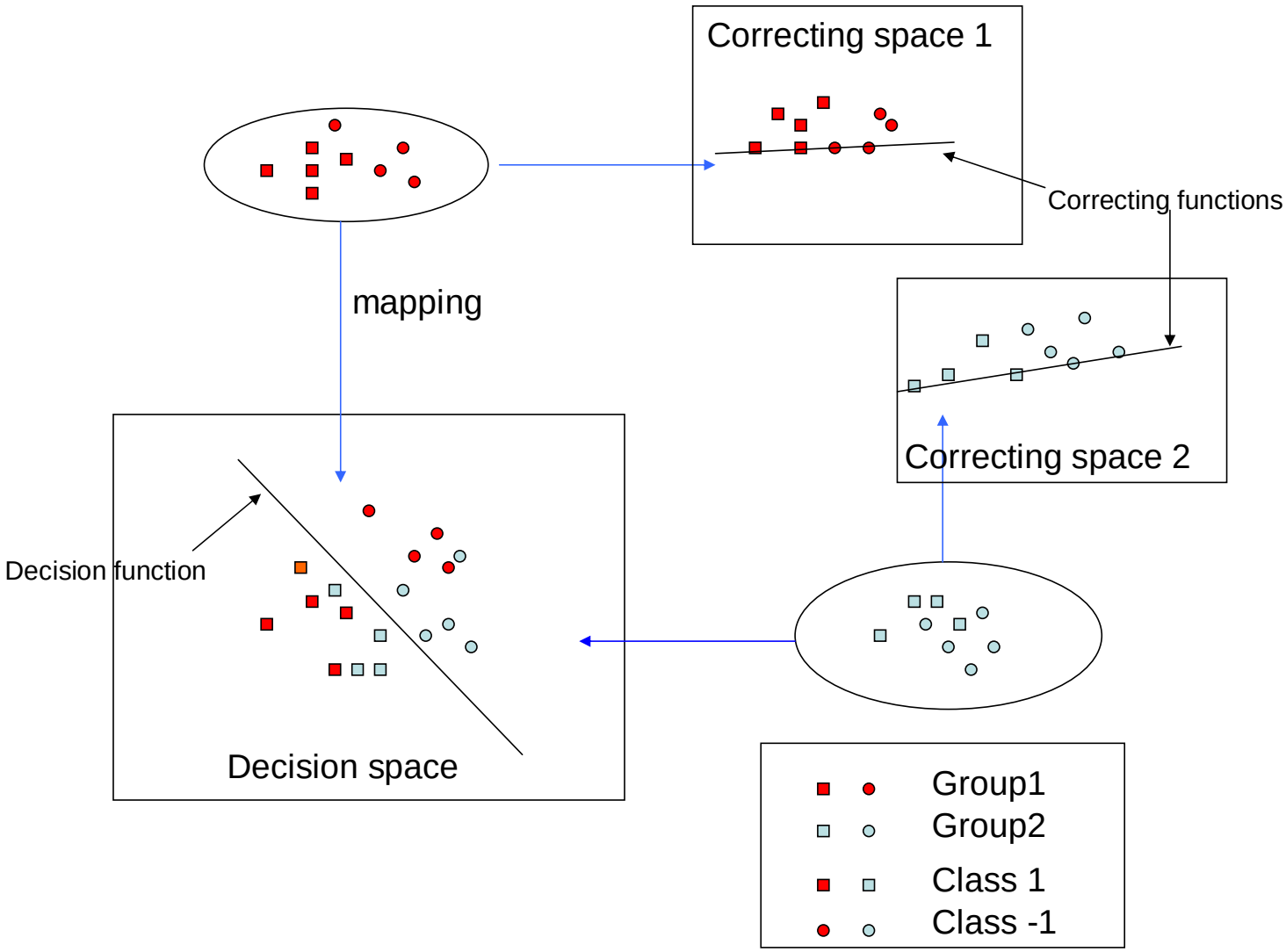
- Decision space/function ~ the same for all groups
- Correcting functions ~ different for each group (but correcting space may be the same)
- SVM+ optimization formulation incorporates:

  - the capacity of decision function $(\mathbf{w}, \mathbf{w})$

  - capacity of correcting functions $(\mathbf{w}_r, \mathbf{w}_r)$ for group *r*

  - relative importance (weight) γ of these two capacities

SVM+ technology



Application Study (Liang and Cherkassky, 2007)

- SVM+ technology is new, and is rather complex. Single empirical comparison study: *fMRI* data
- fMRI data analysis problem (CMU data set)

  - six subjects presented with {picture *or* sentence}

  - fMRI data is recorded over 16 time intervals

  - need to learn binary classifier *fMRI image* → class

- Data preprocessing (Wang et al, 2004)

  - *extract 7 input features* (Regions of Interest) from high–dim. fMRI image → input vector has 16x7=112 components

- Comparison between

  standard SVM (data from all 6 subjects is pooled together)

  SVM+ method (6 groups where each group has 80 samples)

fMRI Application Study (cont'd)

- Experimental protocol: (randomly) split the data
  - 60% ~ training
  - 20% ~ validation (for tuning parameters)
  - 20% ~ test (for estimating prediction error)
- Details of methods used:
  - linear SVM classifier (single parameter *C*)
  - $SVM_\gamma +$ classifier (3 parameters: linear kernel for decision space, RBF kernel for correcting space, and parameter γ)
- Comparison results (over 10 trials):
  - standard SVM ~ ave test error: 22.2 % (st. dev. 3.8%)
  - $SVM_\gamma +$ ~ ave test error: 20.2% (st.dev. 3.1%)

OUTLINE

- Background
- Inductive learning and VC-theory
- Motivation for non-inductive approaches
- Non-inductive learning formulations
- Summary

  Advantages/limitations of non-inductive settings

  Vapnik's Imperative

Advantages+limitations of noninductive settings

- Advantages
  - makes common sense
  - follows methodological framework (VC-theory)
  - yields better generalization (but not always)
  - new directions for research
- Limitations
  - need to formalize application requirements
  - generally more complex learning formulations
  - more difficult model selection (# parameters)
  - few known empirical comparisons (to date)

Vapnik's Imperative

- **Vapnik's Imperative**:

  asking the right question (with finite data)
  → most direct learning problem formulation
- **~ Controlling the goals of inference**, in order to produce a better-posed problem
- **Three types of such restrictions** (Vapnik, 2006)

  - regularization (constraining function smoothness)

  - SRM puts constraints on VC-dim of approx fcts

  - choice of inference models
- **Philosophical interpretation** (of restrictions):

  Find a model that explains well available data and has large falsifiability

References

- 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
- Chapelle, O., Schölkopf, B., and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, 2006
- Popper, K., *The Logic of Scientific Discovery* (2nd ed.), New York: Harper Torch Books,1968
- Cherkassky, V. and Y. Ma, Data complexity, margin-based learning and Popper's philosophy of inductive learning, in *Data Complexity in Pattern Recognition*, M. Basu and T. Ho , Eds, Springer, 2006
- Weston, J., Collobert, R., Sinz, F., Bottou, L. and V. Vapnik, Inference with the Universum, *Proc. ICML* 2006
- Ando, R. and Zhang, T., A Framework for learning predictive structures from multiple tasks and unlabeled data, J. of Machine Learning Research, 2005
- Evgeniou, T. and Pontil, M., Regularized multi-task learning, *Proc of 10-th Conf. on Knowledge Discovery and Data Mining*, 2004
- Wang, X., Hutchingson, R. and Mitchell, T., Training fMRI classifier to discriminate cognitive states across multiple subjects, *Proc. NIPS*, 2004
- Liang, L. and Cherkassky, V., Learning using structured data: application to fMRI data analysis, *Proc. IJCNN*, 2007