Chapter 3

NONINDUCTIVE METHODS OF INFERENCE: DIRECT INFERENCE INSTEAD OF GENERALIZATION (2000–· · ·)

3.1 INDUCTIVE AND TRANSDUCTIVE INFERENCE

Chapter 10 of *EDBED* distinguishes between two different problems of estimation: estimation of the function and estimation of the values of the function at given points of interest.

(1) *Estimation of the function.* Given training data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell), \tag{3.1}$$

find in the set of admissible functions $f(x, \alpha), \alpha \in \Lambda$ the one which guarantees that its expected loss is close to the smallest loss.

(2) *Estimation of the value of the function at the points of interest.* Given a set of training data (3.1) and a sequence of k test vectors

$$x_{\ell+1}, \ldots, x_{\ell+k}, \tag{3.2}$$

find among an admissible set of binary vectors

$$\{Y_* = (y^*_{\ell+1}, \ldots, y^*_{\ell+k})\}$$

the one that classifies the test vectors with the smallest number of errors. Here we consider

$$x_1, \ldots, x_{\ell+k} \tag{3.3}$$

to be random i.i.d. vectors drawn according to the same (unknown) distribution $P(x)$. The classifications y of the vectors x are defined by some (unknown) conditional probability function $P(y|x)$.

This setting is quite general. In the book we considered a particular setting where the set of admissible vectors is defined by the admissible set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$. In other words, every admissible vector of classification Y_* is defined as follows

$$Y_* = (f(x_1, \alpha_*), \ldots, f(x_k, \alpha_*)) .$$

In the mid-1990s (after understanding the relationship between the pattern recognition problem and the philosophy of inference), I changed the technical terminology [139]. That is, I called the problem of function estimation that requires one to find a function given particular data *inductive inference*. I called the problem of estimating the values of the function at particular points of interest given the observations *transductive inference*.

These two different ideas of inference reflect two different philosophies, which we will discuss next.

3.1.1 Transductive Inference and the Symmetrization Lemma

The mechanism that provides an advantage to the transductive mode of inference over the inductive mode was clear from the very beginning of statistical learning theory. It can be seen in the proof of the very basic theorems on uniform convergence. This proof is based on the following inequality which is the content of the so-called symmetrization lemma (see Basic lemma in *EDBED* Chapter 6, Section A3):

$$P\left\{\sup_{\alpha} |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \leq 2P\left\{\sup_{\alpha} \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| \geq \frac{\varepsilon}{2}\right\}, \quad (3.4)$$

where

$$R^{(1)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)| \quad (3.5)$$

and

$$R^{(2)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} |y_i - f(x_i, \alpha)| \quad (3.6)$$

are the empirical risks constructed using two different samples.

The bound for uniform convergence was obtained as an upper bound of the right-hand side of (3.4).

Therefore the symmetrization lemma implies that to obtain a bound for inductive inference we first obtain a bound for transductive inference (for the right-hand side of (3.4)) and then obtain an upper bound for it.

It should be noted that since the bound on uniform convergence was introduced in 1968, many efforts were made to improve it. However, all attempts maintain some form of the symmetrization lemma. That is, in the proofs of the bounds for uniform convergence the first (and most difficult) step was to obtain the bound for transductive inference. The trivial upper bound of this bound gives the desired result for inductive inference.

This means that transductive inference is a fundamental step in machine learning.

3.1.2 Structural Risk Minimization for Transductive Inference

The proof of the symmetrization lemma is based on the following observation: The following two models are equivalent (see Chapter 10, Section 1 of EDBED):

(a) one chooses two i.i.d. sets[1]

$$x_1, \ldots, x_\ell, \quad \text{and} \quad x_{\ell+1}, \ldots, x_{2\ell};$$

(b) one chooses an i.i.d. set of size 2ℓ and then randomly splits it into two subsets of size ℓ.

Using model (b) one can rewrite the right-hand side of (3.4) as follows

$$P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\} =$$

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \; |\{x_1, \ldots, x_{2\ell}\}\right\}. \tag{3.7}$$

To obtain the bound we first bound the conditional probability

$$P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \; |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda(x_1, \ldots, x_{2\ell}) \exp\left\{-\varepsilon^2 \ell\right\} \tag{3.8}$$

where $\Delta^\Lambda(x_1, ..., x_{2\ell})$ is the number of equivalence classes on the set (3.3). The probability is obtained with respect to the random split data into two parts (training and testing). Then we take the expectation over working sets of size 2ℓ. As a result, we obtain

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2} |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda_P(2\ell) \exp\left\{-\varepsilon^2 \ell\right\}. \tag{3.9}$$

Note that for the transductive model of inference we do not even need to take the expectation over sets of size 2ℓ. We can just use the bounds (3.8).

[1] For simplicity of the formulas we choose two sets of equal size.

Let us consider both models of inference, transductive and inductive from one unified point of view: In both cases we are given a set of functions defined on some space R. We randomly choose the training examples from this space. In the inductive case we choose by sampling from the space, and in the transductive case we choose by splitting the working set into the training and testing parts. We define the values of the function of interest over the domain of definition of the function: In the inductive case in the whole space, and in the transductive case on the working set.

The difference is that in transductive inference the space of interest is discrete (defined on $\ell + k$ elements of the working set (3.3)), while in inductive inference it is R^n.

One can conduct a nontrivial analysis of the discrete space but not the continuous space R^n. This is the key advantage of transductive inference.

3.1.3 Large Margin Transductive Inference

Let $F_1, \ldots, F_N$ be the set of equivalence classes defined by the working set (3.3). Our goal is to construct an appropriate structure on this set of equivalence classes.

In Chapter 2, Section 2.6 we constructed a similar structure on the set of equivalence classes for inductive inference. However, we violated one of the important requirements of the theory: The structure must be constructed *before* the training data appear. In fact we constructed it *after* (in the inductive mode of inference the set of equivalence classes was defined by the training data), creating a *data-dependent structure*. There are technical means to justify such an approach. However, the bound for a data-dependent structure will be worse [138].

In transductive inference we construct the set of equivalence classes using a joint working set of vectors that contain both the training and test sets. Since in constructing the equivalence class we do not use information about how our space will be split into training and test sets we do not violate the statistical requirements.

Let us define the size of an equivalence class F_i by the value of the corresponding margin: We find, among the functions belonging to the equivalence class, the one that has the largest margin[2] and use the value of the margin $\mu(F_i)$ as the size of the equivalence class F_i.

Using this concept of the size of an equivalence class, SVM transductive inference suggests:

Classify the test vectors (3.2) by the equivalence class (defined on the working set (3.3)) that classifies the training data well and has the largest value of the (soft) margin.

Formally, this requires us to classify the test data using the rule

$$y_i = \operatorname{sgn}((w_0, z_i) + b_0), \; i = \ell + 1, \ldots, \ell + k,$$

[2]We consider the hard margin setting just for the sake of simplicity. One can easily generalize this setting to the soft margin situation as described in Section 2.3.4.

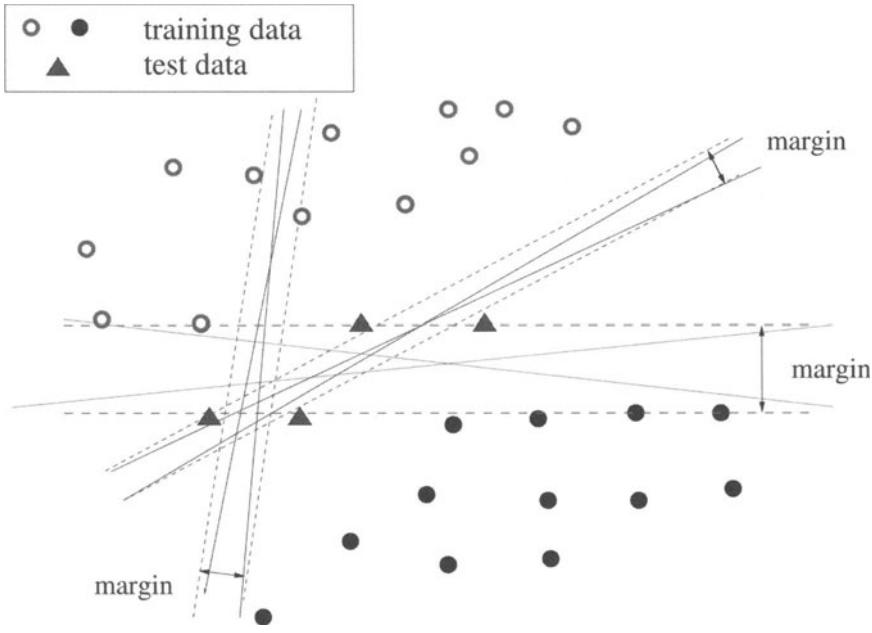

Figure 3.1: Large margin defines a large equivalence class.

where the parameters w_0, b_0 are the ones that minimize the functional

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j),\ C_1, C_2 \geq 0 \tag{3.10}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, \ell \tag{3.11}$$

(defined by the images of the training data (3.1)) and the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0,\ j = \ell + 1, \ldots, \ell + k \tag{3.12}$$

(defined by the set (3.2) and its desired classification $Y_* = (y_{\ell+1}^*, \ldots, y_{\ell+k}^*)$.

One more constraint. To avoid unbalanced solution Chapelle and Zien [174], following ideas of Joachims [154], suggested the following constraint:

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} ((w, z_j) + b) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i. \tag{3.13}$$

This constraint requires that the proportion of test vectors in the first and second categories be similar to the proportion observed in the training vectors.

For computational reasons we will replace the objective function (3.10) with the function

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \xi_i, + C_2 \sum_{s=\ell+1}^{\ell+k} \xi_s^*, \; C_1, C_2 \geq 0 \tag{3.14}$$

Therefore (taking into account kernelization based on Mercer's theorem) we can obtain the following solution of this problem (in the dual space).

The classification rules for the test data in the dual space have the form

$$y_\tau = \text{sgn}(\sum_{i=1}^{\ell} \alpha_i^0 K(x_i, x_\tau) + \sum_{s=\ell+1}^{\ell+k} \beta_s y_s^* K(x_s, x) + b_0), \; \tau = \ell+1, \ldots, \ell+k,$$

where the coefficients $\alpha_i^0, \beta_s^0, b_0$ and desired classifications of test data are the solution of the following problem: Maximize (over α, β, y^*) the functional

$$W(\alpha, \beta, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{s=\ell+1}^{\ell+k} \beta_s - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$- \sum_{i=1}^{\ell} \sum_{s=\ell+1}^{\ell+k} \alpha_i y_i \beta_s y_s^* K(x_i, x_s) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_s y_s^* \beta_t y_t^* K(x_s, x_t)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=\ell+1}^{\ell+k} y_s^* \beta_s = 0,$$

the constraints

$$0 \leq \alpha_i \leq C_1, \; i = 1, \ldots, \ell$$

$$0 \leq \beta_s \leq C_2, \; s = \ell+1, \ldots, \ell+k,$$

and the constraint (3.13):

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} \left(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_j) + \sum_{t=\ell+1}^{\ell+k} \beta_t y_t^* K(x_t, x_j b_0 \right) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i.$$

Note that this problem does not have a unique solution. This makes transductive inference difficult. However, whenever one can maximize the functional well, one obtains an improvement over inductive SVMs.

3.1.4 Examples of Transductive Inference

Here are examples of real-life problems solved using transductive inference.

1. Prediction of molecular bioactivity for drug discovery [146]. The KDD CUP-2001 competition on data analysis methods required the construction

of a rule for predicting molecular bioactivity using data provided by the DuPont Pharmaceutical company. The data belonged to a binary space of dimension 139,351, which contained a training set of 1909 vectors, and a test set of 634 vectors.

The results are given here for the winner of the competition (among the 119 competitors who used traditional approaches), SVM inductive inference and SVM transductive inference.

Winner's accuracy	68.1 %
SVM inductive mode accuracy	74.5 %
SVM transductive mode accuracy	82.3 %

It is remarkable that the jump in performance obtained due to a new philosophy of inference (transductive instead of inductive) was larger than the jump resulting from the reinforcement of the technology in the construction of inductive predictive rules.

2. TEXT CATEGORIZATION [138]. In a text categorization problem, using transductive inference instead of inductive inference reduced the error rate from 30% to 15%.

REMARK. The discovery of transductive inference and its advantages over inductive inference is not just a technical achievement, but a breakthrough in the philosophy of generalization.

Until now, the traditional method of inference was the *inductive–deductive* method, where one first defines a general rule using the available information, and then deduces the answer using this rule. That is, one goes from *particular to general* and then from *general to particular.*

In transductive mode one provides direct inference from *particular to particular*, avoiding the ill-posed part of the inference problem (inference from particular to general).

3.1.5 TRANSDUCTIVE INFERENCE THROUGH CONTRADICTIONS

Replacing the maximal margin generalization principle with the maximal contradiction on the Universum (MCU) princple leads to the following algorithm: Using the working set (3.3) create a set of equivalence classes of functions, then using the Universum (2.67) calculate the size of the equivalence classes by the number of contradictions.

The recommendation of SRM for such a structure would be:

> *To classify test vectors (3.2), choose the equivalence class (defined on the working set (3.3)) that classifies the training data (3.1) well and has the largest number of contradictions on the Universum.*

The idea of maximizing the number of contradictions on the Universum can have the following interpretation:

> *When classifying the test vectors, be very specific; try to avoid extra generalizations on the Universum (2.67).*

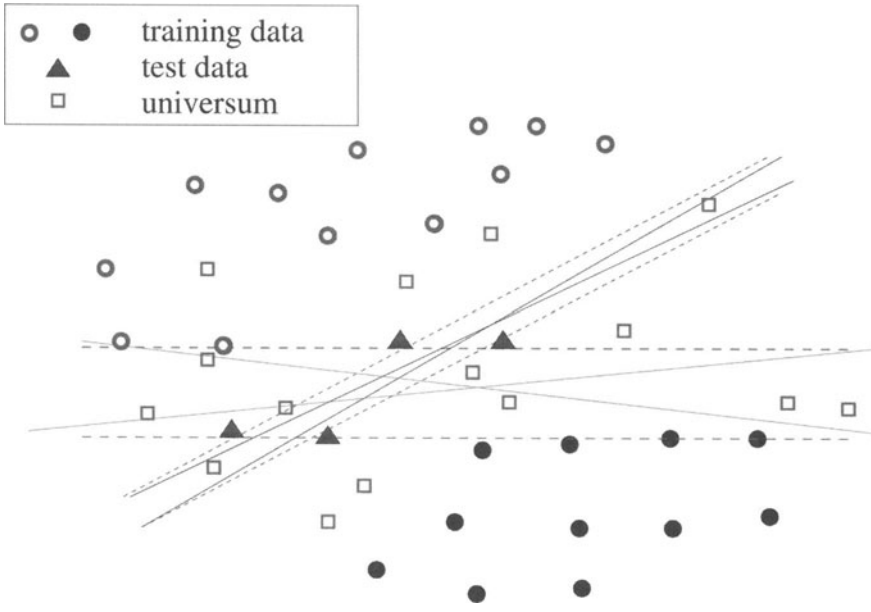

Figure 3.2: Large number of contradictions on Universum (boxes inside the margin) defines a large equivalence class.

From a technical point of view, the number of contradictions takes into account the inhomogeneity of image space, especially when the input vectors are nonlinearly mapped into feature space.

Technically, to implement transductive inference through contradictions one has to solve the following problem.

Given the images of the training data (3.1), the images of the test data (3.2), and the images of the Universum (2.67), construct the linear decision rule

$$I(x) = \theta[(w_0, z) + b_0],$$

where the vector w_0 and threshold b_0 are the solution of the following optimization problem: Minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j) + C_3 \sum_{s=1}^{u} \theta(\xi_s^*), \quad C_1, C_2, C_3 \geq 0 \tag{3.15}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \; i = 1, \ldots, \ell \tag{3.16}$$

(defined by the images of the training data (3.1)), the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0, \; j = \ell + 1, \ldots, \ell + k \tag{3.17}$$

(defined by the set (3.2) and the desired vector $(y^*_{\ell+1}, \ldots, y^*_{\ell+k})$), and the constraints

$$|(z^*_s, w) + b| \le a + \xi^*_s, \quad \xi^*_s \ge 0, \ s = 1, \ldots, u, \ \ a \ge 0 \tag{3.18}$$

(defined by the images of the Universum (2.67)).

As before (for computational reasons), we replace $\theta(\xi)$ in the objective function with ξ. Therefore we minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{j=\ell+1}^{\ell+k} \xi_j + C_3 \sum_{s=1}^{u} \xi^*_s, \quad C_1, C_2, C_3 \ge 0 \tag{3.19}$$

subject to the constraints (3.16), (3.17), and (3.18).

Dual Form Solution

The solutions to all of the above problems in the dual space of Lagrange multipliers can be unified as follows. Find the function

$$f(x) = \sum_{i=1}^{\ell} \alpha^0_i y_i K(x, x_i) + \sum_{t=\ell+1}^{\ell+k} \beta^0_t y^*_t K(x, x_t) + \sum_{m=1}^{u} (\mu^0_m - \nu^0_m) K(x, x^*_m) + b_0 \tag{3.20}$$

whose test classifications y^*_j and coefficients $\alpha^0, \beta^0, \mu^0, \nu^0, b_0$ maximise the functional

$$W(\alpha, \beta, \gamma, \mu, \nu, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{t=\ell+1}^{\ell+k} \beta_t - a \sum_{n=1}^{u} (\mu_n + \nu_n) \tag{3.21}$$

$$-\frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_t y^*_t \beta_s y^*_s K(x_t, x_s)$$

$$-\frac{1}{2} \sum_{m,n=1}^{u} (\mu_m - \nu_m)(\mu_n - \nu_n) K(x^*_m, x^*_n) - \sum_{i=1}^{\ell} \sum_{t=\ell+1}^{\ell+k} \alpha_i y_i \beta_t y^*_t K(x_i, x_t)$$

$$-\sum_{i=1}^{\ell} \sum_{m=1}^{u} \alpha_i y_i (\mu_m - \nu_m) K(x_i, x^*_m) - \sum_{m=1}^{u} \sum_{t=\ell+1}^{\ell+k} (\mu_m - \nu_m) \beta_t y^*_t K(x^*_m, x_t)$$

subject to the constraints

$$0 \le \alpha_i \le C_1, \tag{3.22}$$

$$0 \le \beta_t, \le C_2, \tag{3.23}$$

$$0 \le \mu_m, \nu_m \le C_3, \tag{3.24}$$

and the constraint

$$\sum_{i=1}^{\ell} \alpha_i y_i + \sum_{t=\ell+1}^{\ell+k} \beta_t y^*_t + \sum_{m=1}^{u} (\mu_m - \nu_m) = 0. \tag{3.25}$$

In particular, when $C_2 = C_3 = 0$ we obtain the solution for the conventional SVM, when $C_2 = 0$ we obtain the solution for inductive SVMs with the Universum, and when $C_3 = 0$ we obtain the solution for transductive SVMs.

Note that just taking into account the Universum ($C_2 = 0$) does not change the convexity of the optimization task. The problem becomes nonconvex (and therefore can have a nonunique solution) only for transductive mode.

It is good to use hint (3.13) when solving transductive problems.

3.2 Beyond Transduction: The Transductive Selection Problem

The transductive selection problem was not discussed in the original Russian edition of *EDBED*. It was written at the last moment for the English translation. In *EDBED* the corresponding section (Chapter 10, Section 13) has a very technical title "The Problem of Finding the Best Point of a Given Set." Here we call this type of inference transductive selection.

3.2.1 Formulation of Transductive Selection Problem

The transductive selection problem is the following: Given the training examples (pairs $(x_i, y_i), x \in R^n$, $y \in \{-1, +1\}$, $i = 1, \ldots, \ell$) and given a working set ($x_j^* \in R^*$, $j = 1, \ldots, m$), find in the working set the k elements that belong to the first class ($y = +1$) with the highest probability.

Here are some examples of the selection problem:

- *Drug discovery.* In this problem, we are given examples of effective drugs $(x_i, +1)$ and examples of ineffective drugs $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k candidates with the highest probability of being effective drugs.

- *National security.* In this problem, we are given examples (descriptions) of terrorists $(x_i, +1)$ and examples of non-terrorists $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k most likely terrorists.

Note that in contrast to general transductive inference, this setting does not require the classification of all candidates[3]. The key to solving the selective inference problem is to create an appropriate factorization of a given set of functions that contains fewer equivalence classes than the factorization for transductive inference. The transductive selective models are the main instrument for solving decision-making problems in high-dimensional spaces. However, this instrument has not yet been developed.

[3]In such problems, the most difficult cases are "border candidates." In transductive selection problems, we exclude this most difficult part of the task (classification of border candidates). Here again we obtain the same advantage that we obtained by replacing the model identification scheme by the prediction scheme and replacing the predictive scheme by the transductive scheme: we replaced a not very well-posed problem by a better-posed problem.

3.3 Directed Ad Hoc Inference (DAHI)

3.3.1 The Idea Behind DAHI

This section discusses *directed ad hoc inference*, inference that occupies an intermediate position between *inductive–deductive* inference and *transductive* inference.

The main idea of DAHI is a reconsideration of the roles of the training and testing stages during the inference process. The classical *inductive–deductive* model of inference contains two different stages:

(1) The training (inductive) stage, where one constructs a rule for classification using the training data, and

(2) The testing (deductive) stage where one classifies the test data using the constructed rule.

The *transductive* model of inference solves the classification problem in one step:

- Given a set of training data and a set of test data, it finds the labels for the test data directly.

DAHI works differently. During the training stage, DAHI looks for a principal direction (concept) used to construct different rules for future inferences. This is different from the inductive stage of inference where the goal is to find one fixed rule. During the test stage DAHI uses this principal direction to construct a specific rule for each given test vector (the ad hoc rule). Therefore, DAHI contains elements of both inductive and transductive inference:

(1) It constructs one general direction of inference (as in inductive inference).

(2) It constructs an individual (ad hoc) rule for each given test example (as in transductive inference).

The idea of DAHI is: *To construct a linear (in feature space) decision rule that has fixed homogeneous terms and individual (for different test vectors) thresholds.*

The problem is how to find thresholds that make inferences more accurate than ones based on one fixed threshold (as in SVM).

From a technical point of view DAHI is a combination of ideas from statistical learning theory (in particular, support vector machines), and from nonparametric statistics (methods for conditional probability estimation).

3.3.2 Local and Semi-Local Rules

To discuss the details of DAHI let us consider the idea of *local algorithms* suggested by nonparametric statistics and in particular the k-nearest neighbors method.

k-NEAREST NEIGHBORS METHOD

According to the k-nearest neighbours method for any point of interest x_0 one chooses from the training data the k-nearest (in a given metric) vectors x_i, $i = 1, \ldots, k$ and classifies the point of interest x_0 depending on which class dominates among these k chosen vectors.

The k-nearest neighbors method can be described as a *local* estimating method. Consider the set of constant-valued functions. For a set of indicator functions it contains only two functions: one takes the value -1; another takes the value 1. Consider the following local algorithm: define the spherical vicinity of the point of interest x_0 based on the given metric and a value for the radius (defined by the distance from a point of interest x_0 to its k nearest neighbors). Then choose from the admissible set of functions the function that minimizes the empirical loss on the training vectors belonging to the vicinity of the point of interest x_0. Finally use this function to classify the point of interest.

This description of the k-nearest neighbors method as a local algorithm immediately allows one to generalize it in two respects:

(1) One can use a richer set of admissible functions (for example, the set of large margin linear decision rules, see Section 2.3)

(2) One can use different rules to specify the value of the radius that defines the locality (not just the distance to the kth nearest neighbor).

In 1992 the idea of local algorithms for pattern recognition was used where (local) linear rules (instead of local constant rules) and VC bounds (instead of the distance to the kth nearest neighbor) were utilized [145]. The local linear rules demonstrated a significant improvement in performance (3.2% error rate instead of 4.1% for digit recognition on the US Postal Service database).

For the regression estimation problem a similar idea was used in the Nadaraya–Watson estimator [147, 148] with a slightly different concept of locality. Nadaraya and Watson suggested considering "soft locality": they introduced a weight function (e.g., a monotonically decreasing nonnegative function from the distance between a point of interest x_0 and elements x_i of training data $f(||x_0 - x_i||)$, $i = 1, \ldots, \ell$), and used this function for estimating the value of interest

$$y_0 = \sum_{i=1}^{\ell} \tau_i(x_0) y_i, \tag{3.26}$$

where coefficients $\tau_i(x_0)$ were defined as follows,

$$\tau_i(x_0) = \frac{f(||x_0 - x_i||)}{\sum_{i=1}^{\ell} f(||x_0 - x_i||)}. \tag{3.27}$$

This concept is a generalization of the hard locality concept. We will use this construction later.

However in all of these methods the concept of locality is the same: it is a sphere (a "soft sphere" in the Nadaraya–Watson method) defined by a given metric with the center at the point of interest.

SEMI-LOCAL RULE

In DAHI we use a new concept of vicinity. We map input vectors x into a feature space z where we specify the vicinity. We consider a cylinder (or more generally a "soft cylinder"; see Section 3.3.4 below) whose axis passes through the image z_0 of the point of interest x_0. The defined vicinity is unbounded in one direction (defined by the axis of the cylinder) and bounded in all other directions. We call such a vicinity a *semi-local* vicinity.

The difference between the local and semi-local concepts of vicinity is the following. In a sphere with a fixed center there are no preferable directions in a feature space, while a cylinder has one preferable direction (along the axis of the cylinder). DAHI uses this direction to define vicinities for all points of interest.

During the training stage DAHI looks for the direction of the cylinder that defines the axis (in feature space) for all possible vicinities (cylinders). To find this direction one can use the methods of statistical learning theory (e.g., SVMs).

During the test stage DAHI uses only data from the (semi-local) vicinity of the point of interest z_0 and constructs a one-dimensional conditional probability function defined on the axis of the cylinder passing in the specified direction w_0 through the point of interest z_0. DAHI then uses this conditional probability $P(y_0 = 1|z_0)$ to classify z_0, where z_0 is the image of the point of interest x_0 in feature space.

Note that DAHI generalizes the SVM idea. In SVM one chooses both the direction w_0 and the threshold b_0 for the decision rule. In DAHI one chooses only the direction w_0, and for any test vector constructs an individual decision rule (threshold).

3.3.3 ESTIMATION OF CONDITIONAL PROBABILITY ALONG THE LINE

To solve the classification part of the problem we estimate the conditional probability $P(y(t) = 1|t)$ that the point t on the axis of a cylinder (passing through the point of interest t_0) belongs to the first class. To do this we have to solve the integral equation

$$\int_a^t P(y = 1|t')dF(t') = F(y = 1, t), \tag{3.28}$$

where both the cumulative distribution function of the point on the line $F(t)$ and the probability function $F(y = 1, t)$ of that point on the line with $t' \leq t$ belong to the first class are unknown, but data (inside cylinder) are given.

Note that when the density function $p(t)$ exists for $F(t)$, the conditional probability

$$P(y = 1|t) = \frac{p(y = 1, t)}{p(t)}$$

defines the solution of Equation (3.28).

To solve this problem given data one must first estimate the cumulative distribution functions along the line and then use these estimates $F_{est}(t)$, $F_{est}(1,t)$ in Equation (3.28) instead of the actual functions $F(\xi)$ and $F(y=1,\xi)$.

$$\int_a^t P(y=1|t')dF_{mp}(t') = F_{emp}(y=1,t). \tag{3.29}$$

This Equation forms an ill-posed problem where not only the right-hand side of the equation is an approximation of the real right-hand side but also the operator is an approximation of the real operator (since we use $F_{emp}(t)$ instead of $F(t)$).

In [140] it is shown that if the approximations $F_{emp}(t)$, and $F_{emp}(y=1,t)$ are consistent then there exists a law $\gamma_\ell = \gamma(\ell)$ such that the Tikhonov regularization method

$$R(P) = \left|\left| \int_a^t P(y=1|t')dF_{emp}(t') - F_{emp}(y=1,t) \right|\right|^2 + \gamma_\ell \Omega(P) \tag{3.30}$$

provides the solutions that converge to the solution of Equation (3.28) as $\ell \to \infty$.

3.3.4 Estimation of Cumulative Distribution Functions

A consistent method of estimating cumulative distribution functions along a line was first suggested by Stute in 1986 [149]. He considered a cylinder of radius r whose axis coincides with the line, projected on this line the vectors z of the training data that were inside the cylinder (suppose that there are $r(\ell)$ such vectors), and constructed a one-dimensional empirical distribution function using these projections:

$$F^*_{r(\ell)}(x) = \frac{1}{r(\ell)} \sum_{i=1}^{r(\ell)} \theta(t - t_i). \tag{3.31}$$

Stute showed that under some general law of choosing the radius of the cylinder (which depends on the number of observations ℓ) with an increasing number of observations, this empirical cumulative distribution function converges with probability one to the desired function. To estimate conditional probability one can use in (3.30) the approximation (3.31) and the approximation

$$F_{r(\ell)}(1,t) = \frac{1}{2r(\ell)} \sum_{i=1}^{r(\ell)} (1+y_i)\theta(t - t_i). \tag{3.32}$$

Also one can estimate a cumulative distribution function along the line in the Nadaraya-Watson style using the distances between images of training vectors and the line passing through the point of interest z_0 in direction w_0 ,

$$d_i(z_0) = \sqrt{|z_i - z_0|^2 - t_0^2}, \tag{3.33}$$

where $t_0 = (z_0, w_0)$ is the projection of the vector z_0 on the direction w_0. Using $d_i(z_0)$ instead of $||x_0 - x_i||$ in (3.27) one obtains the Nadaraya–Watson type approximations of the elements of Equation (3.29):

$$F_{emp}(t) = \sum_{i=1}^{\ell} \tau_i(z_0)\theta(t - t_i), \tag{3.34}$$

$$F_{emp}(y = 1, t) = \frac{1}{2}\sum_{i=1}^{\ell} (1 + y_i)\tau_i(z_0)\theta(t - t_i). \tag{3.35}$$

Both the Stute estimate and modified Nadaraya–Watson estimate are step functions. The difference is that in Stute's estimate there are $r(\ell)$ steps where all values of the step are equal to $1/r(\ell)$ while in the Nadarya–Watson estimate there are ℓ steps but the step values $\tau_i(z_0)$ are different, and depend on the distance between the vector z_i and the line passing through the point z_0 in the direction w_0.

3.3.5 Synergy Between Inductive and Ad Hoc Rules

In DAHI we combine two consistent methods: the SVM method for estimating the direction in feature space, and the method for estimating the conditional probability along the line passing through the point of interest.

However, when the number of training data is not large (and this is always the case in a high-dimensional problem) one needs to provide both methods with additional information: In order to choose a good SVM solution one has to map the input vectors into a "good" Hilbert space (to choose a "good" kernel). In order to obtain a good solution for solving the ill-posed problem of estimating a conditional probability function along the line one has to use a priori information about the admissible set of functions that contain the desired conditional probability function.

By combining the above two methods, one tries to construct a robust classification method that reduces the dependency on a priori information.

This is because:

(1) When one chooses a direction that is "reasonably close" to the one that defines a "good" separating hyperplane, the corresponding conditional probability function belongs to the set of *monotonic* functions (the larger the SVM score is, the larger is the probability of the positive class). Finding a direction that maintains the monotonicity property for the conditional probabilities requires fewer training examples than finding a direction that provides a good classification.

(2) The problem of finding a conditional probability function from the set of monotonic nondecreasing functions is much better posed than the more general problem of finding a solution from the set of continuous nonnegative functions.[4]

Therefore, in the set of monotonic functions one can solve this problem well, using a restricted (small) number of observations.

[4] A set of monotonically increasing (or monotonically decreasing) functions has VC dimension one while a set of continuous nonnegative functions has an infinite VC dimension.

(3) Using the leave-one-out technique one can use the same training data for constructing the main direction and later for constructing conditional probability functions.

The minimization of functional (3.30) in a set of monotonic functions is not too difficult a computational problem. The idea behind DAHI is to use this possible synergy.

Figure 3.3 shows two examples of the binary classification problem: separating digit 3 from digit 5. Two examples of conditional probabilities $P(3|t)$ estimated along the line are presented in Figure 3.3. For each example the figure shows the image of interest, the functions $F_{emp}(t)$ and $F_{emp}(3,t)$, and the solution of Equation (3.29). The position of the point of interest on the line corresponds to an ordinate value of 0. Part (a) of the figure shows the probability that the image is a 3 is 0.34, but in part (b) the probability that the image is a 3 is 0.

3.3.6 DAHI AND THE PROBLEM OF EXPLAINABILITY

The idea of DAHI is appealing from a philosophical point of view since it addresses the question of *explainability* of complex rules [169]. DAHI divides the model of explainability for complex rules into two parts: the "main direction" and the "ad hoc" parts where only the "main direction" part of the rule has to be explained (described by the formal model).

One speculation on the DAHI model of explainability can be given by the example of how medical doctors distinguish between cancer and benign cases. They use principal rules to evaluate the cancer and if the corresponding score exceeds a threshold value, they decide the case is cancer.

The threshold, however, is very individual: it depends on the family history of the patient, and many other factors. The success of a doctor depends on his experience in determining the individual threshold. The threshold can make all the difference in diagnostics. Nevertheless the explainability is mostly related to the "main direction" part of the rule.

3.4 PHILOSOPHY OF SCIENCE FOR A COMPLEX WORLD

3.4.1 EXISTENCE OF DIFFERENT MODELS OF SCIENCE

The limitations of the classical model of science when dealing with the real-life complex world have been discussed for quite some time. For example, according to Einstein, the classical model of science is relevant for a simple world. For a complex world it is inapplicable.

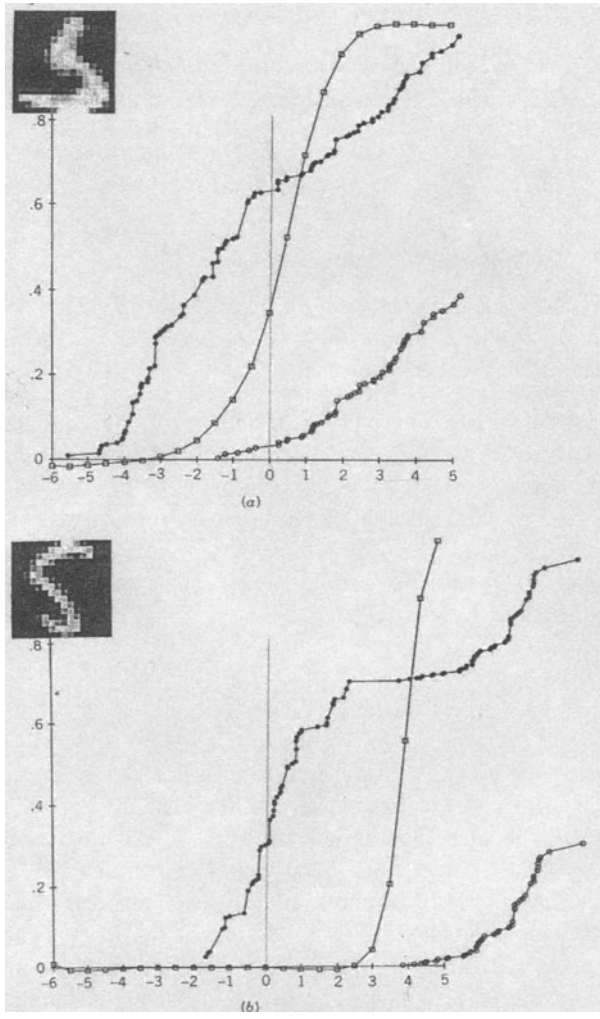

Figure 3.3: Solutions of the integral equation for different data.

- Einstein on the simple world:

  When the solution is simple, God[5] *is answering.*

- Einstein on the complex world:

  When the number of factors coming into play in a phenomenological complex is too large, scientific methods in most cases fail.[6]

One can see the idea of limitation of scientific models and existence of non-scientific ones in the following Richard Feynman's remark (*Lectures on physics*):

> *If something is said not to be a science, it does not mean that there is something wrong with it ... it just means that it is not a science.*

In other words there was an understanding that:

> *Classical science is an instrument for a simple world. When a world is complex, in most cases classical science fails. For a complex world there are methods that do not belong to classical science.*

Nevertheless, the success of the physical sciences strongly influenced the methodology used to analyze the phenomena of a complex world (one based on many factors). In particular, such a methodology was adopted in the biological, behavioural, and social sciences where researchers tried to construct low-dimensional models to explain complex phenomena.

The development of machine learning technology challenged the research in the methodology of science.

3.4.2 Imperative for a Complex World

Statistical learning theory stresses that the main difficulties of solving generalization problems arise because, in most cases, they are ill-posed.

To be successful in such situations, it suggests to give up attempts of solving ill-posed problems of interest replacing them by less demanding but better posed problems. In many cases this leads to renunciation of explainability of obtained solutions (which is one of the main goals declared by the classical science). Therefore, a science for a complex world has different goals (may be it should be called differently).

For solving specific ill-posed problems the regularization technique was suggested [20, 21, 54, 55]. However, to advance high-dimensional problems of inference just applying classical regularization ideas is not enough. The SRM principle of inference is another way to control the capacity of admissible sets of functions. Recently a new general idea of capacity control was suggested in the form of the following imperative [139]:

[5] Here and below Einstein uses the word God as a metaphor for nature.

[6] Great theoretical physicist Lev Landau did not trust physical theories that combine more than a few factors. This is how he explained why: "With four free parameters one can draw an elephant, with five one can draw an elephant rotating its tail."

IMPERATIVE

When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.

According to this imperative:
— Do not estimate a density if you need to estimate a function.
(Do not use the classical statistics paradigm for prediction in a high-dimensional world: Do not use generative models for prediction.)
— Do not estimate a function if you only need to estimate its values at given points. *(Try to perform direct inference rather than induction.)*
— Do not estimate predictive values if your goal is to act well.
(A good strategy of action does not necessary rely on good predictive ability.)

3.4.3 RESTRICTIONS ON THE FREEDOM OF CHOICE IN INFERENCE MODELS

In this Afterword we have discussed three levels of restrictions on the freedom of choice in the inference problem:

(1) *Regularization*, which controls the smoothness properties of the admissible set of functions (it forbids choosing an approximation to the desired function from not a "not smooth enough set of functions").

(2) *Structural risk minimization*, which controls the diversity of the set of admissible functions (it forbids choosing an approximation to the desired function from too diverse a set of functions, that is, from the set of functions which can be falsified only using a large number of examples).

(3) *Imperatives*, which control the goals of possible inferences in order to consider a better-posed problem. In our case it means creating the concept of equivalence classes of functions and making an inference using a large equivalence class (it forbids an inference obtained using a "small" equivalence class).

It should be noted that an understanding of the role of a general theory as an instrument to restrict directions of inference has existed in philosophy for a long time. However, the specific formulations of the restrictions as described above were developed only recently. The idea of using regularization to solve ill-posed problems was introduced in the mid-1960s [21, 55]. Structural risk minimization was introduced in the early 1970s [*EDBED*], and the imperative was introduced in the mid-1990s [139].

In order to develop the philosophy of science for a complex world it is important to consider different forms of restriction on the freedom of choice in inference problems and then analyze their roles in obtaining accurate predictive rules for the pattern recognition problem.

One of the main goals of research in the methodology of analysis of a complex world is to introduce new imperatives and for each of them establish interpretations in the corresponding branches of science.

3.4.4 Metaphors for Simple and Complex Worlds

I would like to finish this part of the Afterword with metaphors that stress the difference in the philosophy for simple and complex worlds. As such metaphors let me again use quotes from Albert Einstein.

Two Metaphors for a Simple World

1. *I want to know God's thoughts.* (*A. Einstein*)
2. *When the solution is simple, God is answering.* (*A. Einstein*)

Interpretation

Nature is a realization of the simplest conceivable mathematical ideas. I am convinced that we can discover, by means of purely mathematical constructions, concepts and laws, connect them to each other, which furnish the key to understanding of natural phenomena. (*A. Einstein.*)

Three Metaphors for a Complex World

First Metaphor

Subtle is the Lord, but malicious He is not. (*A. Einstein*)

Interpretation[7]

Subtle is the Lord — one can not understand His thoughts.

But malicious He is not — one can act well without understanding them.

Second Metaphor

The devil imitates God.[8] (*Medieval concept of the devil.*)

Interpretation

Actions based on your understanding of God's thoughts can bring you to catastrophe.

Third Metaphor

If God does exist then many things must be forbidden. (*F. Dostoevsky*)

Interpretation

If a subtle and nonmalicious God exists, then many ways of generalization must be forbidden. The subject of the complex world philosophy of inference is to define corresponding imperatives (to define what should be forbidden). These imperatives are the basis for generalization in real-life high-dimensional problems.

The imperative described in Section 3.4.2 is an example of the general principle that forbids certain ways of generalization.

[7] Surely what Einstein meant is that the laws of nature may be elusive and difficult to discover, but not because the Lord is trying to trick us or defeat our attempts to discover them. Discovering the laws of nature may be difficult, but *it is not impossible*. Einstein considered comprehensibility of the physical world as a "mystery of the world". My interpretation of his metaphor for a *complex world* given below is different.

[8] This includes the claim that for humans the problem of distinguishing imitating ideas of the devil from thoughts of God is ill-posed.