24 Transductive Inference and Semi-Supervised Learning

Vladimir Vapnik VAPNIK@ATT.NET

This chapter discusses the difference between transductive inference and semi-supervised learning. It argues that transductive inference captures the intrinsic properties of the mechanism for extracting additional information from the unlabeled data. It also shows an important role of transduction for creating noninductive models of inference.[1]

24.1 Problem Settings

Let us start with the formal problem setting for transductive inference and semi-supervised learning.

Transductive Inference: General Setting Given a set of ℓ training pairs,

$$(y_1, x_1), ...(y_\ell, x_\ell), \quad x_i \in \mathbb{R}^d, \; y_i \in \{-1, 1\}, \tag{24.1}$$

and a sequence of k test vectors,

$$x_{\ell+1}, ..., x_{\ell+k}, \tag{24.2}$$

find among an admissible set of binary vectors,

$$\{Y = (y_{\ell+1}, ..., y_{\ell+k})\},$$

1. These remarks were inspired by the discussion, What is the Difference between Transductive Inference and Semi-Supervised Learning?, that took place during a workshop close to Tübingen, Germany (May 24, 2005).

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

$$x_1, ..., x_{\ell+k} \tag{24.3}$$

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 $P(y|x)$.

Below we will call the vectors (24.3) from the training and test sets the *working set* of vectors.

Transductive Inference: Particular Setting In this setting 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:

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

Semi-Supervised Learning Given a set of training data (24.1) and a set of test data (24.2), find among the set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$, the one that minimizes the risk functional

$$R(\alpha) = \int |y - f(x, \alpha)| dP(x, y). \tag{24.4}$$

Therefore, in transductive inference the goal is to classify the given u test vectors of interest while in semi-supervised learning the goal is to find the function that minimizes the functional (24.4) (the expectation of the error).

Semi-supervised learning can be seen as being related to a particular setting of transductive learning. Indeed, if one chooses the function to classify the given test data (24.2) well, why not also use it to classify new unseen data? This looks like a reasonable idea.

However from a conceptual point of view, transductive inference contains important elements of a new philosophy of inference and this is the subject of these remarks.

The transductive mode of inference was introduced in the mid-1970s. It attempts to estimate the values of an unknown function $f(x, \alpha_0)$ at particular points of interest. On the other hand, inductive inference attempts to estimate the unknown function over its entire domain of definition (Vapnik, 2006). In the late 1970s the advantage of transductive inference over inductive inference was shown on real life problems (Vapnik and Sterin, 1977).

The problem of semi-supervised learning was introduced in the mid-1990s (cf. section 1.1.3) and became popular in the early 2000s (Zhou et al., 2004).

24.2 Problem of Generalization in Inductive and Transductive Inference

The mechanism that provides the transductive mode of inference with an advantage over the inductive mode in classification of the given points of interest has been understood since the very first theorems of Vapnik-Chervonenkis (VC) theory were proved.

Suppose that our goal is to find the function that minimizes the functional (24.4). Since the probability measure in (24.4) is unknown we minimize the empirical risk functional

$$R_{emp}(\alpha) = \sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)| \tag{24.5}$$

instead of the risk functional (24.4).

It was shown in (Vapnik and Chervonenkis, 1991), that the necessary and sufficient conditions for consistency (as ℓ increases) of the obtained approximations is the existence of the uniform convergence of frequencies (defined by (24.5)) to their probabilities (defined by (24.4)) over a given set of functions $f(x, \alpha), \alpha \in \Lambda$:

$$P\left\{\sup_{\alpha} |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \longrightarrow 0, \quad \forall \varepsilon > 0. \tag{24.6}$$

In 1968 the necessary and sufficient conditions for uniform convergence (24.6) were discovered (Vapnik and Chervonenkis, 1968, 1971). They are based on the so-called capacity factors. These factors will play an important role in our discussion. We now introduce them.

24.2.1 The VC Entropy, Growth Function, and VC Dimension

Given a set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$ and set of ℓ i.i.d. input vectors

$$x_1, ..., x_\ell, \tag{24.7}$$

consider the value $\Delta^{\Lambda}(x_1, ..., x_\ell)$ that defines the number of different classifications of the set of vectors (24.7) using indicator functions from the set $f(x, \alpha),\ \alpha \in \Lambda$. This is the number of *equivalence classes*[2] of functions on which the set of vectors (24.7) factorizes the set of functions $f(x, \alpha),\ \alpha \in \Lambda$. The number of equivalence classes has the trivial bound

$$\Delta^{\Lambda}(x_1, ..., x_\ell) \leq 2^{\ell}. \tag{24.8}$$

Using the value $\Delta^{\Lambda}(x_1, ..., x_\ell)$ we define the following three capacity concepts.

2. A subset of functions that classify vectors (24.7) in the same way belong to the same equivalence class (with respect to (24.7)).

1. The expectation of the number of equivalence classes,

$$\Delta_P^{\Lambda}(\ell) = E_{x_1,...,x_\ell}\Delta(x_1, ..., x_\ell), \tag{24.9}$$

where the expectation is taken over i.i.d. data (24.7) drawn according to the distribution $P(x)$.
The function

$$H_P^{\Lambda}(\ell) = \ln \Delta_P^{\Lambda}(\ell) \tag{24.10}$$

forms the first capacity concept. It is called the (annealed) *VC entropy.*[3]
The VC entropy depends on three factors:

(a) the set of functions $f(x, \alpha),\ \alpha \in \Lambda$,

(b) the number of vectors ℓ, and

(c) the probability measure $P(x)$.

The condition

$$\lim_{\ell \longrightarrow \infty} \frac{H_P^{\Lambda}(\ell)}{\ell} = 0 \tag{24.11}$$

forms the necessary and sufficient condition for uniform convergence (24.6) *for the fixed probability measure* $P(x)$.

2. The second capacity concept is called *the growth function.* It is defined as

$$G^{\Lambda}(\ell) = \max_{x_1,...,x_\ell} \Delta^{\Lambda}(x_i, ..., x_\ell). \tag{24.12}$$

The value of the growth function depends on two factors:
a. the set of functions $f(x, \alpha),\ \alpha \in \Lambda$, and
b. the number of observations ℓ.
The condition

$$\lim_{\ell \longrightarrow \infty} \frac{\ln G^{\Lambda}(\ell)}{\ell} = 0 \tag{24.13}$$

forms the necessary and sufficient condition for uniform convergence that is *independent of the probability measure* (for all probability measures).

3. The third capacity concept is called *the VC dimension.*[4]
We say that a set of functions $f(x, \alpha),\ \alpha \in \Lambda$ has VC dimension h if the largest number ℓ for which the equality

$$G^{\Lambda}(\ell) = 2^{\ell} \tag{24.14}$$

holds true is equal to h. If this equality is true for any ℓ we say that the VC

3. The abbreviation for Vapnik-Chervonenkis entropy.
4. The abbreviation for Vapnik-Chervonenkis dimension.

dimension equals infinity. In other words

$$h = \max_{\ell}\{\ell : \ G^{\Lambda}(\ell) = 2^{\ell}\}. \tag{24.15}$$

The VC dimension depends only on one factor: (a) the set of functions. VC dimension characterizes the diversity of this set of functions.
A finite VC dimension is the necessary and sufficient condition for uniform convergence which is independent of the probability measure.

In 1968 we proved the important bound (Vapnik and Chervonenkis, 1968)

$$\ln G^{\Lambda}(\ell) \le h\left(\ln\frac{\ell}{h} + 1\right). \tag{24.16}$$

This bound allows one to upper-bound the growth function with a standard function that depends on one parameter, the VC dimension.

We have therefore obtained the following relationship:

$$H_P^{\Lambda}(\ell) \le \ln G^{\Lambda}(\ell) \le h\left(\ln\frac{\ell}{h} + 1\right). \tag{24.17}$$

24.3 Structure of the VC Bounds and Transductive Inference

One of the key results of VC theory is the following bound:

$$P\left\{\sup_{\alpha}|R(\alpha) - R_{emp}(\alpha)| \ge \varepsilon\right\} \le \exp\{H_P^{\Lambda}(2\ell) - \varepsilon^2\ell\}. \tag{24.18}$$

One can rewrite this expression in the following form: with probability $1-\eta$ simultaneously for all α the inequality

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{H_P^{\Lambda}(2\ell) - \ln\eta}{\ell}} \tag{24.19}$$

holds true. Note that this inequality depends on the distribution function $P(x)$.

Since this inequality is true simultaneously for all functions of the admissible set, the function that minimizes the right-hand side of (24.19) provides the guaranteed minimum for the expected loss (24.4).

Taking into account (24.17) one can upper-bound (24.19) using the second capacity concept, the growth function:

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{\ln G^{\Lambda}(2\ell) - \ln\eta}{\ell}}. \tag{24.20}$$

This bound is true for any distribution function (i.e. for the worst distribution function). However it is less accurate (for a specific case $P(x)$) than (24.19).

One can also upper-bound (24.19) and (24.20) using the third capacity concept,

the VC dimension

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\ln \frac{2\ell}{h} + 1) - \ln \eta}{\ell}}. \tag{24.21}$$

The good news about this bound is that it depends on just one parameter h and not on some integer function $G^{\Lambda}(\ell)$. However (24.21) is less accurate than (24.20) which is less accurate than (24.19).

Transductive inference was inspired by the idea of finding better solutions using the more accurate bound (24.19) instead of the bounds (24.20) and (24.21) used in inductive inference.

24.4 The Symmetrization Lemma and Transductive Inference

Bounds (24.18) and (24.19) were obtained using the so-called symmetrization lemma.

Lemma. The following inequality holds true:

$$P\left\{\sup_{\alpha} |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \leq 2P\left\{\sup_{\alpha} \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\}, \tag{24.22}$$

where

$$R^{(1)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)| \tag{24.23}$$

and

$$R^{(2)}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} |y_i - f(x_i, \alpha)| \tag{24.24}$$

are the empirical risk functionals constructed using two different samples.

The bound (24.18) was obtained as an upper-bound of the right-hand side of (24.22).

Therefore, from the symmetrization lemma it follows that to obtain a bound for inductive inference we first obtained a bound for transductive inference (for the right-hand side of (24.22)) and then upper-bounded that.

It should be noted that since the bound (24.18) was introduced in 1968, a lot of efforts were made to improve it. However in all attempts the key element remained the symmetrization lemma. That is, in all 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.

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

To get the bound (24.18) let us bound the right-hand side of (24.22). Two

fundamental ideas were used to obtain this bound:

1. The following two models are equivalent: (a) one chooses two i.i.d. sets:[5]

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

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

2. Using model (b) one can rewrite the right-hand side of (24.22) as follows:

$$\begin{gathered} P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \tfrac{\varepsilon}{2}\right\} = \\ E_{\{x_1,...,x_{2\ell}\}} P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \tfrac{\varepsilon}{2} \mid \{x_1, ..., x_{2\ell}\}\right\}. \end{gathered} \tag{24.25}$$

To obtain the bound we first bound the conditional probability,

$$\begin{gathered} P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \tfrac{\varepsilon}{2} \mid \{x_1, ..., x_{2\ell}\}\right\} \leq \\ \Delta^\Lambda(x_1, ..., x_{2\ell}) \exp\left\{-\varepsilon^2 \ell\right\}, \end{gathered} \tag{24.26}$$

and then take the expectation over working sets of size 2ℓ. As a result, we obtain

$$\begin{gathered} E_{\{x_1,...,x_{2\ell}\}} P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \tfrac{\varepsilon}{2}\right\} \leq \\ E\Delta^\Lambda_P(2\ell) \exp\left\{-\varepsilon^2 \ell\right\} = \exp\left\{H^\Lambda_P(2\ell) - \varepsilon^2 \ell\right\}. \end{gathered} \tag{24.27}$$

This bound depends on the probability measure $P(x)$ (it contains the term $H^\Lambda_P(2\ell)$). To obtain a bound which is independent of the probability measure we upper-bound $H^\Lambda_P(2\ell)$ by $G^\Lambda(2\ell)$ (see (24.17)). Since $G^\Lambda(2\ell)$ is independent of the probability measure we obtain the bound

$$P\left\{\sup_\alpha |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \leq G^\Lambda(2\ell) \exp\left\{-\varepsilon^2 \ell\right\} \tag{24.28}$$

on uniform convergence that is independent of the probability measure.

Therefore, from the symmetrization lemma and (24.17) we obtained the bound (24.28). Note, however, that in order to obtain this bound we twice used a rough estimate: the first time when we used the symmetrization lemma, and the second time when we used the function $G^\Lambda(2\ell)$ instead of the function $H^\Lambda_P(2\ell)$.

24.5 Bounds for Transductive Inference

The inequality (24.26) is the key element for obtaining a VC bound for transductive inference.

Indeed, this inequality is equivalent to the following one: with probability $1 - \eta$

5. For simplicity of the formulas we choose two sets of equal size.

simultaneously for all functions $f(x,\alpha),\ \alpha \in \Lambda$, the inequality

$$\frac{1}{\ell}\sum_{i=\ell+1}^{2\ell}|y_i - f(x,\alpha)| \le \sum_{i=1}^{\ell}|y_i - f(x,\alpha)| + \sqrt{\frac{\ln \Delta^{\Lambda}(x_1,...,x_{2\ell}) - \ln \eta}{\ell}} \quad (24.29)$$

holds true, where probability is defined with respect to splitting the set $\{x_1,...,x_{2\ell}\}$ into two subsets:

1. one that is used in the training set $x_1,...,x_\ell$ and
2. one that forms the test set $x_{\ell+1},...,x_{2\ell}$.

Note that this concept of probability is different from the one defined for inductive inference and which requires the i.i.d. distribution of the elements $x_1,...,x_{2\ell}$. The concepts of probability will be equivalent if an element of the working set is i.i.d. according to some unknown *fixed* probability distribution function. If it is not, then all formal claims are still correct but the concept of probability is changing. In this sense we discuss in section 24.11.1 the idea of adaptation in transductive inference.

But even in the i.i.d. case the bound for transduction is more accurate than (24.20) and (24.21) used in inductive inference. However, the main advantage of transduction over induction appears when one implements the structural risk minimization principle.

24.6 The Structural Risk Minimization Principle for Induction and Transduction

In the 1970s the structural risk minimization (SRM) principle was introduced. Its goal was to find the function that minimizes the right-hand side of inequality (24.19). In order to achieve this goal the following scheme was considered.

Prior to the appearance of the training set, the set of admissible functions is organized as a structure. The nested subsets of functions (called the elements of the structure) are specified:

$$S_1 \subset S_2 \subset ... \subset S_B \subset S = \{f : f(x,\alpha),\ \alpha \in \Lambda\}, \quad (24.30)$$

where subset S_k has a fixed capacity (say VC dimension $h = k$).

The minimization of the right-hand side of inequality (24.29) can then be performed over two terms of the inequality. One first chooses the element of the structure (controlling the second term through the value of h_k) and then the function in the chosen element of the structure (controlling the first term).

It was shown that the SRM principle is strongly uniformly consistent (Devroye et al., 1996), (Vapnik, 1998). This means that when the sample size ℓ increases, the error of the function selected by the SRM principle converges toward the best possible error. However in order to find a good solution using a finite (limited) number of training examples one has to construct a (smart) structure which reflects prior knowledge about the problem of interest. In creating such a structure transductive inference offers some additional opportunities with respect to inductive

inference.

The SRM principle for transductive inference can be introduced as follows (Vapnik, 2006): *Prior to splitting the given working set* $x_1, ..., x_{2\ell}$ into the two subsets that define the elements of the training and test sets, one constructs the structure on the *finite* number $N = \Delta^{\Lambda}(x_1, ..., x_{2\ell})$ of equivalence classes $F_1, ..., F_N$ that are the result of factorization of the given set of functions over the given 2ℓ vectors.[6]

Let such a structure be

$$S_1^* \subset S_2^* \subset, ..., \subset S_B^* \subset S^* = \{F_1, ..., F_N\}, \tag{24.31}$$

where the subset S_k^* contains N_k equivalence classes of functions from $f(x, \alpha),\ \alpha \in \Lambda$.

The opportunity to construct a "smart" structure on the elements of the equivalence classes is a key advantage of SRM for transductive inference over SRM for inductive inference.

The new development in SRM for transductive inference comes from the consideration of the different "sizes" of the equivalence classes. The idea of creating a smart structure on the set of equivalence classes due to their size remains the hierarchical Bayesian approach. In this approach one can distinguish two (several) levels of hierarchy: Suppose that we are given a priori information $P(\alpha)$ on the set of admissible functions (before the set of vectors $x_1, ..., x_{2\ell}$ appear). After these vectors appear one can calculate prior information for equivalence classes $\mu(F_1), ..., \mu(F_N)$ as an integral

$$\mu(F_k) = \int_{F_k} dP(\alpha).$$

Using this prior information one can construct a "smart" structure where the first element contains N_1 equivalence classes with the largest values $\mu(F_i), i = 1, ..., N$, the second element contains N_2 equivalence classes with the largest value $\mu(F_i)$, and so on.

Note that for transductive inference the construction of such a structure for a given working set is a prior process since we do not use both the split of our x vectors into the training and test subsets, and information about the classification of the training data.[7]

6. The functions that take the same values on the working set of vectors $x_1, ..., x_{2\ell}$ form one equivalence class (with respect to the working set).

7. One can unify transductive and inductive inference as follows: In both cases one is given a set of functions defined on some space. One uses the training examples from this space to define the values of the function of interest for the whole space of definition of the function. The difference is that in transductive inference the space of interest is discrete (defined on the working set (24.3)) while in inductive inference it is $\mathbb{R}^d$. One can conduct a nontrivial analysis of the discrete space but not the space $\mathbb{R}^d$. This defines the key factor of the advantage of transductive inference.

For any element S_k of the structure, simultaneously for all equivalence classes belonging to this element, with probability $1-\eta$ the following inequality holds true:

$$\frac{1}{\ell}\sum_{i=\ell+1}^{2\ell}|y_i - F_r(x_i)| \le \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - F_r(x_i)| + \sqrt{\frac{\ln N_k - \ln\eta}{\ell}}, \quad F_r \in S_k. \quad (24.32)$$

The probability is defined with respect to a random split of the set of vectors (24.3) into two subsets: training and test vectors.[8]

Therefore, to minimize the number of errors on the test vectors (the left-hand side of (24.32)) we have to choose the element of the structure S_k (it defines the value of the second term in the right-hand side of (24.32)) and the equivalence class belonging to this element (it defines the value of the first term in the right-hand side of (24.32)).

24.7 Combinatorics in Transductive Inference

When constructing structures on the set of equivalence classes in discrete space one can play combinatorial tricks. This is impossible when constructing a structure on the set of functions defined in the whole space.

Suppose we are given a working set of size 2ℓ which forms our discrete space. Suppose in this space we have N equivalence classes $F_1, ..., F_N$ of functions $f(x, \alpha),\ \alpha \in \Lambda$.

Consider 2ℓ new problems described by 2ℓ discrete spaces: $S^1, ..., S^{2\ell}$, where the discrete space S^r is defined by working vectors (24.3) from which we removed the vector x_r. For each of these spaces we can construct a set of equivalence classes and a corresponding structure on this set. For each of these classes with probability $1-\eta$ the inequality (24.32) holds true and therefore simultaneously for all $2\ell+1$ problems the inequality (24.32) is true with probability $1-(2\ell+1)\eta$. Therefore with probability $1-\eta$ simultaneously for all $2\ell+1$ problems the inequality

$$R_1(F_i^s) \le R_2^s(F_i^s) + \sqrt{\frac{\ln N_k^s - \ln\eta + \ln(2\ell+1)}{\ell-1}}, \quad F_r \in S_k \quad (24.33)$$

holds true, where the term $\ln(2\ell+1)$ is due to our combinatorial games with one element of the working set. One can find an analogous bound for a combinatorial game with k elements of the working set.

Combinatorial games allow one to introduce a very deep geometric concept of equivalence classes (see (Vapnik, 2006, 1998) for details).

8. One can obtain a better bound (see (Vapnik, 1998)).

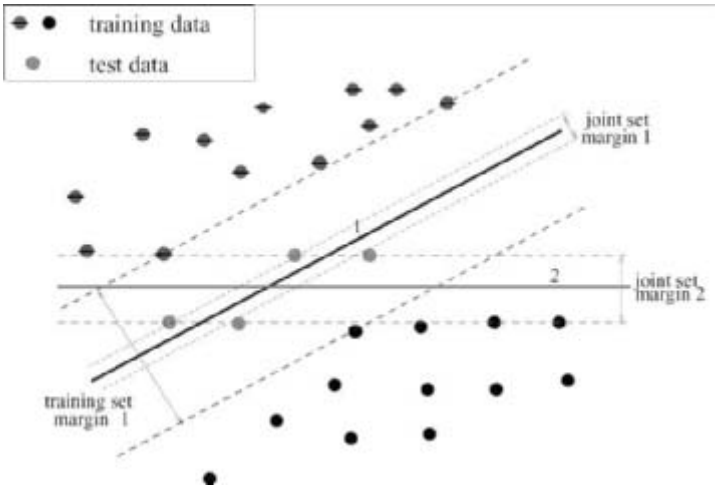

Figure 24.1 The large-margin hyperplane obtained using only the training set does not belong to the largest equivalence class defined on the working set.

24.8 Measures of the Size of Equivalence Classes

We have not yet discussed how to measure the size of equivalence classes. In this section we will discuss two possibilities. We could either

1. use a measure that reflects the VC dimension concept for the set of linear (in a feature space) indicator functions: the value of a margin for the equivalence class, or

2. measure the size of equivalence classes using the most refined capacity concept: the VC entropy.

Using the size of the margin for equivalence class. With the appearance of support vector machines (SVMs) the important problem became the following: given a working set of vectors (24.3) construct a structure on the equivalence classes of linear functions.

Let us measure the size $\mu(F_i)$ of an equivalence class F_i by the value of the corresponding margin.[9] Any equivalence class separates working vectors (24.3) into two classes. Let us find among the functions belonging to the equivalence class one that has the largest distance (margin) to the closest vector of the set (24.3). We use this distance as the measure $\mu(F_i)$ for the size of the equivalence class F_i. This measure and how it differs from the SVM are illustrated in figure 24.1.

Using this concept of the size of an equivalence class the SVM transductive algorithms were suggested Vapnik (2006).

9. There is a direct connection between the value of the margin and the VC dimension defined on the set of equivalence classes (see (Vapnik, 1998, chapter 8)).

The recommendation of SRM for transductive inference would be:
To classify test vectors (24.2) choose the equivalence class (defined on the working set (24.3)) that classifies the training data well and has the largest value of the (soft) margin.

This idea is widely used in constructing transductive SVM algorithms (see chapter 6).

The universum concept: To construct a measure on the size of the equivalence class based on the most refined capacity concept, the VC entropy, the following idea was introduced in (Vapnik, 1998). Suppose that for a given working set of data (24.3) we construct additionally a new set of data,

$$x_1^*, ..., x_u^*, \quad x^* \in \mathbb{R}^d, \tag{24.34}$$

called the universum. Using the working set (24.3) we will create a set of equivalence classes of functions, and using the universum (24.34) we will evaluate the size of the equivalence classes.

The universum plays the role of prior information in Bayesian inference. It describes our knowledge of the problem we are solving. There exist, however, important differences between prior information in Bayesian inference and prior information given by the universum. In Bayesian inference, prior information is information about the relationship of the functions in the set of admissible functions to the desired one. The universum is information about a relationship between the working set and a set of possible problems. For example, for the digit recognition problem it can be some vectors whose images resemble a digit. It defines a style of digits for the recognition task.

Using the value of the VC entropy defined on the universum. Consider now the set of equivalence classes defined by the working set (24.3). Let us measure the size $\mu(F_k)$ of the equivalence class F_r by the value $\ln \Delta^{F_r}(x_1^*, ..., x_v^*)$. This defines the logarithm of the number of different separations of the vectors from the universum (24.34) by the functions belonging to this equivalence class. This measure defines the diversity of the functions from the equivalence class. The size of the equivalence classes decreases with the index in the structure.

The recommendation of SRM for transductive inference would be:
To classify test vectors (24.2) choose the equivalence class (defined on the working set (24.3)) that classifies the training data (24.1) well and has the largest value of the VC entropy (the largest diversity) on the universum (24.34).

Using the number of contradictions on the universum. Unfortunately it is not easy to estimate the values of the VC entropy of equivalence classes on the universum. Therefore we simplify this measure. Let us consider the vector x_i^* as one that contradicts equivalence class F_r if in class F_r there are functions that classify this vector as belonging to the first category as well as functions that classify x_i^* as belonging to the second category.

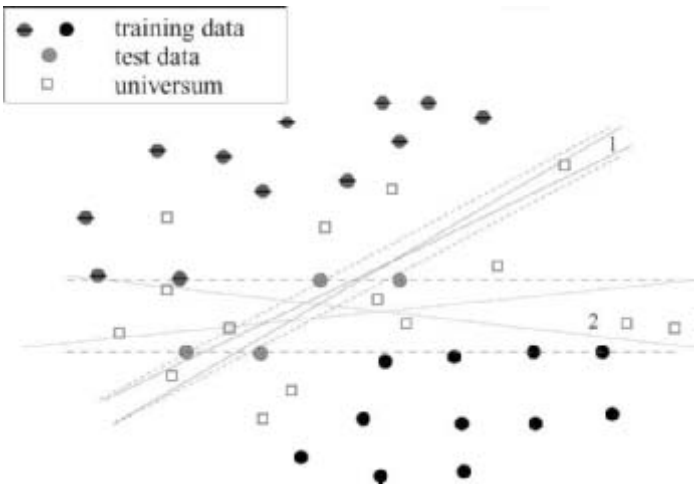

Figure 24.2 The largest number of contradictions on the universum defines the largest equivalence class.

Let us calculate the size $\mu(F_r)$ of an equivalence class F_r by the number t_r of contradictions that the universum has on this class (cf. figure 24.2).

The recommendation of SRM for such a structure would be:
To classify test vectors (24.2) choose the equivalence class (defined on the working set (24.3)) that classifies the training data (24.1) well and has the largest number of contradictions on the universum (24.34).

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

24.9 Algorithms for Inductive and Transductive SVMs

One can translate the discussions of inductive and transductive methods of inference into the following SVM algorithms. In SVM algorithms one first maps input vectors x into vectors z of Hilbert space Z obtaining the images of the training data and test data:

$$(y_1, z_1), ..., (y_\ell, z_\ell), \tag{24.35}$$

$$z_{\ell+1}, ..., z_{\ell+k}, \tag{24.36}$$

and then constructs the optimal separating hyperplane in the feature (Hilbert) space.

24.9.1 SVMs for Inductive Inference

Given the images (24.35) of the training data (24.1) construct the large-margin linear decision rule (Vapnik, 1995):

$$I(x) = \theta[(w, z) + b],$$

where the vector w and threshold b are the solution of the following convex quadratic optimization problem: Minimize the functional

$$R(w) = (w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i), \quad C_1 \geq 0 \tag{24.37}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, ..., \ell \tag{24.38}$$

(defined by the images of the training data (24.35)) where we have denoted

$$\theta(\xi_i) = \begin{cases} 1, & \text{if } \xi_i > 0 \\ 0, & \text{if } \xi_i = 0 \end{cases} .$$

24.9.2 SVMs for Inductive Inference Using the Universum

Given the images (24.35) of the training data (24.1), images (24.36) of the test data (24.2), and the images

$$z_1^*, ..., z_u^* \tag{24.39}$$

of universum (24.34), construct the linear decision rule

$$I(x) = \theta[(w, z) + b],$$

where the vector w and threshold b are the solution of the following convex quadratic optimization problem: Minimize the functional

$$R(w) = (w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{s=1}^{u} \theta(\xi_s^*), \quad C_1, C_2 \geq 0 \tag{24.40}$$

subject to the constraints

$$y_i((z_i, w) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, ..., \ell \tag{24.41}$$

(defined by the images of the training data (24.35)) and the constraints

$$|(z_s^*, w) + b| \leq a + \xi_s^*, \quad \xi_s^* \geq 0, \ s = 1, ..., u, \quad a \geq 0 \tag{24.42}$$

defined by the images (24.39) of the universum (24.34).

24.9.3 SVM for Large-Margin Transductive Inference

Given the images (24.35) of the training data (24.1) and the images (24.36) of the test data (24.2) construct the large-margin linear decision rule for transductive inference,

$$I(x) = \theta[(w, z) + b],$$

where the vector w and threshold b are the solution of the following optimization problem: Minimize the functional

$$R(w) = (w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j), \quad C_1, C_2 \geq 0 \tag{24.43}$$

subject to the constraints

$$y_i((z_i, w) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, ..., \ell \tag{24.44}$$

(defined by the images (24.35) of the training data (24.1)) and the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0, \ j = \ell + 1, ..., \ell + k \tag{24.45}$$

(defined by the images (24.36) of the test data (24.2)) and its desired classifictions $y_{\ell+1}^*, ..., y_{\ell+k}^*$.

One more constraint. To avoid unbalanced solution, Capelle and Zien, following ideas of Thorsten Joachims, suggested the following constraint (Chapelle and Zien, 2005)):

$$\frac{1}{u} \sum_{j=\ell+1}^{\ell+k} ((w, z_j) + b) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i. \tag{24.46}$$

This constraint requires that the test data have about the same proportion of vectors from the two classes as was observed for the training data.

24.9.4 SVM for Transductive Inference Based on Contradictions on the Universum

Given the images (24.35) of the training data (24.1), the images (24.36) of the test data (24.2), and the images (24.39) of the universum (24.34), construct the linear decision rule

$$I(x) = \theta[(w, z) + b],$$

where the vector w and threshold b are the solution of the following optimization problem: Minimize the functional

$$R(w) = (w,w)+C_1\sum_{i=1}^{\ell}\theta(\xi_i)+C_2\sum_{j=\ell+1}^{\ell+k}\theta(\xi_j)+C_3\sum_{s=1}^{u}\theta(\xi_s^*),\quad C_1,C_2,C_3\geq 0 \tag{24.47}$$

subject to the constraints

$$y_i((z_i,w)+b)\geq 1-\xi_i,\quad \xi_i\geq 0,\ i=1,...,\ell \tag{24.48}$$

(defined by the images of the training data (24.35)), the constraints

$$y_j^*((z_j,w)+b)\geq 1-\xi_j,\quad \xi_j\geq 0,\ j=\ell+1,...,\ell+k \tag{24.49}$$

(defined by the images of the test data (24.36)) and its desired classification, and the constraints

$$|(z_s^*,w)+b|\leq a+\xi_s^*,\quad \xi_s^*\geq 0,\ s=1,...,v,\quad a\geq 0 \tag{24.50}$$

(defined by the images (24.39) of the universum (24.34)).

24.9.5 Standard Implementation of the SVM Algorithms

To simplify the optimization problems of the described algorithms the step function $\theta(\xi)$ was replaced by the linear function ξ in the objective functionals (24.37), (24.40), (24.43), and (24.47). Therefore the following algorithms were obtained (Vapnik, 1995, 1998):

- *Large-margin inductive SVM:*

Minimize the functional

$$R(w) = (w,w)+C_1\sum_{i=1}^{\ell}\xi_i,\quad C_1\geq 0 \tag{24.51}$$

subject to the constraints (24.38).

- *Large-margin inductive SVM with the universum*

Minimize the functional

$$R(w) = (w,w)+C_1\sum_{i=1}^{\ell}\xi_i,+C_2\sum_{s=1}^{u}\xi_s^*,\quad C_1\geq 0 \tag{24.52}$$

subject to the constraints (24.41) and (24.42).

- *Large-margin transductive SVM:*

Minimize the functional

$$R(w) = (w,w)+C_1\sum_{i=1}^{\ell}\xi_i+C_2\sum_{j=\ell+1}^{\ell+u}\xi_j,\quad C_1,C_2\geq 0 \tag{24.53}$$

Table 24.1 Test errors of SVMs trained without and with universum.

# of train. examples.	250	500	1,000	2,000	3,000
Test Err. SVM (%)	2.83	1.92	1.37	0.99	0.83
Test Err. SVM+U_1 (%)	2.43	1.58	1.11	0.75	0.63
Test Err. SVM+U_2 (%)	1.51	1.12	0.89	0.68	0.60
Test Err. SVM+U_3 (%)	1.33	0.89	0.72	0.60	0.58

subject to the constraints (24.44) and (24.45). One can also use hint (24.46).

■ *Maximal contradictions on the universum transductive SVM:*
Minimize the functional

$$R(w) = (w, w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{j=\ell+1}^{\ell+k} \xi_j + C_3 \sum_{s=1}^{u} \xi_s^*, \quad C_1, C_2, C_3 \geq 0 \quad (24.54)$$

subject to the constraints (24.48), (24.49), (24.50). One can also use hint (24.46).

24.9.6 Experiments with the Universum

In the summer of 2005, R. Collobert and J. Weston conducted the first experiments on training SVMs with a universum. They demonstrated that

a. SVMs plus a universum can significantly improve performance even in the inductive mode ($C_2 = 0$ in inequality (24.54));

b. for small training sets it is very important how the universum is constructed. For large sets it is less important.

Using the NIST database they discriminated digit 8 from digit 5 using a conventional SVM and an SVM trained with three different universum environments. Table 24.1 shows for different sizes of training data the performance of conventional SVMs and the performance of SVMs trained using universums U_1, U_2, U_3. In all cases the parameter $a = .01$, and the parameters C_1, C_2 and the parameter of Gaussian kernel were tuned using the tenfold cross-validation technique.

For these experiments three different universums (each containing 5000 examples) were constructed as follows:

■ U_1 Select random digits from the other classes (0,1,2,3,4,6,7,9).

■ U_2 Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then for each pixel of the artificial image choosing with probability 1/2 the corresponding pixel from the image 5 or from the image 8.

■ U_3 Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then constructing the mean of these two digits.

24.10 Semi-Supervised Learning

Analyzing the problem of semi-supervised learning as the minimization of the functional (24.4) using training data (24.1) and unlabeled data (24.2), one has to create a clear *statistical* model that allows one to show where (from the formal point of view) one can expect to get an advantage using unlabeled data.

From the theoretical point of view such a statistical model for semi-supervised learning would be more sophisticated than the one described for transductive inference, since it would require additional reasoning in the style of the symmetrization lemma.

Also, from the point of view of possible mechanisms for generalization it looks restricted: Density is defined in input space. It does not depend on the mapping from the input space to the feature space. As we mentioned earlier, a nonlinear mapping can create a large anisotropy in feature space. Using the universum one can take into account this anisotropy, evaluating the size of equivalence classes and using it for classification. If the results obtained in the digit recognition experiments are more or less general, transductive inference can have a more interesting structure than just taking into account density properties.

Therefore it might be a good idea to consider the semi-supervised model as a particular transductive model described at the beginning of this chapter. As such, one first chooses the best equivalence class to perform transductive inference and then chooses from this equivalence class some function which one uses to classify new data that do not belong to the working set.

Such a position allows one to concentrate on the core problem of extracting additional information from the unlabeled data in order to classify them.

24.11 Conclusion: Transductive Inference and the New Problems of Inference

There are two reasons to consider the transductive mode of inference as we have described it above. The first reason is that it is an extremely useful tool for practical applications (see Weston et al. (2003b) and chapter 6).

24.11.1 Adaptation to the Test Data

Transductive inference also contains elements of *adaptation* to new data which we did not discuss since it is not easy to formalize. Back in the 1970s in the very first article devoted to the application of transductive inference (Vapnik and Sterin, 1977), we used data from one medical clinic to classify patients from another clinic. Transduction significantly improved performance.

Another example would be zip code recognition where transductive inference suggests *simultaneously recognizing* all digits of a zip code in contrast to recognizing every digit separately as in inductive inference. It is easy to imagine a situation

where given the training data and an unknown zip code the recognition of any fixed digit of a zip code depends on recognition of the rest of the digits of the zip code. That is, the rule is constructed for the specific zip code. For another zip code one constructs another rule (which might reflect the adaptation to different handwriting). One can find many such examples.

24.12 Beyond Transduction: Selective Inference

The second reason for considering transductive inference is that it forms the simplest model of noninductive inference. These inferences are based on the same general model as inductive inference: the SRM principle. The theory of transduction describes (in the framework of the SRM principle) the mechanisms that provide the advantage of transductive inference over inductive inference.

There also exist models of inferences that go beyond transduction. In particular, *selective inference*:

Given ℓ training examples,

$$(x_i, y_1), ..., (x_\ell, y_\ell), \tag{24.55}$$

and u candidate vectors,

$$x_{\ell+1}, ...x_{\ell+u}, \tag{24.56}$$

select among the u candidates the k vectors with the highest probability of belonging to the first class. Examples of selective inference include:

- *Discovery of bioactive drugs:* Given a training set (24.55) of bioactive and non-bioactive drugs, select from the u candidates (24.56) the k representatives with the highest probability of belonging to the bioactive group.
- *National security:* Given training set (24.55) of terrorists and nonterrorists, select from the u candidates (24.56) the k representatives with the highest probability of belonging to the terrorist group.

Note that selective inference requires a less demanding solution than transductive inference: it does not require classification of the most difficult (border) cases.

Selective inference is the basis for solving high-dimensional decision-making problems. To analyze the selective inference problem one can use the same SRM principle but with a different concept of equivalence classes.

24.12.1 Transductive Inference and the Imperative for Inference in a Complex World

Lastly, the philosophy of transductive inference reflects the general imperative for inference in a complex (high-dimensional) world (Vapnik, 1995), which in fact defines an advantage of the predictive learning models (machine learning techniques) with respect to the generative learning models (classical statistics techniques) (cf. section 1.2.4):

Solving a problem of interest, do not solve a more general (and therefore worse-posed) problem as an intermediate step. Try to get the answer that you really need but not a more general one.

- Do not estimate a density if you need to estimate a function.
(Do not use classical generative models; use ML predictive models.)
- Do not estimate a function if you need to estimate values at given points.
(Try to perform transduction, not induction.)
- Do not estimate predictive values if your goal is to act well.
(A good strategy of action can rely just on good selective inference.)