Chapter 2
Supervised Learning

Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany

Abstract Supervised learning accounts for a lot of research activity in machine learning and many supervised learning techniques have found application in the processing of multimedia content. The defining characteristic of supervised learning is the availability of annotated training data. The name invokes the idea of a 'supervisor' that instructs the learning system on the labels to associate with training examples. Typically these labels are class labels in classification problems. Supervised learning algorithms *induce* models from these training data and these models can be used to classify other unlabelled data. In this chapter we ground or analysis of supervised learning on the theory of risk minimization. We provide an overview of support vector machines and nearest neighbour classifiers – probably the two most popular supervised learning techniques employed in multimedia research.

2.1 Introduction

Supervised learning entails learning a mapping between a set of *input* variables $\mathscr{X}$ and an *output* variable $\mathscr{Y}$ and applying this mapping to predict the outputs for unseen data. Supervised learning is the most important methodology in machine learning and it also has a central importance in the processing of multimedia data. In this chapter we focus on kernel-based approaches to supervised learning. We review support vector machines which represent the dominant supervised learning technology these days – particularly in the processing of multimedia data. We also review nearest neighbour classifiers which can (loosely speaking) be considered

Pádraig Cunningham
University College Dublin, Dublin, Ireland, e-mail: padraig.cunningham@ucd.ie

Matthieu Cord
LIP6, UPMC, Paris, France, e-mail: matthieu.cord@lip6.fr

Sarah Jane Delany
Dublin Institute of Technology, Dublin, Ireland, e-mail: sarahjane.delany@comp.dit.ie

a kernel-based strategy. Nearest neighbour techniques are popular in multimedia because the emphasis on similarity is appropriate for multimedia data where a rich array of similarity assessment techniques is available.

To complete this review of supervised learning we also discuss the ensemble idea, an important strategy for increasing the stability and accuracy of a classifier whereby a single classifier is replaced by a *committee* of classifiers.

The chapter begins with a summary of the principles of statistical learning theory as this offers a general framework to analyze learning algorithms and provides useful tools for solving real world applications. We present basic notions and theorems of statistical learning before presenting some algorithms.

2.2 Introduction to Statistical Learning

2.2.1 Risk Minimization

In the supervised learning paradigm, the goal is to infer a function $f : \mathscr{X} \rightarrow \mathscr{Y}$, the classifier, from a sample data or training set $\mathscr{A}_n$ composed of pairs of (input, output) points, $\mathbf{x}_i$ belonging to some feature set $\mathscr{X}$, and $y_i \in \mathscr{Y}$:

$$\mathscr{A}_n = ((\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)) \in (\mathscr{X} \times \mathscr{Y})^n.$$

Typically $\mathscr{X} \subset \mathbb{R}^d$, and $y_i \in \mathbb{R}$ for regression problems, and y_i is discrete for classification problems. We will often use examples with $y_i \in \{-1, +1\}$ for binary classification.

In the statistical learning framework, the first fundamental hypothesis is that the training data are independently and identically generated from an unknown but fixed joint probability distribution function $P(\mathbf{x}, y)$. The goal of the learning is to find a function f attempting to model the dependency encoded in $P(\mathbf{x}, y)$ between the input $\mathbf{x}$ and the output y. $\mathscr{H}$ will denote the set of functions where the solution is sought: $f \in \mathscr{H}$.

The second fundamental concept is the notion of error or *loss* to measure the agreement between the prediction $f(\mathbf{x})$ and the desired output y. A loss (or *cost*) function $L : \mathscr{Y} \times \mathscr{Y} \rightarrow \mathbb{R}^+$ is introduced to evaluate this error. The choice of the loss function $L(f(\mathbf{x}), y)$ depends on the learning problem being solved. Loss functions are classified according to their regularity or singularity properties and according to their ability to produce convex or non-convex criteria for optimization.

In the case of pattern recognition, where $\mathscr{Y} = \{-1, +1\}$, a common choice for L is the misclassification error:

$$L(f(\mathbf{x}), y)) = \frac{1}{2}|f(\mathbf{x}) - y|.$$

This cost is singular and symmetric. Practical algorithmic considerations may bias the choice of L. For instance, singular functions may be selected for their ability to provide *sparse* solutions. For unsupervised learning developed in Chap. 3.6, the