10

Model Selection

Given an input space $\mathcal{X}$, an output space $\mathcal{Y}$, a class $\mathcal{H}$ of functions $h\colon \mathcal{X} \to \mathcal{Y}$, a loss function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, and data $S = \{(x_i, y_i)\}_{i=1}^n$, it is natural to attempt to find a suitable $h \in \mathcal{H}$ by solving the **Empirical Risk Minimization** (ERM) problem

$$\begin{aligned} &\text{minimize} \quad \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i) \\ &\text{subject to} \quad h \in \mathcal{H}. \end{aligned} \tag{A}$$

This is an example of a **constrained optimization problem**. The function to be minimized is the **objective function** and a solution $\hat{h}_n$ of (A) is called a **minimizer**. Several problems can arise related to this minimization problem.

1. The problem (A) may be hard to solve. This can be the case if the class $\mathcal{H}$ is large, the number of samples n is large, or when the objective function is not differentiable or not even continuous.

2. Do we even want to solve (A)? If the class $\mathcal{H}$ is large we may find a minimizer $\hat{h}_n$ that fits the data well but does not *generalize* well. Conversely, if $\mathcal{H}$ is small then we may be able to find a near-optimal h in $\mathcal{H}$, but the class $\mathcal{H}$ may not be powerful enough to approximate the Bayes classifer.

Since a certain generalization error is unavoidable, we can often replace (A) with a surrogate that is computationally easier to handle and provides a solution that is close enough to the one we are looking for. The choice of such an approximation is also informed by the choice of $\mathcal{H}$. We therefore first study the problem of finding a suitable class $\mathcal{H}$, also known as **model section**.

Model Selection

Consider now the set of inputs again as a set of pairs of random variables $S = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, so that $\hat{h}_n$ is a random variable. Ideally we would like

the generalization risk $R(\hat{h}_n)$ to be close to the Bayes risk $R(h^*)$, which is the *best possible* generalization risk. Recall the decomposition

$$R(\hat{h}_n) = R(h^*) + \underbrace{R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h)}_{\text{Estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} R(h) - R(h^*)}_{\text{Approximation error}} \tag{10.1}$$

of the excess risk. It is thus tempting to choose a class $\mathcal{H}$ that is so large that it contains the Bayes classifier, getting rid of the approximation error. Such a class $\mathcal{H}$, however, is likely to have *high capacity*, i.e., a large VC dimension or Rademacher complexity. For example, if the VC dimension of $\mathcal{H}$ is d, then we obtained the VC bound

$$R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \le 2\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right| \le 2\left(\sqrt{\frac{2d\log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}\right)$$

with probability $1-\delta$. This bound, on its own, would encourage choosing a class $\mathcal{H}$ with a very small VC dimension. One way to address this issue is to consider a nested family of sets $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ of increasing complexity, and optimize not only over $h \in \mathcal{H}_k$, but also over k, taking into account both the approximation and the estimation error. We illustrate this using an example.

Example 10.1. Let $T \subset \mathbb{R}^2$ be a disk and let (X, Y) be a pair of random variables on $\mathbb{R}^2 \times \{0, 1\}$ such that

$$f_T(x) = \mathbb{E}[Y|X = x] = \begin{cases} 1 & x \in T \\ 0 & x \notin T \end{cases}$$

Consider the unit loss function, so that $R(h) = \mathbb{P}\{h(X) \neq Y\}$ for some function $h\colon \mathbb{R}^2 \to \{0, 1\}$. The function f_T is the *Bayes classifier*, as it satisfies $R(f_T) = R^* = 0$. For any hypothesis set $\mathcal{H}$ we can combine (10.1) and the bound (**??**) to get

$$R(\hat{h}_S) \le \underbrace{2\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}(h)\right|}_{\text{Bound on stimation error}} + \underbrace{\inf_{h\in\mathcal{H}} R(h)}_{\text{Approximation error}} .$$

Let $\mathcal{H}_k$ denote the set of indicator functions of regions bounded bys convex polygons with at most k sides (see Figure 10.1). To bound the estimation error, we use the fact that the VC dimension of the class of convex k-gons is $2k + 1$ (see Problem Set 4), and get the bound

$$\sup_{h\in\mathcal{H}_k} |\hat{R}(h) - R(h)| \le \sqrt{\frac{(4k+2)\log(n)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{10.2}$$

with probability $1-\delta$ (we simplified the logarithmic term in the first part of the bound). For the approximation error, we look at the well-known problem of approximating

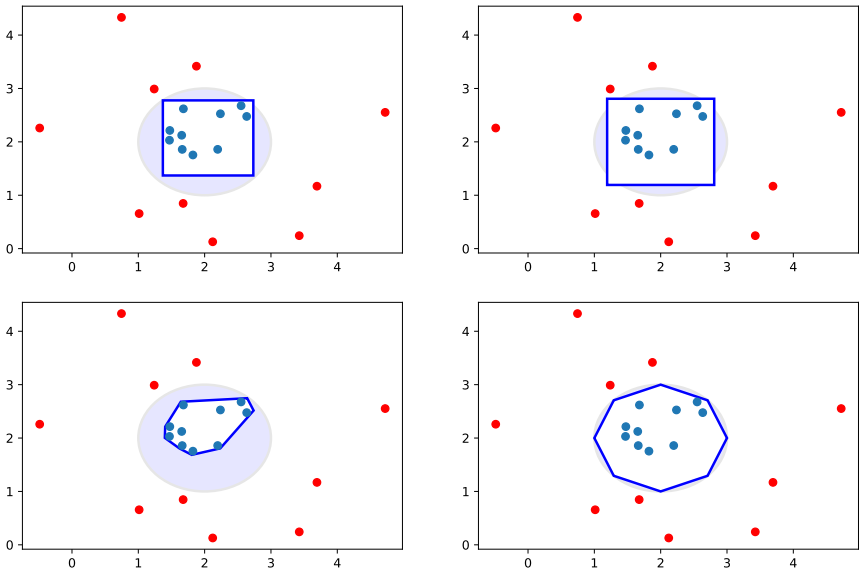

Figure 10.1: Learning a circle with polygons. The left panel shows the *estimation error* when trying to learn the shape using polygons with at most 4 and with at most 8 sides from the data. This is the error typically incurred by empirical risk minimization. The right panel illustrates the *approximation error*. This error measures how good we can approximate the ground truth with our function class.

the circle with a regular polygon. The area enclosed by a regular k-gon inscribed in a circle of radius r is $r^2(k/2)\sin(2\pi/k)$, so the area of the complement in the disk is

$$\pi r^2 - r^2(k/2)\sin(2\pi/k) = O(k^{-2}), \tag{10.3}$$

where the equality follows from the Taylor expansion of the sine function. If the underlying probability distribution on $\mathbb{R}^2$ is the uniform distribution on a larger set or can be approximated as such, then (10.3) gives an upper bound for the approximation error (it can be less), and combined with (10.2) illustrates the estimation-approximation trade-off.

The larger the number of sides of the polygons, the smaller the approximation error becomes, but the estimation error can become large due to *overfitting*. Thus even if the unknown shape we want to learn is a circle, if the number of samples is small we may be better off restricting to simpler models! This also has the additional advantage of saving computational cost.

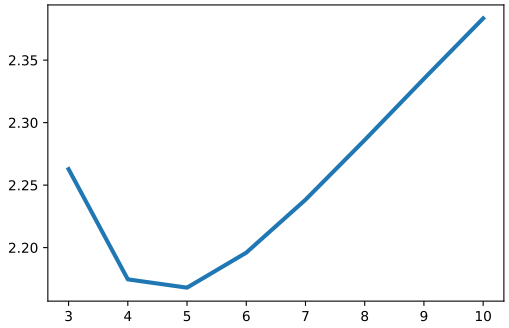

Figure 10.2: Upper bound on the generalization risk that combines the approximation and the estimation error

Structural Risk Minimization and Crossvalidation

One method of dealing with the approximation-estimation trade-off in a principled way is **structural risk minimization** (SRM). In this model, the parameter k enters into the optimization problem to be solved. Assume a sequence of sets of classifiers $\mathcal{H}_k \subset \mathcal{H}_{k+1}$, $k \in \mathbb{N}$, of increasing complexity. Instead of considering the empirical risk $\hat{R}(h)$, we consider, for each class $\mathcal{H}_k$ and $h \in mH_k$, a function $F_k(h)$. This function will be the empirical risk plus some **regularization term**:

$$F_k(h) = \hat{R}(h) + \mathcal{R}_n(\mathcal{H}_k) + \sqrt{\frac{\log k}{n}},$$

where $\mathcal{R}_n(\mathcal{H}_k)$ denotes the Rademacher complexity of $\mathcal{H}_k$. A **structural risk minimizer** (SRM) is then a minimizer of this function over k and h:

$$\hat{h}_n \in \underset{k\in\mathbb{N},\ h\in\mathcal{H}_k}{\arg\min}\ F_k(h).$$

Based on this definition, one can prove bounds on the generalization risk $R(\hat{h}_n)$ of the structural risk minimizer. One drawback of this method is the assumption that the set $\mathcal{H}$ can be expressed as a union of subsets of increasing complexity. An alternative, more practical approach is **crossvalidation**. In this approach, the training set S is subdivided into a smaller training set and a *validation set*. In a nutshell, the ERM problem is solved for different parameters k on the training set, and the one that performs best on the validation set is chosen.

Notes

The problem of model selection is one of the fundamental problems of statistics. The underlying philosophical principle is Occam's razor, or the law of parsimony, which

states that among models with equal explanatory power, one should chose the simplest one. Structural risk minimization was introduced by Vapnik and Chervonenkis. Our treatment of the subject is based on [1, Chapter 4].

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2012.