Chapter 4

Bias Variance: Yuille, Coughlan, Kersten, Schrater

Statistical learning theory is a large subject. It deals not only with probability estimation but also with learning functional relationships between variables (regression), and perhaps most importantly, discrimination and classification. It has a large overlap with classical probability theory and statistics, neural networks, and parts of computer science. In this chapter we will only need some basic results from statistical learning theory and will concentrate mainly on learning probability distributions.

A fundamental problem of statistical learning is to determine how much data is required to learn distributions (for example) to a sufficient degree of accuracy. How many example, for example, are needed to learn a histogram? In this chapter we will state some results on this topic based on theoretical results on statistical estimation from a set of large samples. The basis for these theoretical results, in particular large deviation theory, will be discussed in the next chapter. We add that much of the advanced theoretical material on learning, such as Vapnik-Chervonekis theory, is also based on large deviation theory results (we will briefly introduce VC theory in a section of this chapter but refer to Vapnik for further details).

An important related problem in learning theory is the so-called "curse of dimensionality". The difficultly arises when we seek to learn probability distributions on spaces with high dimensions. The problem is that the amount of data required may grow exponentially with the dimensionality of the data. So while learning a probability distribution on one variable may require a limited amount of data but learning a probability distribution in ten dimensions may require an exorbitant amount of data. This issue will keep arising throughout the chapter. It motivates a variety of techniques, such as dimensionality reduction, where one seeks to represent the data in a lower dimensional space without losing any information.

A useful distinction to keep in mind is the difference between *generalization* and *memorization*. In statistical learning theory[**?**] generalization is the ability to predict the behaviour of novel data after training the system on a dataset. It contrasts with memo-

rization which is simply the ability to describe the dataset well. Generalization suggests the ability to understand the hidden structure of the data rather than simply mimicking its form. An intelligent parrot is able to memorize sentences[1] but, so far, none have shown the ability to produce novel realistic sentences (i.e. to generalize).

Our strategy is to first describe the particular learning applications that we are most relevant to this book while illustrating the points made above. At the end of the chapter we will abstract out the key points and show how they are generalized within the statistical learning theory framework. We stress that there are several important differences between the way that humans, or animals, may learn and statistical learning theories.

Firstly, statistical learning problems are always formulated so that it is clear exactly what the task is (e.g. classify signs on paper as handwritten numbers). By contrast, an important aspect of biological learning, from the perspective of the experimental subject, may involve learning (understanding) what the task is (see next subsection for an elaboration on this issue).

Secondly, in biological vision there is the notion of whether certain "learned" skills can be *transfered*. For example, if the subject learns to discriminate between two signals in one part of the visual field, then can the subject transfer this ability and discriminate between the same signals in a different part of the visual field? Or will the observer still be able to discriminate between them if they are rotated? (These questions are of interest because they throw light on where in the visual system the learning takes place. If learning is very specific to position in the visual field then this suggests that it may be performed early in the visual system and perhaps in V1.) Such questions are rarely studied in the statistical learning community because, for an artificial system, it is often very straightforward to build such transference directly into the system[2]. Rather confusingly, the term "transfer" is sometimes called "generalization" in the biological learning community but, as described above, "generalization" has a rather different, and mathematically precise, definition in statistical learning. To avoid confusion we will only use "generalization" in the statistical learning sense and use "transference" for biological vision.

Thirdly, any biological learning system is far more complex than the systems currently studied in statistical learning theory and involves some phenomena which have only partially been explored in statistical learning. For example, there is some evidence that biological systems learn better when they are started off with simple examples. Although there have been some attempts in the statistical learning community to build in the use of "hints" there is, as yet, no consensus about how to model this. Another effect is

[1]My favourite example is the parrot who was found lost in London. When picked up the parrot said "I'm Pete and my phone number is 348-1780. Have you got that, mate? Please call my boss.".

[2]Neural network learning models which build this tranferance into the network, either explicitly or by choice of representation, appear to be far more effective than neural networks which hope that transference will arise as an emergent property given enough data samples. See LeCun for details.

that biological systems often show the "Eureka effect" where learning suddenly improves. Such effects have been found in statistical learning (Sombolinsky) and theories of one-shot learning have been discussed by Valiant but, again, there is no full consensus on this topic. There is also strong evidence that much of early biological learning takes places in an orchestrated fashion whereby certain visual skills are learnt during critical time frames (for example, binocular stereo appears to develop between the period of 4-12 weeks in infants). There is no direct analog of this in statistical learning theory at present.

4.1 Learning as Empirical Risk Minimization

The basic formulation of statistical learning theory assumes that the data is generated by an (unknown) probability distribution $P(z)$. We are given a set of functions to account for the data parameterized by α. Depending on the task (learning probabilities, classification, or regression) we determine a loss function $L(z, \alpha)$ which describe how well our model fits the data z with parameter value α. We then seek α to minimize the total risk function:

$$R(\alpha) = \int L(z, \alpha) P(z) dz. \tag{4.1}$$

So far, this is standard decision theory as described in Chapter 2. The main difference is that *we do not know the distribution* $P(z)$. Instead we have a set of samples $\{z_i\}$ which are, hopefully, representative of $P(z)$.

We can then determine the *empirical risk* which is

$$R_{emp}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} L(z_i, \alpha). \tag{4.2}$$

It can be shown (Glivenko-Cantelli Theorem) that in the limit as the number of samples tends to infinity the minimum of the empirical risk $R_{emp}(\alpha)$ will tend to the minimum of the true risk $R(\alpha)$. For mathematical afficianados, this convergence is "in probability" which is weaker than "almost surely". A major question of learning theory is precisely how quickly does this convergence occur. This obviously determines the number of samples required in order to learn. A key concept here is the *Vapnik-Chervonenkis (VC) dimension* which is a measure of the capacity of the function. The larger the VC dimension then the more the function can represent but, conversely, the more data is required for convergence. We will return to this issue later in the chapter.

Observe that this definition does not make explicit use of a teacher. If a teacher is available then the each data element z_i can be divided into an input $\vec{x}_i$ and an output y_i, where the teacher specifies y_i. If no teacher is available then the data is simply an input $\vec{x}_i$.

For classification, we write $R(\alpha) = \int L(y, \phi(x, \alpha)) P(y, x) dy dx$ where $L(y, \phi) = 0$, if $y = \phi$ and 1 otherwise. Here $\phi(x, \alpha)$ is a decision rule that assigns category $y =$

$\phi(x, \alpha)$ to data x. (Both y and ϕ can take k distinct values). The empirical risk is $R_{emp}(\alpha) = (1/N) \sum_{i=1}^{N} L(y_i, \phi(x, \alpha))$, where the input is a random independent sample of pairs $\{(y_i, x_i) : i = 1, ..., N$ is given.

For regression, we have $R(\alpha) = \int (y - f(x, \alpha))^2 P(y, x) dy dx$. The empirical risk function is $R_{emp}(\alpha) = (1/N) \sum_{i=1}^{N} (y_i - f(x_i, \alpha))^2$ where $\{y_i, x_i : i = 1, ..., N\}$ is the sample data.

For probability estimation, the risk function is usually set to be:

$$R(\alpha) = -\int P(x) dx \log P(x|\alpha). \tag{4.3}$$

As above, we obtain an empirical risk function $R_{emp}(\alpha) = -(1/N) \sum_{i=1}^{N} \log P(x_i|\alpha)$ where the training data is $\{x_i : i = 1, ..., N\}$. As we will show, such standard procedures as estimating the mean and variance of a set of data samples can be formulated as special cases of this framework.

Another way to think about learning probability distributions is in terms of minimizing the Kullback-Leibler divergence between the true distribution $P(x)$ and a model distribution $P(x|\alpha)$. More precisely, minimizing the risk, see equation (4.3), is equivalent to selecting $\alpha^* = \arg\min_\alpha D(P(x)||P(x|\alpha))$ (exercise for reader). Moreover, minimizing the *empirical risk* can also be re-expressed in terms of selecting $\alpha^*_{emp} = \arg\min_\alpha D(P_{emp}(x)||P(x|\alpha))$ where $P_{emp}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$ where $\{x_i : i = 1, ..., N\}$ are the set of data samples. (In other words, $P_{emp}(x)$ is the empirical probability distribution).

We now describe the bias variance dilemma for regression. This illustrates a key aspect of learning theory: the trade-off between memorizing the samples and generalizing to predict the behaviour of novel samples. (These trade-offs apply for all forms of learning but are particularly simple for regression, as we will see, because of the nature of the loss function).

For regression the risk is $R(\alpha) = \int (y - f(x, \alpha))^2 P(y, x) dy dx$. We can write the loss function per sample x as $R(\alpha : x) = \int (y - f(x, \alpha))^2 P(y|x) dy$ (i.e. $R(\alpha) = \int R(\alpha; x) P(x) dx$).

Now suppose that we estimate α from a set of data-samples $D = \{(x_i, y_i)\}$. Whatever our choice of estimator (e.g. MAP) *the estimate α_D will be a function $\alpha(\{(x_i, y_i)\})$ of the samples D and hence is a random variable* with distribution $P(\alpha_D)$ determined from the distribution $P(x, y)$ and the form of the estimator. The detailed form of $P(\alpha_D)$ are unimportant for the following argument. (In a later section of this chapter we will see that $P(\alpha_D)$ will tend to a Gaussian distribution as the number of samples used to estimate it becomes large. In the full limit it will becomes a Dirac delta function).

We now ask: what is the expected risk for predicting sample x averaged over our estimated α_D? More precisely, we are concerned with

$$\int R(\alpha_D; x) P(\alpha_D) d\alpha = \int \int (y - f(x, \alpha))^2 P(y|x) dy P(\alpha_D) d\alpha_D. \tag{4.4}$$

We can re-express the right-hand-side as a sum of three positive terms which can be interpreted in terms of three types of error. To do this we define $\hat{y}(x) = \int dyyP(y|x)$. We then express:

$$(y-f(x,\alpha))^2 = (y-\hat{y}(x)+\hat{y}(x)-f(x,\alpha))^2 = (y-\hat{y}(x))^2+(\hat{y}(x)-f(x,\alpha))^2+2(y-\hat{y}(x))(\hat{y}(x)-f(x,\alpha)). \tag{4.5}$$

Substituting into the equation for $\int R(\alpha_D;x)P(\alpha_D)d\alpha$ gives:

$$\int R(\alpha_D;x)P(\alpha_D)d\alpha = \int (y-\hat{y}(x))^2P(y|x)dy + \int (\hat{y}(x)-f(x,\alpha))^2P(\alpha_D)d\alpha_D, \tag{4.6}$$

where the cross terms have dropped out (because $\int P(y|x)dy\{y-\hat{y}(x)\} = 0$) and we have integrated out α_D and y for the first and second terms on the right hand side respectively.

We now perform a similar procedure on the second term on the right hand side by defining $\tilde{f}(x) = \int f(x:\alpha_D)P(\alpha_D)d\alpha_D$ and re-expressing:

$$\begin{aligned}(\hat{y}(x)-f(x,\alpha))^2 = (\hat{y}(x)-\tilde{f}(x)+\tilde{f}(x)-f(x,\alpha))^2 = (\hat{y}(x)-\tilde{f}(x))^2 \\ +(\tilde{f}(x)-f(x,\alpha))^2 + 2(\hat{y}(x)-\tilde{f}(x))(\tilde{f}(x)-f(x,\alpha)).\end{aligned} \tag{4.7}$$

We substitute this into the expression for $\int R(\alpha_D;x)P(\alpha_D)d\alpha$ and obtain (using $\int\{f(x:\alpha_D)-\tilde{f}(x)\}P(\alpha_D)d\alpha_D = 0$):

$$\int R(\alpha_D;x)P(\alpha_D)d\alpha = \int (y-\hat{y}(x))^2P(y|x)dy+\int (\tilde{f}(x)-f(x,\alpha))^2P(\alpha_D)d\alpha_D+(\hat{y}(x)-\tilde{f}(x))^2. \tag{4.8}$$

So the expected risk has three non-negative components. The first component, $\int(y-\hat{y}(x))^2P(y|x)dy$, occurs because it is impossible to predict y from x exactly because y is a random variable with distribution $P(y|x)$. The second component, $\int(\tilde{f}(x)-f(x,\alpha))^2P(\alpha_D)d\alpha_D$, is the *variance* and is the contribution to the expected risk caused by the fluctuations in the estimation of α_D from data samples. The third term, $(\hat{y}(x)-\tilde{f}(x))^2$, is the *bias* and is the error arising from whether the set of function $f(x,\alpha)$ can approximate the best estimate $\hat{y}(x)$.

This illustrates the so-called bias variance dilemma. There is a trade-off between the two terms which depends on the set of functions $f(x,\alpha)$ chosen to estimate y from x (and also on the estimator chosen to estimate α_D). By restricting the set of functions to be very small then we can ensure that the variance is small – but the bias may be enormous because it may be impossible to estimate y from x accurately using such a function. Conversely, if we pick a large set of functions $f(x,\alpha)$ then we can achieve small bias but have large variance. The problem is that incorrect models (functions) lead to large

biases while truly model-free learning suffers from large variance. Model-free approaches require large amounts of data and are slow to converge, due to high variance. But using models to reduce the variance means that we increase the risk of bias. This, of course, is the standard trade-off in learning theory and emphasizes yet again that how well we can approximate the true underlying function depends strongly on how much data we have available. The VC-dimension theory give a framework for this.