Statistical Theories of Model Fitting

B. D. Ripley

Department of Statistics, University of Oxford

Abstract. Statistics has long been concerned with fitting and choosing good models for data. In this paper we discuss the statistical theory pertinent to the problem of generalization, and the reasons perceived by statisticians why they have not already solved the problem. This includes statistical decision theory, Bayes factors, probability scoring, shrinkage methods, model-averaging, C_p, AIC and allied indices, and the use of performance bounds.

The assessment of generalization performance is also discussed, an important task much neglected in the neural networks and machine learning literature.

Statistics has been concerned with the fitting of models for centuries, and several competing theories have been developed over at least decades. Have these solved the problem of generalization? Some statisticians certainly think they have a good a solution as is possible. We will explore the parts of the argument most relevant to the machine learning and neural network community.

The most pertinent reference for much of this material is Ripley (1996, Chapter 2); general modern references on statistical decision theory include Robert (1994) and Bernardo and Smith (1994).

1 Explanation *vs* Prediction

Most of classical applied statistics is concerned with finding the true explanation (*scientific inference*); if this can be done the selected model will generalize well. Here parsimony is a key concept. Much of theoretical statistics is concerned with *decision making*, in essence predicting the unknown state of the current problem (and that includes mainstream Bayesian statistics). Most people agree that the two approaches are different and that both have their place: there is no agreement about their relative importance.

Other fields have different emphases, but Donald Michie has always stressed the need for explanation in machine learning/intelligence. In many fields of application one has to explain the concept before it will be accepted. Contemplate the problems of medical diagnosis, ZIP code recognition, control of a guided missile and credit scoring to reflect on the different emphases.

Consider the problem of linear regression, and suppose we have lots of possible regressors. Should we

- Include them all in the regression?

– Select a subset that we are convinced has a real effect?
– Include them all, but down-weight those that are less convincing?

The answer depends on the purpose (and possibly also on the computational resources). If the purpose is explanation we should probably use the second, but use the third for prediction. Linear regression has traditionally be taught from the viewpoint of explanation, which reflects its importance in that role in scientific and medical research.

The relative importance of explanation and prediction is not always clear-cut in human decision-making. I heard the following anecdote from Professor Jim Alty, then of the University of Strathclyde:

> JA was visiting Cairo, and took a taxi. He noticed that the taxi driver did not stop at red traffic lights.
> Later, JA noticed that the driver *did* sometimes stop at red traffic lights, and that these traffic lights were also manned by policeman. This spawned a new hypothesis:
>
> > 'Taxi drivers in Cairo only stop at those traffic lights which show red and are manned by a policeman.'
>
> Later still, the taxi went past a red traffic light which *was* manned by a policeman. Now JA can not fathom out the rule, and asks the driver for an explanation.
>
> > 'Ah, but that is obvious. . . . .'

The explanation (. . .) is given at the end of the paper.

This was given as an example of a task for a machine learning system. It shows some interesting features:

(i) It includes 'learning by query'. Human learning allows us to ask for explanations, to ask for the labels of examples we construct (a classic paradigm of science) and to ask for counter-examples to current hypotheses.
(ii) The induced rule has considerable *generalization*—to all taxi drivers from one, from a few traffic lights to all, for example.
(iii) Explanation is important—a 'black box' solution such as a neural network would be unacceptable.
(iv) A precise rule may need unobserved information.

We do need to be careful that the theory we apply actually solves the real-life problem, and not a subtly different abstraction of it.

2 Supervised Classification

For simplicity and brevity we will concentrate on supervised classification. Suppose we have a problem in which we are asked to identify a pattern as belonging to one of K classes; use C to denote the unknown true class. Our decision must be to report one of the classes or $\mathcal{D}$ for *don't know* or *'doubt'*. We may also want or allow the *outlier* decision $\mathcal{O}$ *'none of these'*. Non-negative losses $L(k,l)$ are incurred if decision l is taken when class k is actually true. A common loss specification is

$$L(k,l) = \begin{cases} 0 & \text{if } l = k \text{ (correct decision)}, \\ 1 & \text{if } l \neq k \text{ and } l \in \{1,\dots,K\} \\ & \quad\quad \text{(wrong decision)}, \\ 0 < d < 1 & \text{if } l = \mathcal{D} \text{ (being in doubt)}, \end{cases} \tag{1}$$

which penalizes all mistakes equally, but penalizes indecision rather less.

Finally, we are given a measurement (the *features*) $\boldsymbol{x} \in \mathcal{X}$ on which to base the decision.

The goal is to minimize the *risk* R, the expected loss for a future example, for a *classifier* $c : \mathcal{X} \to \{1,\dots,K,\mathcal{D}\}$. The minimum achievable is the *Bayes risk*. We have the following solution.

Proposition 1. *The classification rule which minimizes the total risk under loss* (1) *is*

$$c(\boldsymbol{x}) = \begin{cases} k & \text{if } p(k \,|\, \boldsymbol{x}) = \max_{l \leqslant K} p(l \,|\, \boldsymbol{x}) \\ & \quad\quad \text{and this exceeds } 1-d, \\ \mathcal{D} & \text{if each } p(k \,|\, \boldsymbol{x}) \leqslant 1-d, \end{cases}$$

and for a general loss function is

$$c(\boldsymbol{x}) = \begin{cases} k & \text{if this attains } \min_{l \leqslant K} \sum_j L(j,l) p(j \,|\, \boldsymbol{x}) < d, \\ \mathcal{D} & \text{otherwise.} \end{cases}$$

Proof. We have

$$R(\widehat{c}) = \mathsf{E}\big[\mathsf{E}[L(C,\widehat{c}(\boldsymbol{x})) \,|\, X]\big] = \int_{\mathcal{X}} \mathsf{E}[L(C,\widehat{c}(\boldsymbol{x})) \,|\, X = \boldsymbol{x}]\, p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

where $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$ is the marginal density for X. It suffices to minimize the conditional expectation, which we can write as $\sum_{k=1}^{K} L(k,c)\, p(k \,|\, \boldsymbol{x})$, with respect to c, for each $\boldsymbol{x}$.

Also, any admissible classifier (one which cannot be beaten for every sensible loss function L by another classifier) is a Bayes rule for some prevalences (π_k) of the classes.

Consequences We have an apparent separation of tasks:

1. Find the posterior probabilities $p(c \,|\, \boldsymbol{x})$.
2. Use these with the current loss function L_{ij} to classify.

and in particular 'training' is separated from decision making, unlike empirical risk minimization. This can be useful where losses are case-dependent, for example depend on a patient's age and physical condition.

Often it is only important to know $p(\cdot \mid \boldsymbol{x})$ in certain regions ($\lesseqgtr 0.5$ for a two-class problem with equal losses).

Note that we have not yet mentioned outliers, sometimes known as the problem of *novelty detection*. The only sound way to approach this seems to be to treat outliers as a separate class $\mathcal{O}$, in which case $p_{\mathcal{O}}(\boldsymbol{x})$ needs to be known.

The problem with using proposition 1 is that the posterior probabilities $p(c \mid \boldsymbol{x})$ are known only in very unusual problems (some do occur in genetic counselling and stock breeding). Usually, we have to estimate the posterior probabilities from examples. How we do so is the key to good generalization, since if we can find good estimates, proposition 1 suggests that we will have a classifier that is close to optimal.

2.1 The predictive approach

Suppose we are given a training set $\mathcal{T}$ of examples with their correct classifications, and a family $p(x, c; \theta)$ of possible joint probabilities. Suppose we have a probability distribution $p(\theta)$ over the family. The risk $R(c, \theta)$ now depends on θ; we allow c to depend on $\mathcal{T}$ and consider the goal of minimizing the overall risk

$$R(C) = \mathsf{E}\, L\big(C, c(X, \mathcal{T})\big) = \int R(C, \theta) p(\theta)\, \mathrm{d}\theta$$

Then we can still find the optimal classifier by

Proposition 2. *The classifier that minimizes the overall risk under loss* (1) *is*

$$\widetilde{c}(\boldsymbol{x}) = \begin{cases} k & \textit{if } \widetilde{p}(k \mid \boldsymbol{x}) = \max_l \widetilde{p}(l|\boldsymbol{x}) \\ & \qquad \textit{and this exceeds } 1 - d, \\ \mathcal{D} & \textit{if each } \widetilde{p}(k \mid x) \leqslant 1 - d, \end{cases}$$

where $\widetilde{p}(k \mid \boldsymbol{x}) = \mathsf{P}\{C = k \mid X = \boldsymbol{x}; \mathcal{T}\}$. *This is also the rule that minimizes the conditional risk given the training data. The extension to other loss functions follows Proposition 1.*

Proof. We condition on the training set $\mathcal{T}$:

$$R(\widehat{c}) = \mathsf{E} L(C, \widehat{c}(X, \mathcal{T})) = \mathsf{E}_{\mathcal{T}} \mathsf{E}\big[L(C, \widehat{c}(X, \mathcal{T})) \bigm| \mathcal{T}\big]$$

and the conditional expectation is the total risk in the conditional distribution. Now apply proposition 1 to the conditional total risk.

Remarks:

- The predictive probability $\widetilde{p}$ can be manipulated by the usual rules as it is just a conditional distribution. In terms of the parameters θ, the predictive probabilities are given by

$$\widetilde{p}(c \mid \boldsymbol{x}) = \int p(c \mid \boldsymbol{x}; \theta) p(\theta \mid \mathcal{T}, \boldsymbol{x}) \, \mathrm{d}\theta$$

- This proposition has been known as folklore for a long time: the approach was advocated by Geisser and by Aitchison & Dunsmore in the early and late 1960's. (See Aitchison and Dunsmore, 1975 and Geisser, 1993 for later book-length accounts by those authors.)
- The distribution $p(\theta)$ need not be a prior belief, but it needs to be a reasonable weighting.
- The distribution $p(\theta)$ must be *proper*, that is have total mass one. 'Vague' priors with infinite mass are not allowed.
- θ can be very general; it can index probability models as well as parameters and 'weights'. Thus the predictive approach mandates averaging over the predictions by different models, weighting the models by the *Bayes factor* $p(\mathcal{T} \mid \text{model})$.

It is reasonable to accept that proposition 2 is *the* answer to supervised classification (and it has analogues for other decision problems). The difference from most of current statistical practice is that it takes the uncertainty in θ into account. (Unlike the much more common *'plug-in'* approaches, which use proposition 1 after estimating the probabilities, usually after estimating θ by maximum likelihood and using $p(c \mid \boldsymbol{x}; \widehat{\theta})$.)

Why is the predictive approach not more widely used?

- It lacks an eloquent evangelist.
- It *can* be hard to find an acceptable $p(\theta)$.
- It may impose a large computational load to compute $\widetilde{p}(c \mid \boldsymbol{x})$.

We now explore some of the insights it can give to model fitting in machine learning and neural networks.

2.2 Model Averaging

Let us write out the details if θ incorporates several models. Suppose we have m models M_i. Then

$$\begin{aligned}\widetilde{p}(c \mid \boldsymbol{x}) &= \sum_i \mathsf{P}(c \mid \boldsymbol{x}, \mathcal{T}, M_i) \mathsf{P}(M_i \mid \boldsymbol{x}, \mathcal{T}) \\ &\propto \sum_i \mathsf{P}(c \mid \boldsymbol{x}, \mathcal{T}, M_i) p(\mathcal{T}, \boldsymbol{x} \mid M_i) \mathsf{P}(M_i)\end{aligned}$$

We can recover the constant, since $\sum_c \widetilde{p}(c \mid \boldsymbol{x}) \equiv 1$.

Geisser (1993, §4.1) argues that choosing one model is acceptable only if the loss function incorporates a principle of extreme parsimony: only one model may be used. (Remember that the implicit goal is good prediction, not explanation.)

We usually assume that the distribution of $\boldsymbol{x}$ does not depend on the model or its parameters. Suppose model M_i has parameters ψ. Then we have

$$\widetilde{p}(c \mid \boldsymbol{x}) \propto \sum_i \mathsf{P}(M_i) p(\mathcal{T} \mid M_i) \int p_i(c \mid \boldsymbol{x}; \psi) p_i(\psi \mid \mathcal{T}) \, \mathrm{d}\psi$$

where

$$p(\mathcal{T} \mid M_i) = \int \ell_i(\mathcal{T}; \psi) p_i(\psi) \, \mathrm{d}\psi, \qquad \ell_i(\mathcal{T}; \psi) = \prod_{j \in \mathcal{T}} p_i((\boldsymbol{x}_j, c_j) \mid \psi)$$

where $\ell_i(\mathcal{T}; \psi)$ is the likelihood assuming model M_i.

Note that $p(\mathcal{T} \mid M_i)$ is in general a probability density and not a probability, so only ratios of $p_i(\mathcal{T} \mid M_i)$ for different models i are properly defined: such ratios are known as *Bayes factors*.

Another viewpoint Suppose we believe that one of the models M_i is true. Then our belief that it is model i after seeing the training set $\mathcal{T}$ is

$$\mathsf{P}(M_i \mid \mathcal{T}) \propto p(\mathcal{T} \mid M_i) \mathsf{P}(M_i)$$

If we know how we would act if model i were true, the theory say that we should hedge our bets by averaging across actions, weighting our average by the current beliefs in the truth of each model. This has been used to much effect in time-series forecasting by Harrison, Stevens and West (see, for example, West and Harrison, 1997).

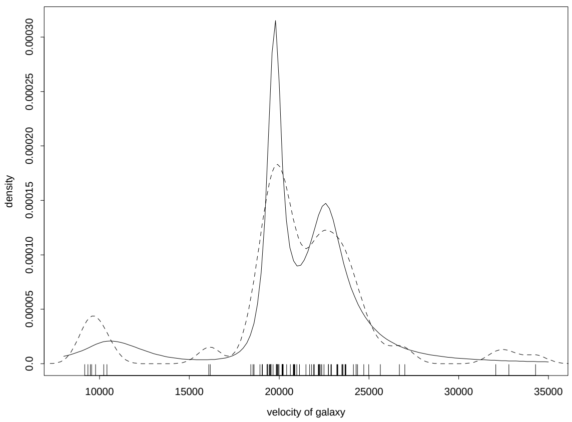
There has been much recent interest in incorporating 'model uncertainty' into statistical analyses. For example, Draper (1995, p. 48) has a graph of predictions of oil prices for 1981–90 made in 1980. The analysts were all confident, differed considerably from each other, and were all way off! Almost all of the uncertainty is in the 'correct' model for oil price movements, and the analysts' model all seem to be incorrect as prices went down when all the analysts predicted them to rise.

Note that there is an implicit assumption in predictive theory that the model $p(x, c; \theta)$ is adequate, so for some value of θ it matches closely the true mechanism for producing examples. This seems not to be the case for the oil-price prediction study, and is perhaps the biggest potential drawback in applications of the predictive approach. Using models which are universal approximators can help guard against this, but there will remain many other assumptions (such that the examples in the training set are correctly labelled).

Priors The prior *must be proper*. Draper (1995) says

> In small sample situations with vague prior information about the parameters, care must be taken, if improper priors are used, to avoid undefined constants [in approximation (10)]; see, for example, Spiegelhalter and Smith (1982) for an approach to solving this problem. An alternative solution would involve the use of proper but relatively uninformative priors whose specification is guided by preposterior analysis.

This is not at all respected in the neural networks literature, and may have serious consequences.



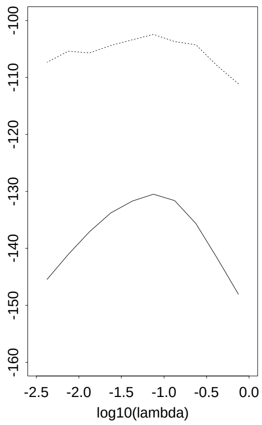


Fig. 1. (left) Logspline (solid line: Kooperberg and Stone, 1992) and kernel density (dashed: Simonoff, 1996) estimates for the data on the velocities of 82 galaxies. The bandwidth of the kernel estimate was chosen by the Sheather and Jones (1991) method.

Fig. 2. (Right) Log Bayes factor for λ (solid) and log predictive scores (dashed). Natural logs were used for the y axis ($e^{10} \approx 22,000$).

Figure 1 shows data used by Roeder (1990) on the velocities of 82 galaxies (of 83 in the original data set). The substantive scientific question in the original investigation is how many galaxy clusters these represent. Richardson and Green (1997); Stephens (1997) and others earlier (including Roeder) have considered a mixture of normals for this dataset. How many? Richardson and Green use a vague prior for the means and an informative one for the variances. Matthew Stephens (Stephens, 1997) has shown

$$p(\mathcal{T} \mid k \text{ normals}) \propto \frac{n!\,k!}{(n+k-1)!}$$

for $n = 82$ examples in $\mathcal{T}$, otherwise independent of the dataset, in the limit of a vague prior. This is an enormous bias to a single cluster caused by the (mis-)specification of $p(\theta)$.

With a more realistic prior, Stephens (1997) demonstrates that model choice is possible but tricky. One issue is that the assumption of normality for each cluster is not a good one, and a longer-tailed distribution gives clearer results. Once again, if our specification of $p(c \mid \boldsymbol{x}; \theta)$ is not adequate even the predictive approach cannot give us good estimates of the posterior probabilities.

A neural network example Figure 2 shows an example from Mark Mathieson (see Mathieson, 1998). A neural network with 3 classes was fitted to 300 data points and the issue is how to average over the amount of weight decay, λ. The solid line shows approximate Bayes factors computed by a method described later. This is fairly convincing evidence for $\lambda \approx 0.1$, but even here some model averaging would be useful. The dashed line validates this conclusion: see later.

2.3 Analogies with linear regression

How do these ideas apply to linear regression? Suppose we have a vector $\boldsymbol{x}$ of regressors to predict a random variable Y, and we use squared error as our measure of loss. There are analogous theorems in this setting.

Suppose we have n observations on $p \geqslant 3$ variables. Then the p-dimensional mean is not admissible under squared-error loss (Stein, 1956). It can be beaten by a shrinkage estimator, which has a bias but a smaller variance.

In *subset selection* we explore all subsets of size $k \leqslant p$. Conventionally one is chosen after a (partial) search of the space of models. However, the predictive approach says we should average over models. Since regression is linear, $\widetilde{\mathsf{E}}[y \mid \boldsymbol{x}] = \widehat{\beta}_i^T \boldsymbol{x}$ for model i where some elements of $\hat{\beta}_i$ are constrained to be zero. (We have averaged over the uncertainty in the coefficients; their posterior mean is the usual estimate.) Thus

$$\widetilde{\mathsf{E}}[y \mid \boldsymbol{x}] = \widehat{\beta}^T \boldsymbol{x} \quad \text{where} \quad \beta = \sum_i \beta_i \mathsf{P}(M_i \mid \mathcal{T})$$

The posterior probability may be spread over many models: Moulton (1991) reports an example with 12 regressors in which the top 800 models out of $2^{12} = 4096$ are needed to account for 90% of the posterior probability.

Ridge regression was originally (Hoerl and Kennard, 1970a,b) designed to overcome collinearity; it adds a penalty proportional to the squared length of β. (Thus it is a very special case of weight decay.) This is also an approximate Bayesian procedure with a Gaussian prior on β.

Both model averaging and ridge regression down-weight regressors whose effect is uncertain. This is appropriate when the goal is prediction not explanation. The classic approach in applied statistics to control over-fitting and hence ensure reasonable generalization has been subset selection (and more generally model selection), aiming at a model with good properties of explanation. This is usually good for prediction, but it can often be bettered by down-weighting methods.

2.4 Balancing datasets

Suppose that we have two classes and loss matrix

$$\begin{bmatrix} 0 & \ell \\ 1 & 0 \end{bmatrix}$$

that is, declaring 1 as 2 is ℓ times as expensive as declaring 2 as 1. Then the optimal rule is to declare class 2 if $\ell\,\widetilde{p}(1 \mid \boldsymbol{x}) > \widetilde{p}(2 \mid \boldsymbol{x})$, or

$$\widetilde{p}(2 \mid \boldsymbol{x}) > 1/(\ell + 1) = p_{\text{crit}}$$

So we need to know $\widetilde{p}(2 \mid \boldsymbol{x})$ accurately around p_{crit} (only). For some 'plug-in' models (including logistic regression) we will do this better by weighted training. Suppose we increase the relative proportion of class 2 by ℓ: this would change p_c to 0.5. However, it is better to do this by down-weighting class 1 cases by this factor in the likelihood.

Strictly speaking, balancing is not advantageous in the predictive approach, although it will be for many of the approximations we discuss. The plug-in approach tends to estimate posterior probabilities $p(c \mid \boldsymbol{x})$ that are too extreme (near zero or near one; see figure 4), which is less harmful after balancing.

3 Non-Bayesian Approaches to Model Selection

People do find ways to choose subsets in linear regression, for example, and this has an extensive literature. The significance-test approach of stepwise selection suffers from multiple testing and will only work for nested models (although there is some work of Cox which embeds non-nested models into a larger family). There has been a long series of suggested criteria such as Mallows' C_p, FPE, PRESS, $\ldots$, which aim to estimate the future performance and minimize the estimate. (Structural risk minimization (Vapnik, 1995) aims to bound the future performance, and to minimize the bound.)

The major player is Akaike's AIC ('An Information Criterion', named in Akaike, 1974 from a 1973 idea). This takes

$$-2 \log \ell(\mathcal{T}; \widehat{\theta}) + a \times \text{number of parameters}$$

for $a = 2$ and ℓ the likelihood. Many other values of a have been proposed, including $a = 0.5, 1, 1.5$ and $a = \log n$, sometimes called BIC or SBC (after Schwarz (1978), who proposed it for linear regression). Akaike has proposed more than one motivation, but the ideas are famously obscure. My own view follows.

Suppose we consider an independent sample of size n and let q be the number of parameters. We use log-probability scoring (see later) to measure the fit of $p(X;\widehat{\theta})$ to the true density $p(X)$. Thus on the training set we use the deviance $2\sum_i \bigl\{\log p(x_i) - \log p(x_i;\widehat{\theta})\bigr\}$ and similarly on the test set. Then the expected deviance on the test set is $2n$ times the Kullback-Leibler divergence $d(p, p_\theta) = \int p(x) \log p(x)/p(x;\theta)\,\mathrm{d}x$. Then under conditions, the expected difference between the deviance on the test and training sets is approximately $2q$, and the deviance plus $2q$ is an approximately unbiased estimate of the test-set deviance. These conditions are

- The model is true, that is $p = p_\theta$ for some θ,
- the log likelihood is dominated by a single peak, and
- that peak can be approximated by a second-order Taylor series.

So minimizing AIC can be seen as trying to minimize the (expected) deviance on a test set of the same size as the training set. However, the size of the test set is not relevant, as the expected deviance is proportional the size.

Stone (1979), Moody (1991, 1992), Murata *et al.* (1991, 1993, 1994) considered removing the first condition (which is silly in model selection, as not all models could be true). They found that we need to replace q by $q^* = \operatorname{trace}\bigl[KJ^{-1}\bigr]$ where

$$J = -\mathsf{E}_p \frac{\partial^2 \log p(X_i;\theta_0)}{\partial\theta\,\partial\theta^T}, \qquad K = \mathsf{Var}_p \frac{\partial \log p(X_i;\theta_0)}{\partial\theta}$$

and θ_0 is the 'least-false' parameter value that minimizes $d(p, p_\theta)$. Murata *et al.*'s NIC is AIC with q replaced by q^*, and Moody's $p_{\text{eff}} = q^*$. Unfortunately,

- the other two conditions are rarely fulfilled, and
- the variability of the differences in AIC is $O_p(1)$ even if the models are nested, and can be $O_p(\sqrt{n})$. (O_p denotes 'order in probability' and measures the size of the random fluctuations.)

We may hope that AIC will pick a model which is at least big enough to pick out the salient features, and not too much too large. It does seem to do so in time-series analysis. This bias towards a model that is at least large enough is important for prediction, but may not be satisfactory for explanation.

We return to BIC/SBC, which despite their names are a non-Bayesian approach. If we use Laplace's method (Evans and Swartz, 1995; Ripley, 1996) on the Bayes factor we find

$$\begin{aligned} p(\mathcal{T}\,|\,M) &= \int p(\mathcal{T}\,|\,\theta)\, p(\theta)\,\mathrm{d}\theta = \int \exp -E(\theta)\,\mathrm{d}\theta \qquad (2)\\ &\approx \exp -E(\widehat{\theta}) \int \exp[-\tfrac{1}{2}(\theta-\widehat{\theta})^T V^{-1}(\theta-\widehat{\theta})]\,\mathrm{d}\theta \\ &= \exp -E(\widehat{\theta})\,(2\pi)^{q/2}|V|^{1/2} \end{aligned}$$

where V is the Hessian of $E(\theta) = -\log \ell(\theta; \mathcal{T}) - \log p(\theta)$. Thus

$$\log p(\mathcal{T} \mid M) \approx \log \ell(\widehat{\theta}; \mathcal{T}) + \log p(\widehat{\theta}) + \tfrac{q}{2} \log 2\pi + \tfrac{1}{2} \log |V|$$

If we suppose θ has a prior which we may approximate by $N(\theta_0, V_0)$, we have

$$\log p(\mathcal{T} \mid M) \approx \log \ell(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) - \tfrac{1}{2} \log |V_0| + \tfrac{1}{2} \log |V|$$

and V^{-1} is the sum of V_0^{-1} and the Hessian H of the log-likelihood at $\widehat{\theta}$. Thus

$$\log p(\mathcal{T} \mid M) \approx \log \ell(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) - \tfrac{1}{2} \log |H|.$$

If we assume that the prior is very diffuse we can neglect the second term, so the penalty on the log-likelihood is $-\frac{1}{2} \log |H|$. For a random sample of size n from the assumed model this might be roughly proportional to $-(\frac{1}{2} \log n)\, q$ provided the parameters are all identifiable. This leads to BIC. There are a number of variations in which different terms in the expansion of $\log p(\mathcal{T} \mid M)$ are dropped.

What is n? Why should the total sample size be appropriate for all parameters? (An example is study of election forecasting models, in which different parties have very different sample sizes.) There may be no substitute for trying to do the integration in (2) as accurately as we can.

4 Predictive Fitting of a Neural Network

How can we apply the predictive ideas to a feed-forward neural net (even with fixed number of hidden units and amount of regularization)? The issue is to approximate the integration

$$\widetilde{p}(c \mid x) = \int p(c \mid \boldsymbol{x}; \psi) p(\psi \mid \mathcal{T})\, \mathrm{d}\theta$$

where $p(\psi \mid \mathcal{T})$ is often a very complex multimodal function of ψ. Various approaches have been tried.

1. Markov chain Monte Carlo (MCMC) ideas—but our tests on problems with known symmetries suggest a lack of convergence after a day's CPU time on an Sun UltraSparc 2/200. In any case, it is very hard to demonstrate that an MCMC method has converged enough to give reliable results. MCMC methods may work but they are rarely *proved* to work. Further, they can appear to have converged but in fact have not (Ripley, 1988).
2. Evans and Swartz (1995) give many other approximate methods of Bayesian computation. Many are based on finding the local maxima of $p(\psi \mid \mathcal{T})$.
   (a) Find the local maxima, use Laplace's method to weight the predictions at the modes.

(b) As (a), but simulate values from the mixture of normals implied by the multiple Laplace approximation, and average.
(c) As (b), but use this as an envelope for importance sampling.
(d) Use those samples to re-fit a mixture of normals and iterate ....
(e) Use a local Gibbs sampler within each mode to find its weight and samples for averaging.

3. Fit the net say, 25 times, and average the $p(c \mid \boldsymbol{x}; \widehat{\theta})$ over the $\widehat{\theta}$ found.

Approach 3 is simple, and we have found it hard to beat by a significant margin. However, it does not cover model-averaging.

A testbed procedure in my group (see for example Mathieson, 1998, and used for the Bayes factors in figure 2) is

A. Find as many local maxima of $p(\psi \mid \mathcal{T})$ as we can in a reasonable time. (The proper prior rules out very flat regions.)
B. Identify equivalent local maxima by symmetries.
C. For each local maximum do
 (i) Find the Hessian, explore the curvature in the principal directions.
 (ii) Use this to get a better local quadratic approximation to $p(\psi \mid \mathcal{T})$, and use that to delimit the peak by an ellipsoid or box.
 (iii) Run a Gibbs sampler locally within the peak, updating the coordinate system used during the 'burn-in' phase.
 (iv) Use harmonic averaging to find the relative probability mass to assign to this peak.
 (v) Average predictions over samples from the Gibbs sampler.
D. Check for overlaps in the peak regions, merge if necessary.
E. Take a weighted average of peak-specific predictions.

This is computationally expensive, but is being used to find the best possible answer with which to compare approximations to the full predictive approach. It can also be used to find Bayes factors between separate models and to average across models, as the weights have the correct relativities between models.

Harmonic averaging has been used earlier (Ripley, 1992; DiCiccio *et al.*, 1997). Suppose we sample ψ from $g(\psi) \propto p(\psi \mid \mathcal{T})$. Then

$$\mathsf{E}\frac{s(\psi)}{p(\psi \mid \mathcal{T})} = \int \frac{s(\psi)}{p(\psi \mid \mathcal{T})} g(\psi)\, \mathrm{d}\psi = A \int s(\psi)\, \mathrm{d}\psi, \quad A = 1 \Big/ \int p(\psi \mid \mathcal{T})\, \mathrm{d}\psi$$

If we know $\int s(\psi)\, \mathrm{d}\psi$ (say by choosing s to be a conventional density) we can estimate A by $1/\mathsf{P}(\text{peak} \mid \mathcal{T})$. Thus our estimate is based on

$$\mathsf{P}(\text{peak} \mid \mathcal{T}) = 1 \Big/ \mathsf{E}\Big[\frac{s(\psi)}{p(\psi \mid \mathcal{T})} \;\Big|\; \mathcal{T}\Big]$$

We choose the density s to match $p(\psi \mid \mathcal{T})$ with lighter tails, and estimate the mean by the average over samples.

5 Assessing Generalization

It can be very helpful to have some idea of the *Bayes risk*, the smallest possible expected loss. (Too many people assume that it is zero!) This is a measure of the 'overlap' of the classes.

We obviously need a very flexible classifier. We can do something with k-nearest neighbour rules. A (k, ℓ)-nn rule declares a class if it has at least ℓ representatives amongst the k nearest points in $\mathcal{T}$, otherwise it declares 'doubt' $\mathcal{D}$. We have (risk-averaging is defined later):

Proposition 3. *(Ripley, 1996, pp. 196–7) In the large-sample theory the means of the risk-averaged* $(3, 2)$*-nn rule and the error rate of the* $(2, 2)$*-nn rule are equal and provide a lower bound for the Bayes risk. The risk-averaged estimator has smaller variance.*

The large-sample theory is the limit as the size of the training set grows: this ensures that the k-nn rules are really local and that in particular $p(\cdot \mid \boldsymbol{x})$ does not vary much over the search for k neighbours.

This suggests estimating a lower bound for the Bayes risk by running the 3-nn classifier on the training set and reporting 1/3 the number of occasions on which the neighbours are two of one class, one of another (and of course one of the neighbours will be the training-set example itself). If the distances are tied, we can average over ways of breaking the tie, since this will be equivalent to averaging over perturbations of the points.

In the theory this result does not depend on the metric used to measure the distances; in practice it does. With small samples this method tends to underestimate the lower bound on the Bayes risk (which is the safe direction).

5.1 Uniform Bounds

The theory pioneered by Vapnik and represented by Devroye *et al.* (1996) concerns universal procedures, which will with enough data get arbitrarily close to the Bayes risk. Universal strong consistency (risk converging to the Bayes risk with probability one over sequences of independent identically distributed training sets) can be achieved by k-nn rules for $k = o(n)$ and various other simple rules.

These limit results are proved by proving uniform bounds, and it has been suggested that these be used directly. From a published paper:

> 'If our network can be trained to classify correctly a fraction $1 - (1 - \beta)\epsilon$ of the n training examples, the probability that its error—a measure of its ability to generalize—is less than ϵ is at least $1 - \delta$.'

(authorship suppressed to protect the guilty).

Let pmc denote the true error rate of a classifier, and $\widehat{\mathsf{pmc}}$ the error rate on a training set of size n. The theorems are of the form

$$\mathsf{P}\left\{\widehat{\mathsf{pmc}}(g) < (1 - \beta)\mathsf{pmc}(g) \ \textit{and}\ \mathsf{pmc}(g) > \epsilon \text{ for any } g \in \mathcal{F}\right\} < \delta \qquad (3)$$

for $n > n_0(\epsilon, \delta, \mathcal{F})$ for a class of classifiers $\mathcal{F}$. The probability is over the random choice of training set of size n. This says that however we choose the classifier in the class, our observed error rate on the training set will be close to true error rate of the classifier, *for most training sets*.

The authors of the quote have (like many lawyers and judges) confused $\mathsf{P}(A \text{ and } B)$ with $\mathsf{P}(B \mid A)$, for their quote means

$$\mathsf{P}\big\{\mathsf{pmc}(\widehat{c}) > \epsilon \;\big|\; \widehat{\mathsf{pmc}}(\widehat{c}) < (1-\beta)\mathsf{pmc}(\widehat{c})\big\} < \delta$$

This is only implied by (3) (with a different δ) if we can bound below the probability of the conditioning term. We only have one training set, so how could we possibly do that? Sadly, this is not an isolated example.

These bounds apply to any classifier: we also need results that relate $\mathsf{pmc}(\widehat{c})$ to the Bayes risk. That is harder: for example Devroye *et al.* prove consistency for a projection pursuit regression trained by empirical risk minimization — and we have no idea how to do that in practice.

It *would* be useful to have bounds on the errors in proposition 3.

6 Practical Ways of Performance Assessment

The only really convincing way we have to compare generalization ability is to try out various classifiers. We need a *test* set unused for any other purpose.

- This is an experiment, so it should be designed and analysed. Even specialist statisticians forget this. Cohen (1995), Kleijnen (1987) and Kleijnen and van Groenendaal (1992) discuss these issues in a way that may be accessible to the neural network community.
- Comparisons between classifiers will normally be more precise that assessments of the true risk.
- Standard errors of estimates decrease at $O(n^{-1/2})$ so we often need a very large test set to measure subtle differences in performance. For example, if we count errors, we need differences of at least 5 and often 15 or more for statistical significance (Ripley, 1994; Ripley, 1996, p. 77).

6.1 Validation

The same ideas can be used to choose within classes of classifiers (for example the number of hidden units). This is called *validation* and needs a separate validation set. A validation set can also be used to optimize over the weights in model averaging (suggested as long ago as Stone, 1974).

Cross-validation and cross-testing With small datasets, separate training, validation and test sets seem wasteful. Can we reuse just one set? In V-fold cross validation we divide the data into V roughly equal pieces (how?).

Use one for validation, $V-1$ for training, and rotate. Similarly for testing: if we need both we nest or rotate.

What size should V be? The original proposal (Lunts and Brailovsky, 1967; Lachenbruch, 1967; Stone, 1974) was $V=n$, leave-one-out CV (often confused with the distinct concept of *jackknifing*). This is often too expensive, and also too variable. The V-fold variant was proposed by Toussaint and Donaldson (1970) and more generally by Breiman *et al.* (1984). I usually use $V=5$ or 10.

6.2 How Do We Estimate Performance?

In supervised classification we should be interested in the achieved risk, as that was the objective we set out to minimize. To estimate this on a test set we count errors or add up losses. That can be a rather crude measure with high variance.

For validation we can choose whatever measure we wish. There is a lot to be said for *log-probability scoring*, that is

$$-\textstyle\sum_i \log \widehat{p}(c_i \mid \boldsymbol{x}_i)$$

summed over the test set. This is estimating the crucial part of the Kullback-Leibler divergence $d(p,\widehat{p})$ and has a lower variability than error-counting. We return to the example of figure 2. This used a test set of size 1200, and log-probability scoring showed much more discrimination than the error count. Note that the Bayes factor on the training set and the log-probability score on the test set have a similar but not identical shape. Fortunately they both favour the same range of λ. Note that these curves have different purposes and we would only expect the positions of their modes to be close, not their exact shapes.

Smoothed or risk-averaged estimates Suppose we see an example with features $\boldsymbol{x}$. Then we only observe on the test set one of the possible classes c. Our expected loss given $X=\boldsymbol{x}$ is

$$e(\boldsymbol{x}) = \sum_c L\big(c, \widehat{c}(\boldsymbol{x})\big)\, p(c \mid \boldsymbol{x})$$

Thus if we knew the posterior probabilities, we could estimate the risk by averaging $e(\boldsymbol{x})$ over the test set. As we do not know them, we could use our best estimates to give

$$\widetilde{e}(\boldsymbol{x}) = \sum_c L\big(c, \widehat{c}(\boldsymbol{x})\big)\, \widetilde{p}(c \mid \boldsymbol{x})$$

and the resulting estimate of the risk is known as an *smoothed* or *risk-averaged* estimate. The idea goes back to Glick (1978); for more details see Ripley (1996, pp. 75–6). The loss $L\big(c, \widehat{c}(\boldsymbol{x})\big)$ takes a small set of values; $e(\boldsymbol{x})$ is much less variable (often with variance reduced by a factor of 10). So we get a more precise estimator of risk, and we do not even need to know the true classification!

Where is the catch? $\widetilde{p}(c \mid \boldsymbol{x})$ has to be a good estimate of $p(c \mid \boldsymbol{x})$. Can we check that? That is our next topic.

6.3 Calibration

How will we know if we have estimated the $p(c \mid \boldsymbol{x})$ well? One answer is by the performance of our rule. Another is that the probability forecasts are *well-calibrated*; about 30% of the events we say have probability about 0.3 should happen. We can check this on a validation set (figure 3). Fitting the smooth curve to the 'rug' of ticks on the axes is a non-linear curve-fitting problem, for which one's favourite method could be used and the degree of smoothing is usually chosen by eye.

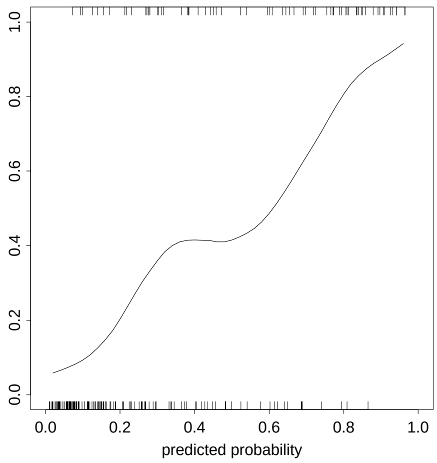


Fig. 3. Calibration plot for a two-class medical diagnosis problem.

More often we see something like figure 4 in which the extreme probabilities (near zero or one) are estimated as too extreme. The reason is that we are not using $\widetilde{p}(k \mid \boldsymbol{x})$ but $\widehat{p}(k, \mid \boldsymbol{x}; \widehat{\theta})$, a *plug-in* estimate in which we assume that the fitted parameters $\widehat{\theta}$ are the true weights (and that the neural network represents the true posterior probabilities). We are ignoring the uncertainty in the $\widehat{\theta}$.

If necessary, we can use such graphs to re-calibrate, but usually they indicate a deeper problem. Calibration plots can help detect over-fitting: an example from Mark Mathieson is shown in figures 5 and 6.

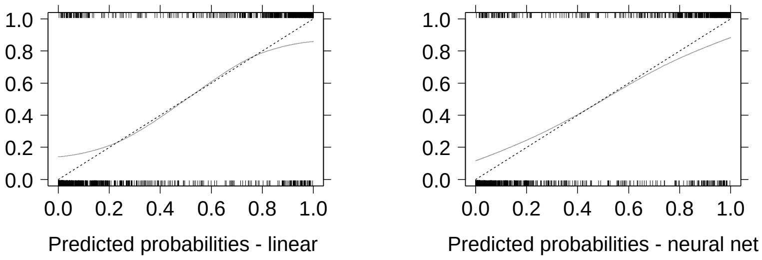


Fig. 4. Calibration plots from a breast cancer prognosis study—the task is to predict years to relapse as one of the classes 0, 1, 2, 3, 4+.

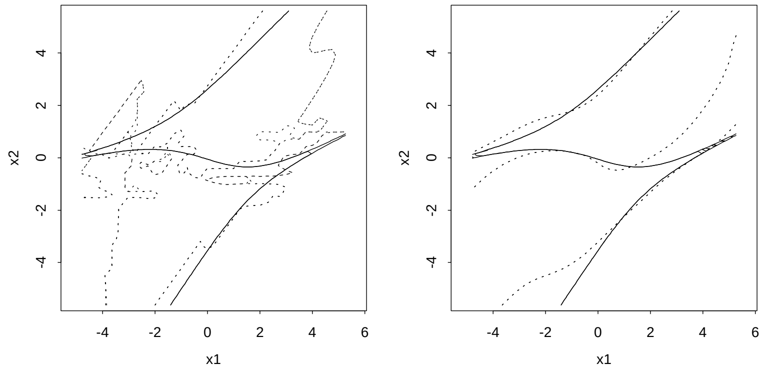


Fig. 5. Three-class fits with the Bayes rule and fits from an ordinal neural network with weight decay parameter λ. $\lambda = 0$ (left) and $\lambda = 0.1$ (right). The true class boundaries are shown as solid lines, the fitted ones as dashed lines.

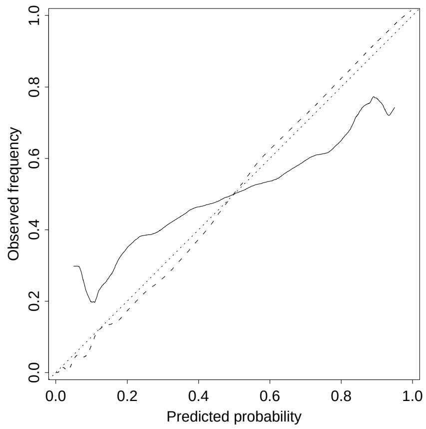


Fig. 6. Calibration plots for the fits in figure 5; the line for $\lambda = 0.1$ is dashed.

There is an extensive literature on probability forecasting: see Dawid (1986) and Cooke (1991). This has usually been for the assessment of human experts, but applies equally in machine learning. There are various aggregate scores that say how well 'experts' are doing on a variety of predictions. *Proper* scoring rules reward 'experts' who truly are expert. The most popular of these is the (half-)Brier score, the sum of squares of one minus the predicted probability of the event that happened, and especially *log-probability scoring*, the sum of minus the natural logarithms of those probabilities.

6.4 Performance Summaries

Suppose we have just two classes, in medical diagnosis normal or diseased. Then class-conditional error rates have conventional names

$$\text{specificity} = \mathsf{P}(\text{predict normal} \mid \text{is normal})$$
$$\text{sensitivity} = \mathsf{P}(\text{predict diseased} \mid \text{is diseased})$$

Considering these separately is usually an admission that losses other than zero-one (1) are appropriate.

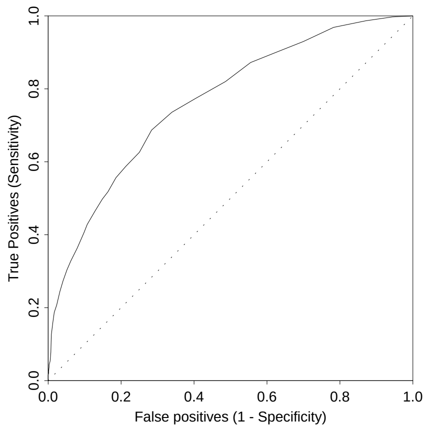


Fig. 7. An ROC curve for the breast-cancer prognosis problem of figure 4.

Suppose our classifier outputs an approximation to $p(\text{diseased} \mid \boldsymbol{x})$. Then we get a family of classifiers by declaring 'diseased' if $p(\text{diseased} \mid \boldsymbol{x}) > \alpha$ and varying $\alpha \in (0, 1)$. The ROC (receiver operating curve) is a plot of sensitivity against $1 - \text{specificity}$ (Hand, 1997, Chapter 7). Figure 7 shows an example ROC. For equal losses we want to minimize the sum of sensitivity and $1 - \text{specificity}$, or $\text{sensitivity} - \text{specificity}$. The best choice of α in that case corresponds to the point at which the ROC has slope 1 (and we know corresponds to $\alpha = 0.5$); in this problem the sensitivity and specificity would be both about 0.8. As the ratio of losses changes the optimal operating point moves along the curve, jumping along any linear sections of the curve.

It is common in medicine to quote the area under the ROC curve. This is a compromise between losses that may be irrelevant to the real problem, where for example high sensitivity may be much more important than high specificity.

Acknowledgements

I am grateful to Mark Mathieson, Ruth Ripley and Matthew Stephens for their comments and for providing examples.

The explanation of the anecdote

'Ah, but that is obvious. He is my brother'

Bibliography

Aitchison, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis.* Cambridge: Cambridge University Press.

Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Bernardo, J. and Smith, A. F. M. (1994) *Bayesian Theory.* Wiley.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Monterey, CA: Wadsworth and Brooks/Cole.

Cohen, P. R. (1995) *Empirical Methods for Artificial Intelligence.* Cambridge, MA: The MIT Press.

Cooke, R. M. (1991) *Experts in Uncertainty. Opinion and Subjective Probability in Science.* New York: Oxford University Press.

Dawid, A. P. (1986) Probability forecasting. In *Encyclopedia of Statistical Sciences* (Eds S. Kotz, N. L. Johnson and C. B. Read), volume 7, pp. 210–218, New York. Wiley.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* New York: Springer.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing bayes factors by posterior simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society series B*, **57**, 45–97.

Evans, M. and Swartz, T. (1995) Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254–272.

Geisser, S. (1993) *Predictive Inference: An Introduction.* New York: Chapman & Hall.

Glick, N. (1978) Additive estimators for probabilities of complete classification. *Pattern Recognition*, **10**, 211–222.

Hand, D. J. (1997) *Construction and Assessment of Classification Rules.* Chichester: Wiley.

Hoerl, A. E. and Kennard, R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b) Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Kleijnen, J. P. C. (1987) *Statistical Tools for Simulation Practitioners.* New York: Marcel Dekker.

Kleijnen, J. P. C. and van Groenendaal, W. (1992) *Simulation: A Statistical Perspective.* Chichester: Wiley.

Kooperberg, C. and Stone, C. J. (1992) Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, **1**, 301–328.

Lachenbruch, P. (1967) An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, **23**, 639–645.

Lunts, A. L. and Brailovsky, V. L. (1967) Evaluation of attributes obtained in statistical decision rules. *Engineering Cybernetics*, **3**, 98–109.

Mathieson, M. J. (1998) *Ordinal Models and Predictive Methods in Pattern Recognition*. D.Phil. dissertation, University of Oxford.

Moody, J. E. (1991) Note on generalization, regularization and architecture selection in nonlinear learning systems. In *First IEEE-SP Workshop on Neural Networks in Signal Processing*, pp. 1–10, Los Alamitos, CA. IEEE Computer Society Press.

Moody, J. E. (1992) The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4. Proceedings of the 1991 Conference* (Eds J. E. Moody, S. J. Hanson and R. P. Lippmann), pp. 847–854, San Mateo, CA. Morgan Kaufmann.

Moulton, B. R. (1991) A Bayesian-approach to regression selection and estimation with application to a price-index for radio services. *Journal of Econometrics*, **49**, 169–193.

Murata, N., Yoshizawa, S. and Amari, S. (1991) A criterion for determining the number of parameters in an artificial neural network model. In *Artificial Neural Networks. Proceedings of ICANN-91* (Eds T. Kohonen, K. Mäkisara, O. Simula and J. Kangas), volume I, pp. 9–14, Amsterdam. North Holland.

Murata, N., Yoshizawa, S. and Amari, S. (1993) Learning curves, model selection and complexity of neural networks. In *Advances in Neural Information Processing Systems 5. Proceedings of the 1992 Conference* (Eds S. J. Hanson, J. D. Cowan and C. L. Giles), pp. 607–614, San Mateo, CA. Morgan Kaufmann.

Murata, N., Yoshizawa, S. and Amari, S. (1994) Network information criterion—determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865–872.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society series B*, **59**, 731–792.

Ripley, B. D. (1988) Uses and abuses of statistical simulation. *Mathematical Programming*, **42**, 53–68.

Ripley, B. D. (1992) Classification and clustering in spatial and image data. In *Analyzing and Modeling Data and Knowledge* (Ed. M. Schader), pp. 93–105, Berlin. Springer.

Ripley, B. D. (1994) Flexible non-linear approaches to classification. In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications* (Eds V. Cherkassky, J. H. Friedman and H. Wechsler), NATO ASI Series F volume 136, pp. 105–126, Berlin. Springer.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Robert, C. P. (1994) *The Bayesian Choice*. Springer.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, **53**, 683–690.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer.

Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society series B*, **44**, 377–387.

Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Ed. J. Neyman), volume 1, pp. 197–206, Berkeley. University of California Press.

Stephens, M. (1997) *Bayesian Methods for Mixtures of Normal Distributions*. D.Phil. dissertation, University of Oxford.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society series B*, **36**, 111–147.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society series B*, **41**, 276–278.

Toussaint, G. and Donaldson, R. (1970) Algorithms for recognizing contour-traced hand-printed characters. *IEEE Transactions on Computers*, **19**, 541–546.

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. Springer.

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. New York: Springer, second edition.