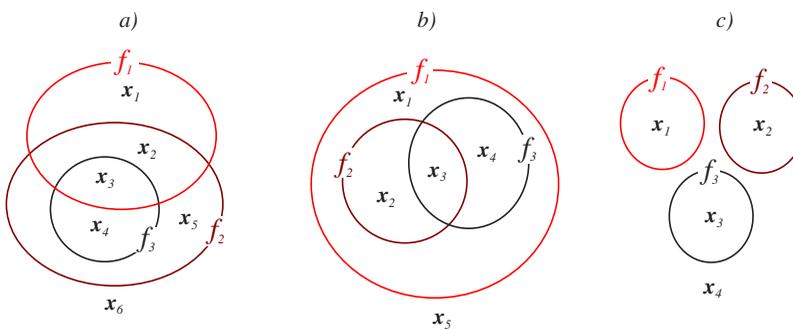
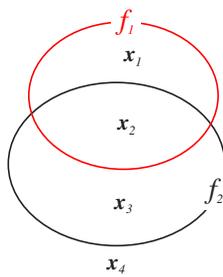


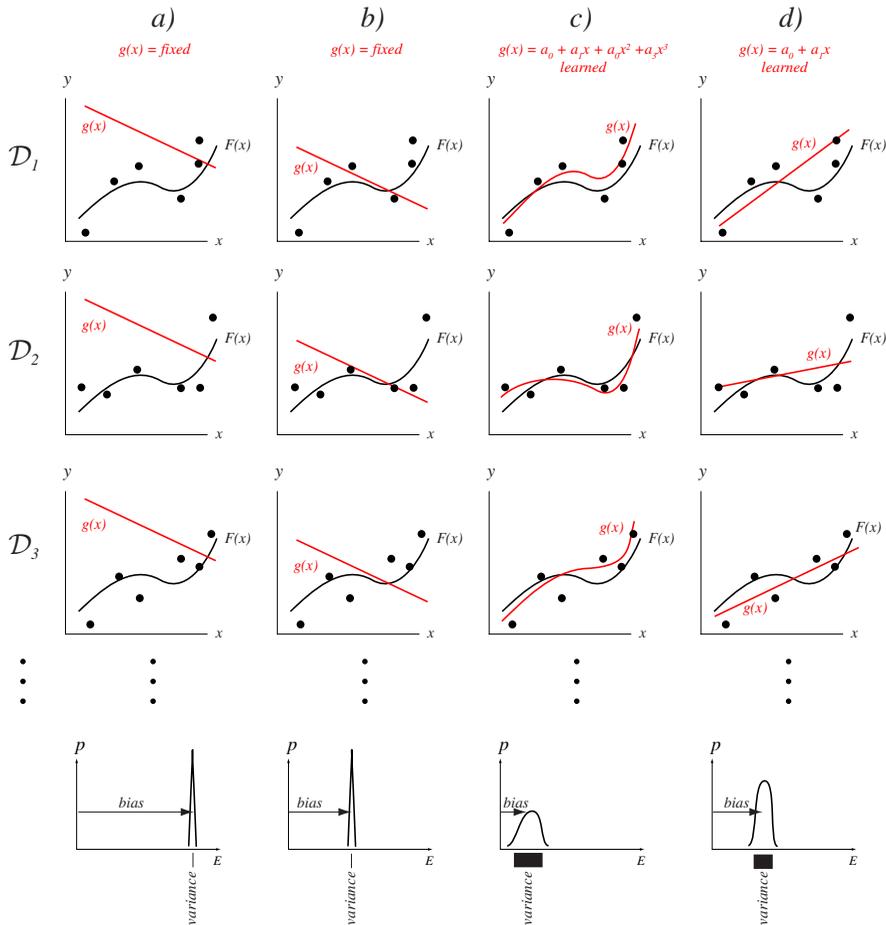
**FIGURE 9.1.** The No Free Lunch Theorem shows the generalization performance on the off-training set data that *can* be achieved (top row) and also shows the performance that *cannot* be achieved (bottom row). Each square represents all possible classification problems consistent with the training data—this is *not* the familiar feature space. A + indicates that the classification algorithm has generalization higher than average, a - indicates lower than average, and a 0 indicates average performance. The size of a symbol indicates the amount by which the performance differs from the average. For instance, part a shows that it is possible for an algorithm to have high accuracy on a small set of problems so long as it has mildly poor performance on all other problems. Likewise, part b shows that it is possible to have excellent performance throughout a large range of problem, but this will be balanced by very poor performance throughout a large range of other problems. It is impossible, however, to have good performance throughout the full range of problems, shown in part d. It is also impossible to have higher-than-average performance on some problems while having average performance everywhere else, shown in part e. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



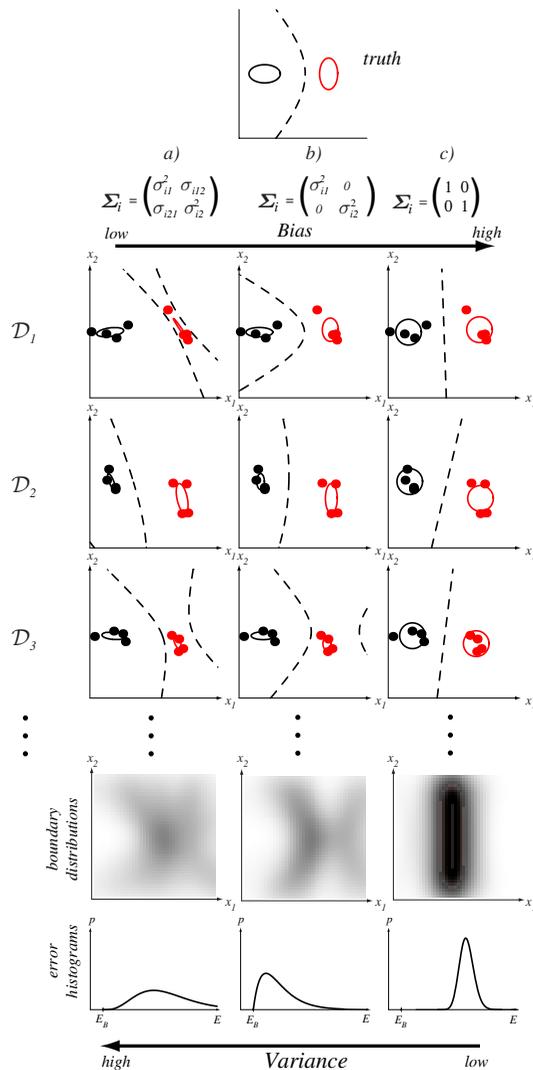
**FIGURE 9.2.** Patterns  $\mathbf{x}_i$ , represented as  $d$ -tuples of binary features  $f_i$ , can be placed in Venn diagram (here  $d = 3$ ); the diagram itself depends upon the classification problem and its constraints. For instance, suppose  $f_1$  is the binary feature attribute `has_legs`,  $f_2$  is `has_right_arm` and  $f_3$  the attribute `has_right_hand`. Thus in part a) a pattern  $\mathbf{x}_1$  denotes a person who has legs but neither arm nor hand;  $\mathbf{x}_2$  a person who has legs and an arm, but no hand; and so on. Notice that the Venn diagram expresses the biological constraints associated with real people: it is impossible for someone to have a right hand but no right arm. Part c) expresses different constraints, such as the biological constraint of mutually exclusive eye colors. Thus attributes  $f_1$ ,  $f_2$  and  $f_3$  might denote `brown`, `green`, and `blue`, respectively, and a pattern  $\mathbf{x}_i$  describes a real person, whom we can assume cannot have eyes that differ in color. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



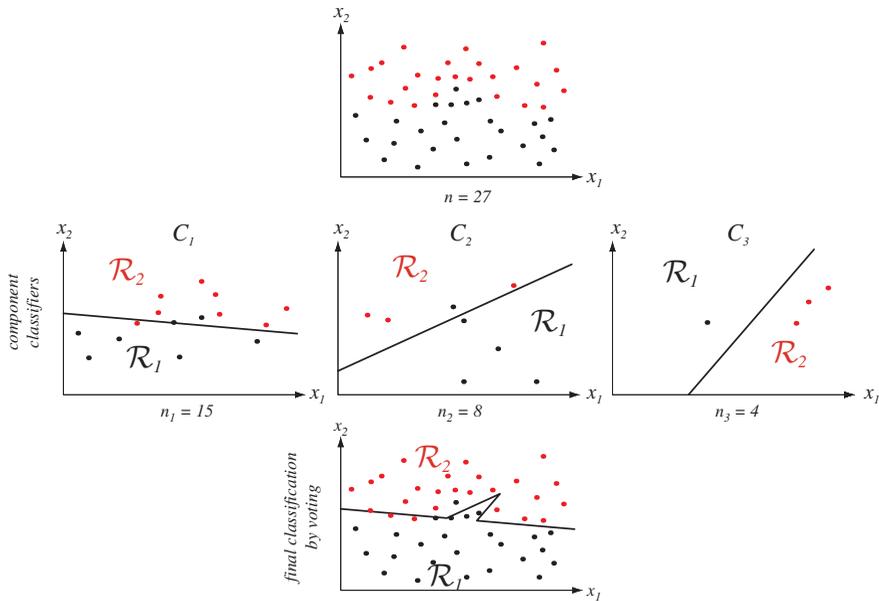
**FIGURE 9.3.** The Venn for a problem with no constraints on two features. Thus all four binary attribute vectors can occur. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



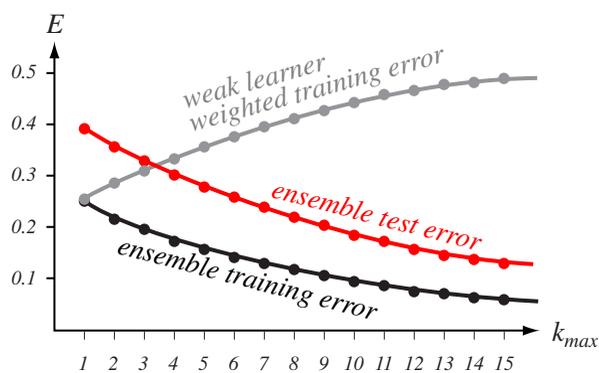
**FIGURE 9.4.** The bias-variance dilemma can be illustrated in the domain of regression. Each column represents a different model, and each row represents a different set of  $n = 6$  training points,  $\mathcal{D}_i$ , randomly sampled from the true function  $F(x)$  with noise. Probability functions of the mean-square error of  $E = \mathcal{E}_{\mathcal{D}}[(g(x) - F(x))^2]$  of Eq. 11 are shown at the bottom. Column a shows a very poor model: a linear  $g(x)$  whose parameters are held fixed, *independent* of the training data. This model has high bias and zero variance. Column b shows a somewhat better model, though it too is held fixed, independent of the training data. It has a lower bias than in column a and has the same zero variance. Column c shows a cubic model, where the parameters are trained to best fit the training samples in a mean-square-error sense. This model has low bias and a moderate variance. Column d shows a linear model that is adjusted to fit each training set; this model has intermediate bias and variance. If these models were instead trained with a very large number  $n \rightarrow \infty$  of points, the bias in column c would approach a small value (which depends upon the noise), while the bias in column d would not; the variance of all models would approach zero. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



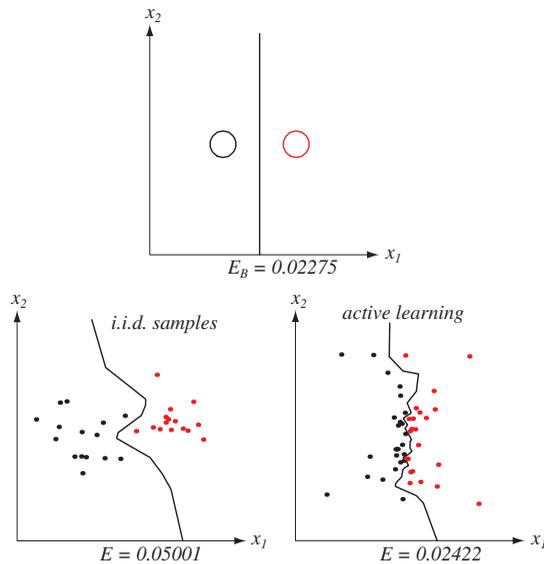
**FIGURE 9.5.** The (boundary) bias-variance trade-off in classification is illustrated with a two-dimensional Gaussian problem. The figure at the top shows the true distributions and the Bayes decision boundary. The figure at the middle shows different learned decision boundaries. Each row corresponds to a different training set of  $n = 8$  points selected randomly from the true distributions and labeled according to the true decision boundary. Column a shows case of a Gaussian model with fully general covariance matrices trained by maximum-likelihood. The learned boundaries differ significantly from one data set to the next; this learning algorithm has high variance. Column b shows the decision boundaries resulting from fitting a Gaussian model with diagonal covariances; in this case the decision boundaries vary less from one row to another. This learning algorithm has a lower variance than the one at the left. Finally, column c shows decision boundaries learning by fitting a Gaussian model with unit covariances (i.e., a linear model); notice that the decision boundaries are nearly identical from one data set to the next. This algorithm has low variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



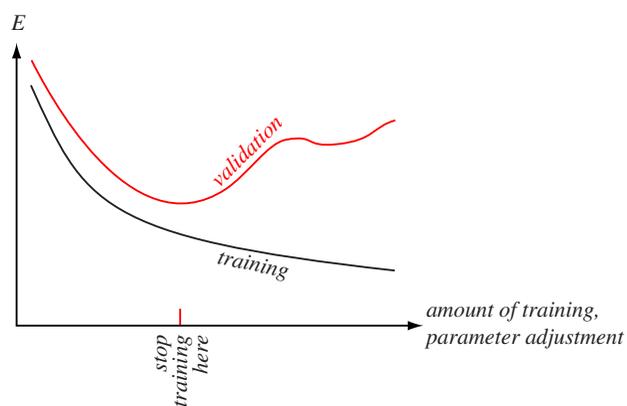
**FIGURE 9.6.** A two-dimensional two-category classification task is shown at the top. The middle row shows three component (linear) classifiers  $C_k$  trained by LMS algorithm (Chapter 5), where their training patterns were chosen through the basic boosting procedure. The final classification is given by the voting of the three component classifiers and yields a nonlinear decision boundary, as shown at the bottom. Given that the component classifiers are weak learners (i.e., each can learn a training set at least slightly better than chance), the ensemble classifier will have a lower training error on the full training set  $\mathcal{D}$  than does any single component classifier. Of course, the ensemble classifier has lower error than a single linear classifier trained on the entire data set. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



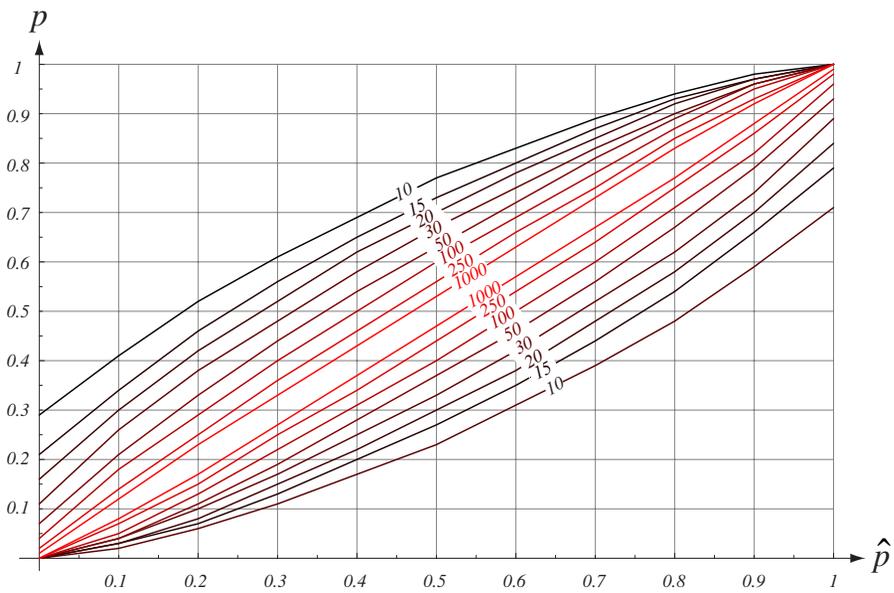
**FIGURE 9.7.** AdaBoost applied to a weak learning system can reduce the training error  $E$  exponentially as the number of component classifiers,  $k_{max}$ , is increased. Because AdaBoost “focuses on” *difficult* training patterns, the training error of each successive component classifier (measured on its own weighted training set) is generally larger than that of any previous component classifier (shown in gray). Nevertheless, so long as the component classifiers perform better than chance (e.g., have error less than 0.5 on a two-category problem), the weighted ensemble decision of Eq. 36 ensures that the training error will decrease, as given by Eq. 37. It is often found that the test error decreases in boosted systems as well, as shown in red. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



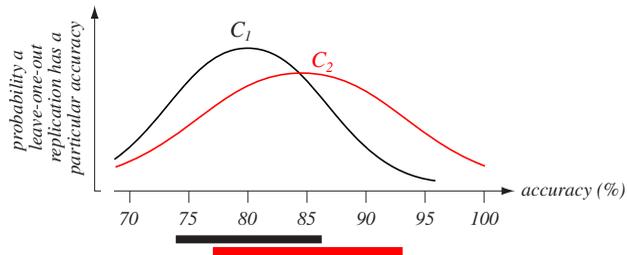
**FIGURE 9.8.** Active learning can be used to create classifiers that are more accurate than ones using i.i.d. sampling. The figure at the top shows a two-dimensional problem with two equal circular Gaussian priors; the Bayes decision boundary is a straight line and the Bayes error  $E_B$  equals 0.02275. The bottom figure on the left shows a nearest-neighbor classifier trained with  $n = 30$  labeled points sampled i.i.d. from the true distributions. Note that most of these points are far from the decision boundary. The figure at the right illustrates active learning. The first four points were sampled near the extremes of the feature space. Subsequent query points were chosen midway between two points already used by the classifier, one randomly selected from each of the two categories. In this way, successive queries to the oracle “focused in” on the true decision boundary. The final generalization error of this classifier,  $E = 0.02422$ , is lower than the one trained using i.i.d. samples,  $E = 0.05001$ . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



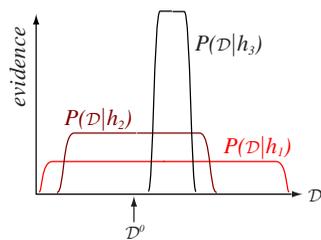
**FIGURE 9.9.** In validation, the data set  $\mathcal{D}$  is split into two parts. The first (e.g., 90% of the patterns) is used as a standard training set for setting free parameters in the classifier model; the other (e.g., 10%) is the validation set and is meant to represent the full generalization task. For most problems, the training error decreases monotonically during training, as shown in black. Typically, the error on the validation set decreases, but then increases, an indication that the classifier may be overfitting the training data. In validation, training or parameter adjustment is stopped at the first minimum of the validation error. In the more general method of cross-validation, the performance is based on multiple independently formed validation sets. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



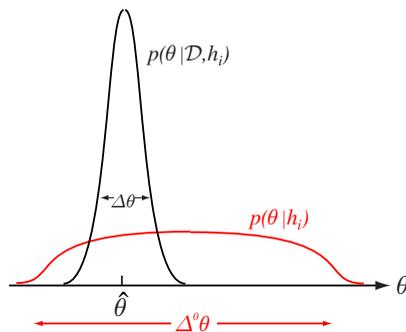
**FIGURE 9.10.** The 95% confidence intervals for a given estimated error probability  $\hat{p}$  can be derived from a binomial distribution of Eq. 38. For each value of  $\hat{p}$ , the true probability has a 95% chance of lying between the curves marked by the number of test samples  $n'$ . The larger the number of test samples, the more precise the estimate of the true probability and hence the smaller the 95% confidence interval. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



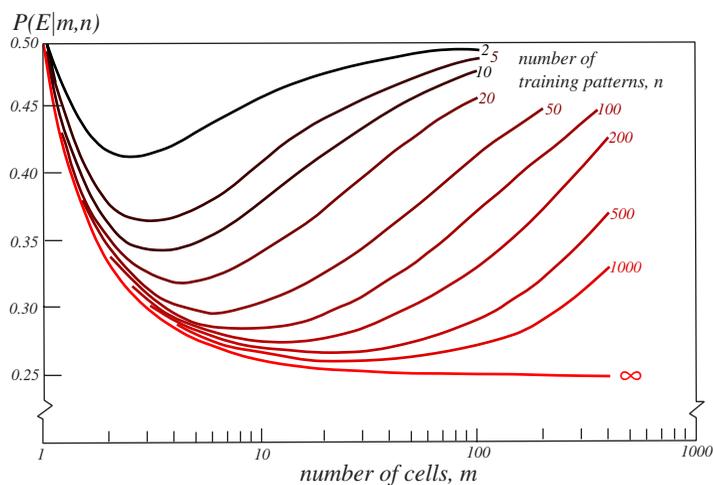
**FIGURE 9.11.** Jackknife estimation can be used to compare the accuracies of classifiers. The jackknife estimate of classifiers  $C_1$  and  $C_2$  are 80% and 85%, and full widths (twice the square root of the jackknife estimate of the variances) are 12% and 15%, as shown by the bars at the bottom. In this case, traditional hypothesis testing could show that the difference is not statistically significant at some confidence level. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



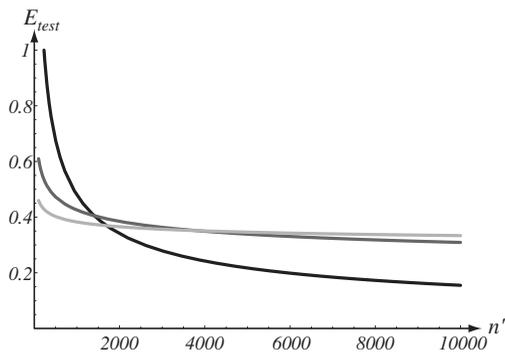
**FIGURE 9.12.** The evidence (i.e., probability of generating different data sets given a model) is shown for three models of different expressive power or complexity. Model  $h_1$  is the most expressive, because with different values of its parameters the model can fit a wide range of data sets. Model  $h_3$  is the most restrictive of the three. If the actual data observed is  $\mathcal{D}^0$ , then maximum-likelihood model selection states that we should choose  $h_2$ , which has the highest evidence. Model  $h_2$  “matches” this particular data set better than do the other two models, and it should be selected. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



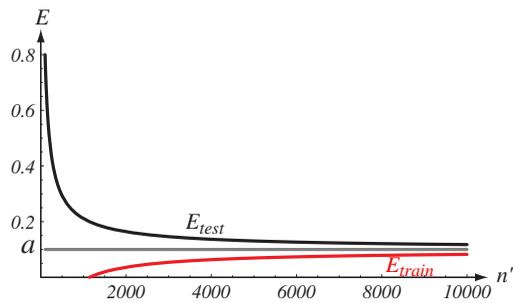
**FIGURE 9.13.** In the absence of training data, a particular model  $h_i$  has available a large range of possible values of its parameters, denoted  $\Delta^0\theta$ . In the presence of a particular training set  $\mathcal{D}$ , a smaller range is available. The Occam factor,  $\Delta\theta/\Delta^0\theta$ , measures the fractional decrease in the volume of the model's parameter space due to the presence of training data  $\mathcal{D}$ . In practice, the Occam factor can be calculated fairly easily if the evidence is approximated as a  $k$ -dimensional Gaussian, centered on the maximum-likelihood value  $\hat{\theta}$ . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



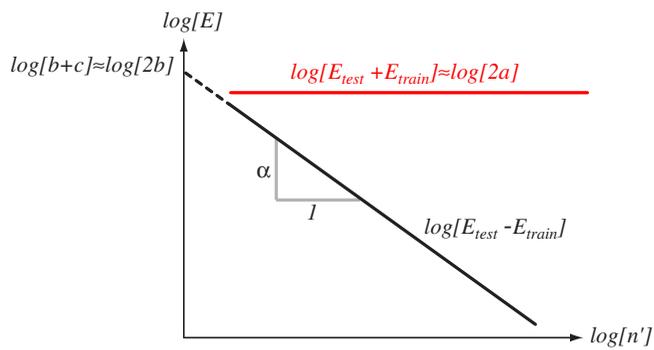
**FIGURE 9.14.** The probability of error  $E$  on a two-category problem for a given number of samples,  $n$ , can be estimated by splitting the feature space into  $m$  cells of equal size and classifying a test point according to the label of the most frequently represented category in the cell. The graphs show the average error of a large number of random problems having the given  $n$  and  $m$  indicated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



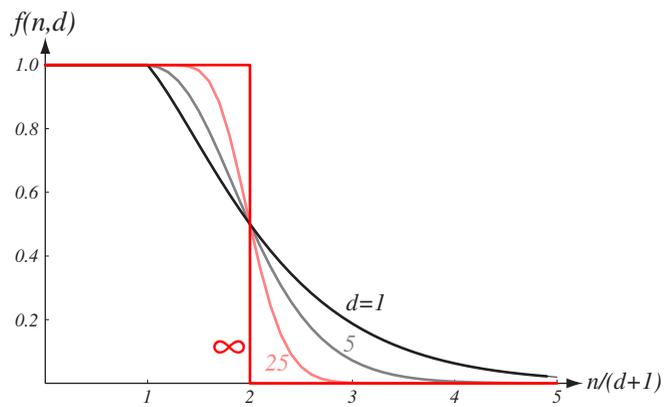
**FIGURE 9.15.** The test error for three classifiers, each fully trained on the given number  $n'$  of training patterns, decreases in a typical monotonic power-law function. Notice that the rank order of the classifiers trained on  $n' = 500$  points differs from that for  $n' = 10000$  points and the asymptotic case. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



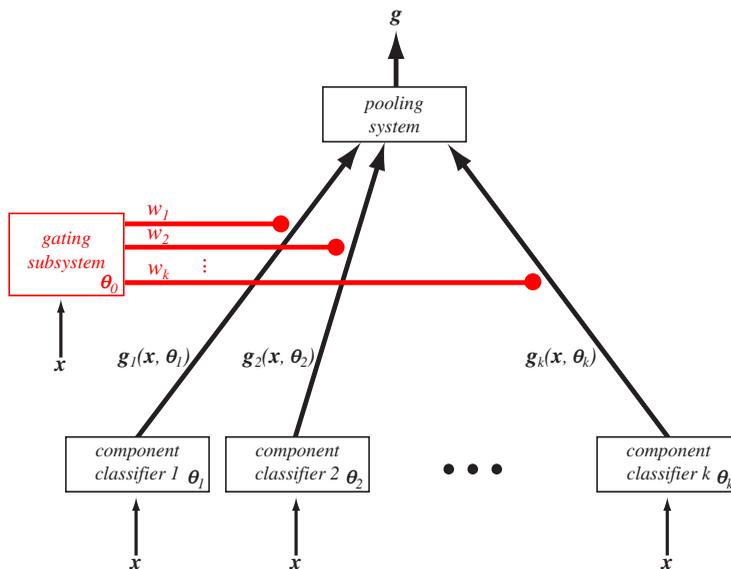
**FIGURE 9.16.** Test and training error of a classifier fully trained on data subsets of different size  $n'$  selected randomly from the full set  $\mathcal{D}$ . At low  $n'$ , the classifier can learn the category labels of the points perfectly, and thus the training error vanishes there. In the limit  $n' \rightarrow \infty$ , both training and test errors approach the same asymptotic value,  $a$ . If the classifier is sufficiently powerful and the training data is sampled i.i.d., then  $a$  is the Bayes error rate,  $E_B$ . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



**FIGURE 9.17.** If the test and training errors versus training set size obey the power-law functions of Eqs. 49 and 50, then the log of the sum and log of the difference of these errors are straight lines on a log-log plot. The estimate of the asymptotic error rate  $a$  is then simply related to the height of the  $\log[E_{test} + E_{train}]$  line, as shown. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



**FIGURE 9.18.** The fraction of dichotomies of  $n$  points in  $d$  dimensions that are linear, as given by Eq. 53. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



**FIGURE 9.19.** The mixture-of-experts architecture consists of  $k$  component classifiers or “experts,” each of which has trainable parameters  $\theta_i$ ,  $i = 1, \dots, k$ . For each input pattern  $\mathbf{x}$ , each component classifier  $i$  gives estimates of the category membership  $g_{ir} = P(\omega_r | \mathbf{x}, \theta_i)$ . The outputs are weighted by the gating subsystem governed by parameter vector  $\theta_0$  and are pooled for ultimate classification. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.