

Statistical Learning Theory

E.P.D. Pednault

**MIT Encyclopedia of the Cognitive Sciences
1998**

Statistical Learning Theory*

Edwin P. D. Pednault
Mathematical Sciences Department
IBM T. J. Watson Research Center
Yorktown Heights, New York 10598

July 22, 1997

Statistical learning theory addresses a key question that arises when constructing predictive models from data—how to decide whether a particular model is adequate or whether a different model would produce better predictions. Whereas classical statistics typically assumes that the form of the correct model is known and the objective is to estimate the model parameters, statistical learning theory presumes that the correct form is completely unknown and the goal is to identify the best possible model from a set of competing models. The models need not have the same mathematical form and none of them need be correct. The theory provides a sound statistical basis for assessing model adequacy under these circumstances, which are precisely the circumstances encountered in machine learning, pattern recognition, and exploratory data analysis.

Estimating the performance of competing models is the central issue in statistical learning theory. Performance is measured through the use of loss functions. The loss $Q(\mathbf{z}, \alpha)$ between a data vector \mathbf{z} and a specific model α (one with values assigned to all parameters) is a score that indicates how well α performs on \mathbf{z} , with lower scores indicating better performance. The squared-error function for regression models, the 0/1 loss function for classification models, and the negative log likelihood for other more general statistical models are all examples of loss functions. The choice of loss function depends on the nature of the modeling problem.

From the point of view of utility theory, α is a decision variable, \mathbf{z} is an outcome, and $Q(\mathbf{z}, \alpha)$ is the negative utility of the outcome given the decision. If the statistical properties of the data were already known, the optimum model would therefore be the α that minimizes the expected loss $R(\alpha)$:

$$R(\alpha) = \mathbb{E}_{\mathbf{z}}[Q(\mathbf{z}, \alpha)] = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}),$$

where $F(\mathbf{z})$ is the probability measure that defines the true statistical properties of the data. $R(\alpha)$ is also referred to as the *risk* of α . In learning situations, $F(\mathbf{z})$ is unknown and one must choose a model based on a set of observed data vectors \mathbf{z}_i , $i = 1, \dots, \ell$, that are assumed to be random samples of $F(\mathbf{z})$. The average loss $R_{\text{emp}}(\alpha, \ell)$ on the observed data is used as an empirical estimate of the expected loss, where

$$R_{\text{emp}}(\alpha, \ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(\mathbf{z}_i, \alpha).$$

$R_{\text{emp}}(\alpha, \ell)$ is also referred to as the *empirical risk* of α .

The fundamental question in statistical learning theory is the following: under what conditions does minimizing $R_{\text{emp}}(\alpha, \ell)$ yield models α that also minimize $R(\alpha)$, since the latter is what

*To appear in the *MIT Encyclopedia of the Cognitive Sciences*. R. Wilson and F. Keil, General Editors. M. Jordan and S. Russell, Advisory Editors, AI/Computer Science.

we actually want to accomplish? This question is answered by considering the accuracy of the empirical loss estimate.

As in classical statistics, accuracy is expressed in terms of confidence regions; that is, how far can $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$ be expected to deviate from $R(\boldsymbol{\alpha})$ and with what probability? One of the fundamental theorems of statistical learning theory shows that the size of the confidence region is governed by the maximum difference between the two losses over all models being considered:

$$\sup_{\boldsymbol{\alpha} \in \mathbf{A}} \left| R(\boldsymbol{\alpha}) - R_{\text{emp}}(\boldsymbol{\alpha}, \ell) \right| ,$$

where \mathbf{A} is a set of competing models. The maximum difference dominates because of the phenomenon of overfitting.

Overfitting occurs when the best model relative to the training data tends to perform significantly worse when applied to new data. This mathematically corresponds to a situation in which the average loss $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$ substantially underestimates the expected loss $R(\boldsymbol{\alpha})$. Although there is always some probability that underestimation will occur for a fixed model $\boldsymbol{\alpha}$, both the probability and the degree of underestimation are increased by the fact that we explicitly search for the $\boldsymbol{\alpha}$ that minimizes $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$. This search biases the difference between $R(\boldsymbol{\alpha})$ and $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$ toward the maximum difference among competing models. If the maximum difference does not converge to zero as the number of data vectors ℓ increases, then overfitting will occur with probability one.

The core results in statistical learning theory are a series of probability bounds developed by Vapnik and Chervonenkis [1, 2, 3] that define small-sample confidence regions for the maximum difference between $R(\boldsymbol{\alpha})$ and $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$. The confidence regions differ from those obtained in classical statistics in three respects. First, they do not assume that the chosen model is correct. Second, they are based on small-sample statistics and are not asymptotic approximations. Third, a uniform method is used to take into account the degree to which overfitting can occur for a given set of competing models. This method is based on a measurement known as the Vapnik-Chervonenkis (VC) dimension.

Conceptually speaking, the VC dimension of a set of models is the maximum number of data vectors for which overfitting is virtually guaranteed in the sense that one can always find a specific model that fits the data exactly. For example, the VC dimension of the family of linear discriminant functions with n parametric terms is n , since n linear terms can be used to exactly discriminate n points in general position for any two-class labeling of the points. This conceptual definition of VC dimension accurately reflects the formal definition in the case of 0/1 loss functions. The formal definition is more general in that it considers arbitrary loss functions and does not require exact fits.

In the probability bounds obtained by Vapnik and Chervonenkis, the size of the confidence region is largely determined by the ratio of the VC dimension h to the number of data vectors ℓ . For example, if $Q(\mathbf{z}, \boldsymbol{\alpha})$ is the 0/1 loss function used for classification, then with probability at least $1 - \eta$,

$$R_{\text{emp}}(\boldsymbol{\alpha}, \ell) - \frac{\sqrt{\mathcal{E}}}{2} \leq R(\boldsymbol{\alpha}) \leq R_{\text{emp}}(\boldsymbol{\alpha}, \ell) + \frac{\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4 R_{\text{emp}}(\boldsymbol{\alpha}, \ell)}{\mathcal{E}}} \right) ,$$

where

$$\mathcal{E} = \frac{4h}{\ell} \left(\ln \frac{2\ell}{h} + 1 \right) - \frac{4}{\ell} \ln \left(\frac{\eta}{4} \right) .$$

Note that the ratio of h over ℓ is the dominant term in the definition of \mathcal{E} and, hence, in the size of the confidence region for $R(\boldsymbol{\alpha})$.

Vapnik [4, 5, 6] has reported probability bounds for other families of loss functions that yield analogous confidence regions based on VC dimension. Bounds also exist for the special case in which the set of competing models is finite (continuous parameters typically imply an infinite number of specific models). These bounds avoid explicit calculation of VC dimension and are useful in validation-set methods. A remarkable property shared by all of the bounds is that they either make no assumptions at all or very weak assumptions about underlying probability distribution $F(\mathbf{z})$. In addition, they are valid for small sample sizes and they depend only on the VC dimension of the set of competing models \mathbf{A} , or on its size, and on the properties of the loss function $Q(\mathbf{z}, \boldsymbol{\alpha})$. All bounds are independent of the mathematical forms of the models—the VC dimension and/or the number of specific models summarizes all relevant information. Thus, the bounds are equally applicable to both nonlinear and nonparametric models, and to combinations of dissimilar model families. This includes neural networks, classification and regression trees, classification and regression rules, radial basis functions, Bayesian networks, etc.

When using statistical learning theory to identify the best model from a set of competing models, the models must first be ordered according to preference. The most preferable model that best explains the data is then selected. The preference order corresponds to the notion of learning bias found in machine learning. No restrictions are placed on the ordering other than it must be fixed prior to model selection. The ordering itself is referred to as a *structure* and the process of selecting models is called *structural risk minimization*.

Structural risk minimization has two components: one is to determine a cutoff point in the preference ordering, the other is to select the best model from among those that occur before the cutoff. As the cutoff point is advanced through the ordering, both the subset of models that appear before the cutoff and the VC dimension of this subset steadily increase. With more models to choose from, the minimum average loss $R_{\text{emp}}(\boldsymbol{\alpha}, \ell)$ for all models $\boldsymbol{\alpha}$ before the cutoff tends to decrease. However, the size of the confidence region for $R(\boldsymbol{\alpha})$ tends to increase because the size is governed by the VC dimension. The cutoff point is selected by minimizing the upper bound on the confidence region for $R(\boldsymbol{\alpha})$, with the corresponding $\boldsymbol{\alpha}$ chosen as the most suitable model given the available data. For example, for classification problems one would choose the cutoff and the associated model $\boldsymbol{\alpha}$ so as to minimize the right hand side of the inequality presented above for a desired setting of the confidence parameter η .

The overall approach is illustrated by the graph in Figure 1. The process balances the ability to find increasingly better fits to the data against the danger of overfitting and thereby selecting a poor model. The preference ordering provides the necessary structure in which to compare competing models. Judicious choice of the ordering enables one to avoid overfitting even in high-dimensional spaces. For example, Vapnik [5, 6] orders models within parametric families according to the magnitudes of the parameters. Each preference cutoff then limits the parameter magnitudes, which in turn limits the VC dimension of the corresponding subset of models. Reliable models can thus be obtained using structural risk minimization even when the number of data samples is orders of magnitude less than the number of parameters.

References

- [1] VAPNIK, V. N., AND CHERVONENKIS, A. JA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, pp. 264–280. Originally published in (1968) *Doklady Akademii Nauk USSR*, **181**(4).

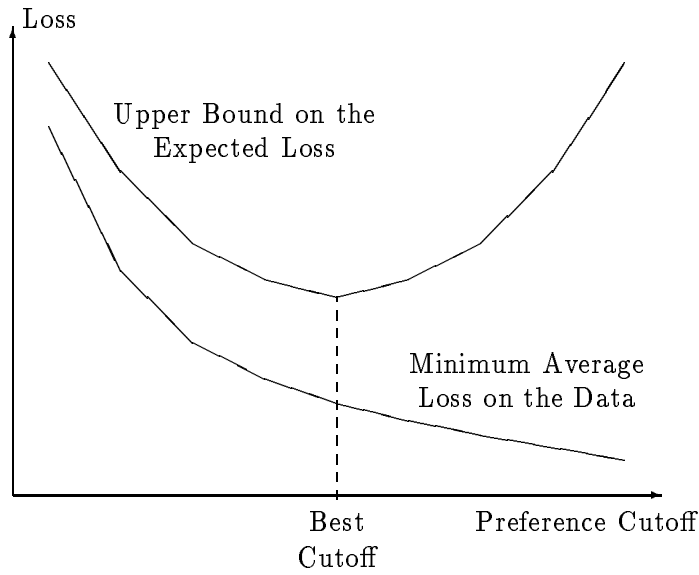


Figure 1: Expected loss and average loss on the data as a function of the preference cutoff.

- [2] VAPNIK, V. N., AND CHERVONENKIS, A. JA. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, **26**, pp. 532–553.
- [3] VAPNIK, V. N., AND CHERVONENKIS, A. JA. (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, **1**, (3), pp. 284–305. Originally published in (1989) *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting*, **2**.
- [4] VAPNIK, V. N. (1982). *Estimation of dependencies based on empirical data*. New York, New York: Springer-Verlag.
- [5] VAPNIK, V. N. (1995). *The nature of statistical learning theory*. New York, New York: Springer-Verlag.
- [6] VAPNIK, V. N. (to appear, 1998). *Statistical learning theory*. New York, New York: Wiley.