Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Receiver Operating Characteristics of Perceptrons: Influence of Sample Size and Prevalence

Ansgar Freking, Michael Biehl, Christian Braun,
Wolfgang Kinzel, and Malte Meesmann

Ref.: WUE-ITP-98-049
e-mail : freking@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de, directory pub/preprint/1998

Receiver Operating Characteristics of Perceptrons: Influence of Sample Size and Prevalence

A Freking†, M Biehl†, C Braun‡,
W Kinzel†, and M Meesmann‡

† Institut für Theoretische Physik, Universität Würzburg, Germany
‡ Medizinische Universitätsklinik Würzburg, Germany

Abstract.

In many practical classification problems it is important to distinguish false positive from false negative results when evaluating the performance of the classifier. This is of particular importance for medical diagnostic tests. In this context, receiver operating characteristic (ROC) curves have become a standard tool.

Here we apply this concept to characterize the performance of a simple neural network. Investigating the binary classification of a perceptron we calculate analytically the shape of the corresponding ROC curves. The influence of the size of the training set and the prevalence of the quality considered are studied by means of a statistical-mechanics analysis.

PACS numbers: 87.10, 07.05, 05.90

1. Introduction

Classification problems in general and medical diagnostic tests in particular are often well suited for the application of neural networks [1, 2, 3]. The rule how to classify an item is generally not available, but can be derived from examples, e.g. patients with a known clinical status. The task for the network is to learn from these examples, i.e. to extract the implicit information in order to classify other items. We focus on medical diagnostic tests, where the aim is to discriminate between absence and presence of a certain disease or risk, yet our analysis is not restricted to this domain.

To gain further insight into the applicability of the neural network approach in a clinical setting we calculated the required size of the training set in relation to the prevalence of the disease considered. For this purpose we provide analytical expressions for the influence of prevalence on the shape of receiver operating characteristic curves.

2. Definition of descriptive measures

For the evaluation of a diagnostic test, it is often important to distinguish false positive from false negative results. In this case the so called generalization error as defined in the context of learning theory is not sufficient to assess the validity of the test.

Table 1. Cross table containing the relative frequencies of positive and negative test results subdivided with respect to the clinical status (disease vs. no disease or risk vs. no risk). a depicts correctly classified diseased persons, d represents correctly classified normals. $b + c$ gives the fraction of false classifications, which is often referred to as *generalization error.*

		Clinical status	
		with disease	without disease
Test result	positive	a	b
	negative	c	d

A common way to summarize the results of a medical test is to list the frequencies of positive and negative test results in a two-by-two cross table where the columns correspond to the clinical status of the patients (see Table 1). In the following we use the convention that all frequencies are counted relatively to the total number of patients, i.e. $a + b + c + d = 1$.

From the cross table one defines the ratios

$$\begin{aligned} \text{SENS} &= \frac{a}{a+c} \qquad & \text{PPV} &= \frac{a}{a+b} \\ \text{SPEC} &= \frac{d}{b+d} \qquad & \text{NPV} &= \frac{d}{c+d} \end{aligned} \tag{1}$$

which are standard measures to describe the performance of a medical diagnostic test. The *sensitivity* (SENS) gives the percentage of correctly classified diseased persons. Thus, a sensitivity of 100% means that any occurrence of the disease is detected by the test. The *specificity* (SPEC) gives the analogue ratio of the correctly classified persons without the disease. Whereas sensitivity and specificity have a more global meaning, physicians might be more interested in the ratios taken with respect to the rows. If the test is positive, the *positive predictive value* (PPV) gives the probability to have the disease. The *negative predictive value* (NPV) tells the reliability of a negative test result. The fraction of diseased persons in the sample is called *prevalence*,

$$\text{PREV} = a + c \tag{2}$$

Usually, diagnostic tests yield a continuous valued quantity which is compared to a threshold value for binary classification. By varying this threshold, the test can be made more or less stringent to meet given requirements with respect to sensitivity or specificity.

A more stringent test gives high specificity, i.e. the fraction of correctly classified healthy persons is high. On the other hand, an increase in specificity usually results in a loss of sensitivity, i.e. more persons with the disease are missed.

This trade-off between specificity and sensitivity is described by the *receiver operating characteristic* (ROC) curve. An ROC curve is the plot of a test's true positive rate, i.e. the fraction $y_{\text{ROC}} = a/(a + c)$ of correctly classified persons with the disease, as a function of its false positive rate, the fraction $x_{\text{ROC}} = b/(b + d)$ of misclassified healthy persons. Using the definitions (1), it is obvious that $y_{\mathrm{ROC}} = \mathrm{SENS}$ and $x_{\mathrm{ROC}} = (1 - \mathrm{SPEC})$.

Each point on the ROC curve corresponds to a certain stringency level of the test. If a diagnostic test makes exclusively positive classifications, the true positive rate is one, but the same applies to the false positive rate. The corresponding point on the ROC curve is the upper right corner, $y_{\mathrm{ROC}} = x_{\mathrm{ROC}} = 1$. In the opposite extreme, there are no positive classifications and therefore one has $y_{\mathrm{ROC}} = x_{\mathrm{ROC}} = 0$.

For a test without predictive power, the ratio between positive classified persons with and without disease would be the same as the corresponding ratio of the whole test population; in this case one has

$$\frac{a}{b} = \frac{a+c}{b+d}$$

and the ROC curve follows the diagonal $y_{\mathrm{ROC}} = x_{\mathrm{ROC}}$. Any meaningful test should yield an ROC curve above this diagonal. By changing the threshold value of the test, one obtains a curve extending between the two corners $y_{\mathrm{ROC}} = x_{\mathrm{ROC}} = 0$ and $y_{\mathrm{ROC}} = x_{\mathrm{ROC}} = 1$.

The given conditions in medical practice often require a certain level of specificity for a test to be useful. If one thinks of a risky or cost expensive treatment, it is clear that the fraction of misclassified healthy persons should not exceed a certain limit. On the other hand, a diagnostic test should achieve some minimal value of sensitivity in order to be effective. Since ROC curves display precisely this interplay, they can be used to determine the threshold value for a certain diagnostic test. Additionally, ROC curves allow for comparison of several tests at the same level of specificity. For a detailed discussion of the use of ROC curves see for instances [4] and references therein.

3. The data model

In our analysis, each patient is represented by an N-dimensional feature vector $\boldsymbol{S} \in \mathbb{R}^{\mathrm{N}}$. Each component of this vector can be thought of as a clinically relevant measure which characterizes the patient (e.g. blood pressure, heart rate). The patient's clinical status is coded by the binary quantity S_0 where $S_0 = +1$ means *with* and $S_0 = -1$ means *without* disease.

Let us now assume, that diseased constellations in feature space can be separated from the healthy ones by a $(N-1)$-dimensional hyperplane. This means, that the actual status S_0 of a patient with feature vector $\boldsymbol{S}$ is given by

$$S_0 = \mathrm{sign}(\boldsymbol{B}\cdot\boldsymbol{S} - \theta), \tag{3}$$

where $\boldsymbol{B} \in \mathbb{R}^{\mathrm{N}}$, $\boldsymbol{B}^2 = 1$, is a unit-vector perpendicular to the hyperplane; for obvious reasons $\boldsymbol{B}$ is called the *rule* vector. θ is related to the prevalence, i.e. the fraction of actually diseased individuals. The components of $\boldsymbol{S}$ are taken to be Gaussian distributed with zero mean and unit variance, therefore the overlap $y = \boldsymbol{B}\cdot\boldsymbol{S}$ of the feature vectors with the rule is Gaussian distributed as well,

$$p(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right). \tag{4}$$

Note, that in case of large N, (4) holds true under more general conditions according to the central limit theorem. Given the distribution (4) of y, the threshold θ is related to the prevalence through

$$\mathrm{PREV} = \Phi(\theta) \qquad \text{with} \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{d}t \, \exp\left(-\frac{1}{2}t^2\right) \tag{5}$$

4. ROC curves of a perceptron

In this section we consider a perceptron with fixed weight vector $\boldsymbol{J} \in \mathbb{R}^{\mathrm{N}}$, $\boldsymbol{J}^2 = 1$. As does the rule vector, $\boldsymbol{J}$ defines a hyperplane in the feature space, and

$$\sigma = \mathrm{sign}(\boldsymbol{J}\cdot\boldsymbol{S} - \gamma) \tag{6}$$

is the test result for the patient $\boldsymbol{S}$. We use the overlap $R = \boldsymbol{J}\cdot\boldsymbol{B}$ as the usual measure for the perceptron's knowledge about the rule. For a given R, the projection $x = \boldsymbol{J}\cdot\boldsymbol{S}$ of a feature vector on the perceptron vector $\boldsymbol{J}$ and the projection $y = \boldsymbol{B}\cdot\boldsymbol{S}$ on the rule vector are jointly Gaussian distributed according to

$$p(x, y) = \frac{1}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{1}{2}\,\frac{x^2 - 2Rxy + y^2}{1-R^2}\right). \tag{7}$$

In order to plot the ROC curves for the perceptron classification, we need to know the entries of the respective cross table. If we want to calculate, say b, which is the relative frequency of $((S_0 = -1) \wedge (\sigma = +1))$, we have to perform the average of $\Theta(-S_0)\Theta(+\sigma)$ over the distribution of feature vectors $\boldsymbol{S}$. Since $\boldsymbol{S}$ enters only through the scalar products x and y, we arrive in an average over (7):

$$\begin{aligned} b &= \int_{\gamma}^{\infty} \mathrm{d}x \int_{-\infty}^{\theta} \mathrm{d}y \; p(x, y) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\gamma}^{\infty} \mathrm{d}x \exp\left(-\frac{1}{2}x^2\right) \Phi\left(\frac{\theta - Rx}{\sqrt{1-R^2}}\right) =: \Psi(\gamma, \theta, R). \end{aligned} \tag{8}$$

All other entries of the cross table can be calculated in the same way and hence can be expressed through the function Ψ defined in (8):

$$a = \Psi(\gamma, -\theta, -R), \qquad c = \Psi(-\gamma, -\theta, R) \qquad \text{and} \qquad d = \Psi(-\gamma, \theta, -R) \tag{9}$$

This allows us to express the quantities in (1) in terms of the perceptron threshold γ, the bias θ which is related to the prevalence (5) and the overlap R:

$$\begin{aligned} \mathrm{SENS} &= \frac{\Psi(\gamma, -\theta, -R)}{\Phi(\theta)} \qquad & \mathrm{PPV} &= \frac{\Psi(\gamma, -\theta, -R)}{\Phi(\gamma)} \\ \mathrm{SPEC} &= \frac{\Psi(-\gamma, \theta, -R)}{\Phi(-\theta)} \qquad & \mathrm{NPV} &= \frac{\Psi(-\gamma, \theta, -R)}{\Phi(-\gamma)} \end{aligned} \tag{10}$$

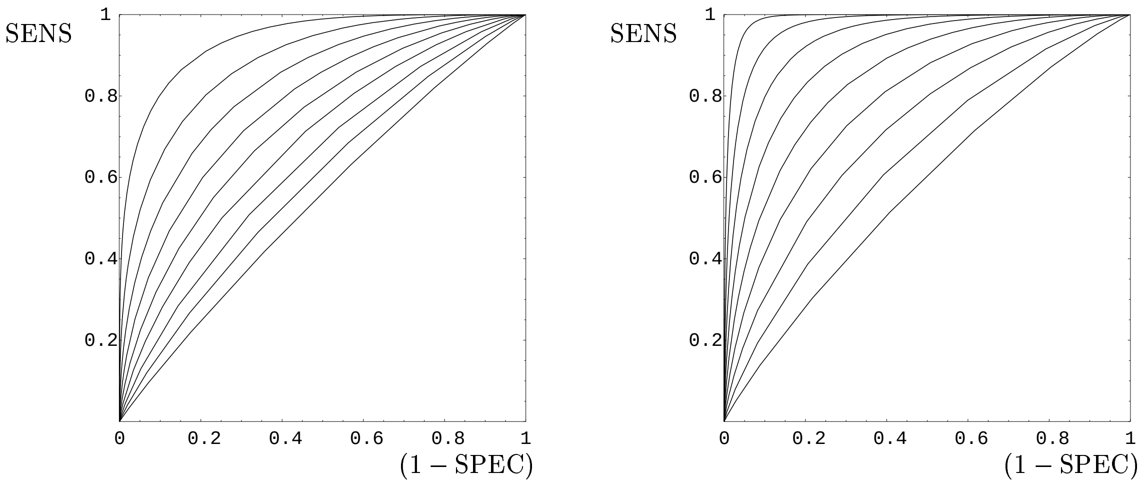

Figure 1. ROC curves of perceptrons with overlaps $R = 0.1, 0.2, \ldots 0.9$ (both panels); higher values of R give larger areas under the curves. For the left panel the prevalence was set to PREV = 50%, in the right panel to PREV = 1%.

For a prevalence of 50%, the curves are symmetric with respect to the line $(0, 1) - (1, 0)$, which reflects the symmetry between *diseased* and *healthy* in this case.

For given external conditions, i.e. θ and R fixed, the performance measures in (10) can only be tuned relative to each other by the choice of the perceptron threshold γ. A higher value of γ gives less positive classifications (cf. (6)) and therefore an increase of SPEC with a concomitant decrease of SENS.

The plot of SENS, i.e. the fraction of positive classifications among the diseased persons, versus (1 − SPEC), the fraction of positive classifications in the healthy subgroup, gives the ROC curve.

As $\gamma \to +\infty$ there are no positive test results and the respective point on the ROC curve is the lower left corner, $\gamma \to -\infty$ equivalently corresponds to the upper right corner. For finite γ the value of R determines the path in between the corners. The higher the value of R, the higher the sensitivity at a certain specificity and the larger the area under the curve.

Figure 1 presents ROC curves for two different prevalences (1% and 50%) and several values of R. In comparing the two plots, we find that a perceptron with a certain overlap R reaches higher sensitivities for all specificities when there are less patients with disease then without disease. The ROC curves for the case PREV = 1% are steeper than the ones for PREV = 50%. On the other hand, it should be harder for the perceptron to achieve a certain knowledge about the rule, if there are less examples of one group at a fixed total amount of training examples.

5. Learning from a training set

Up to now, we have considered a perceptron with fixed weights $\boldsymbol{J}$ and a certain overlap R with the rule. The aim of this section is to include a learning process into the analysis.

The perceptron gains knowledge about the rule by learning from examples provided in the training set. Consequently, we shall analyze the influence of the size of the training set and the prevalence of the quality considered to the performance of the perceptron. This means, we have to calculate the quantity R as a function of the total number P of training examples and the relative frequency of the label $S_0 = +1$. Finally, this gives us access to the desired accuracy measures defined in the previous sections.

We apply the standard statistical mechanics approach and consider the components of the perceptron weight vector $\boldsymbol{J}$ as the N degrees of freedom of a physical system with energy

$$H(\boldsymbol{J}) = \sum_{\mu=1}^{P} \Theta(-\sigma^\mu S_0^\mu) = \sum_{\mu=1}^{P} \Theta\left((\boldsymbol{J}\cdot\boldsymbol{S}^\mu - \gamma)(\boldsymbol{B}\cdot\boldsymbol{S}^\mu - \theta)\right), \tag{11}$$

where the sum extends over all feature vectors $\boldsymbol{S}^\mu$ in the training set. The threshold γ is considered to be fixed during the training process.

In an ensemble of perceptrons at formal temperature $1/\beta$, a configuration $\boldsymbol{J}$ occurs with the corresponding Gibbs–Boltzmann density, hence the term *Gibbs learning* has been coined for this scenario [5]. The quenched average over the randomness contained in the training data is performed using the replica method assuming replica symmetry. A sketch of the calculation is given in the appendix.

The limit $\beta \to \infty$ forces the system into its ground state. For $\gamma = \theta$ the energy of the ground state is $H = 0$, independent of $\alpha = P/N$. This corresponds to an error free classification of a training set of any size. In the limit $N \to \infty$ with finite normalized sample size $\alpha = P/N$, θ can be determined exactly from the prevalence and $\gamma = \theta$ is a valid choice for the perceptron threshold during training. Since we expect, that this choice already gives the largest achievable overlap within the framework of Gibbs learning, we proceed without an explicit optimization of the learning strategy with respect to γ for given θ, and use $\gamma = \theta$ instead. Note that this particular choice of γ is only used to describe a specific training process and does not affect the role of the threshold as described in the previous sections.

The overlap R defined in section 4 now plays the role of an order parameter. Together with the quantity q, which represents the typical overlap of two error free perceptron vectors (A.2), it is sufficient to describe the properties of the system in the thermodynamic limit $N \to \infty$.

In analogy to the case $\gamma = \theta = 0$, which was studied in [6], one can argue, that the unknown vector $\boldsymbol{B}$ coincides with equal probability with any error free $\boldsymbol{J}$. As a consequence the relation $q = R$ holds true which would be violated in more general settings with $\gamma \neq \theta$.

The remaining saddle point condition $d\Gamma/dR = 0$ (cf. (A.4)) yields the so–called learning curve, i.e. R in dependence of α. Figure 2 displays such learning curves for different values of θ. As intuitively clear, the more examples are provided, the better the rule is captured by the perceptron. This works best, if there are equal numbers of examples of both classification types. In case of very few examples for one of the

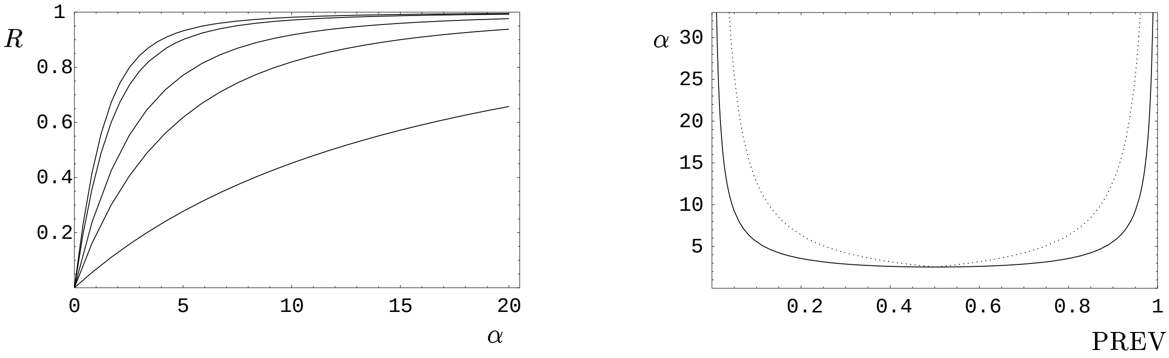

Figure 2. On the left panel, the overlap R of the perceptron vector $\boldsymbol{J}$ with the rule vector $\boldsymbol{B}$ is plotted as a function of the ratio $\alpha = P/N$. The individual curves correspond to the prevalences PREV $= 0.01, 0.05, 0.10, 0.25, 0.50$ from bottom to top. Before having seen any example, i.e. at $\alpha = 0$, $\boldsymbol{J}$ is perpendicular to $\boldsymbol{B}$ and the curves start at $R = 0$ in any case. With increasing α, R increases and approaches asymptotically the value 1 as $\alpha \to \infty$. For a fixed α the largest value of R is achieved when both classes have equal weights, i.e. PREV $= 0.50$. Nevertheless, only very asymmetric class weights cause pronounced reducements. This can also be seen from the right part of the figure. Here the value of α required for a fixed overlap ($R = 0.8$) is plotted against the prevalence (solid line). Whereas the curve is rather flat in the center region it grows dramatically as PREV $\to 0$ or PREV $\to 1$. For comparison, the dotted line shows the value of α required to obtain the same number of examples in the smaller of the subgroups as there are for PREV $= 0.50$.

classes, i.e. for prevalences far away from the balance PREV $= 50\%$, it takes a huge total amount to gain knowledge about the rule. Nevertheless, for ratios from about PREV $= 10\%$ up to PREV $= 90\%$ there is a remarkably weak dependence of α on the prevalence.

A rough but widely used approximation for the dependence of the required sample size on the prevalence is given by the statement, that the smaller subgroup determines the actual sample size. This implies the dependence $\tilde{\alpha} \propto 1/\text{PREV}$ for PREV < 0.50, since the normalized number of examples of the smaller group is given by $(\alpha \cdot \text{PREV})$, in this case. Equivalently the dependence should be $\tilde{\alpha} \propto 1/(1 - \text{PREV})$ in the case PREV > 0.50. The right part of figure 2 shows the value of α necessary to obtain a certain overlap ($R = 0.8$) as a function of PREV; for comparison, $\tilde{\alpha}$ is plotted as well. As expected the approximation $\tilde{\alpha}$ overestimates the actually needed sample size for any value of PREV.

6. How much training data is necessary?

The relation between R, θ and α as derived in the previous section allows for a quantitative description of the influence of sample size and disease prevalence on the performance of a perceptron.

As discussed in section 2, a quantity of interest might be the sensitivity of a test at a given specificity, i.e. the ordinate of the ROC curve at fixed abscissa $(1 - \text{SPEC})$.

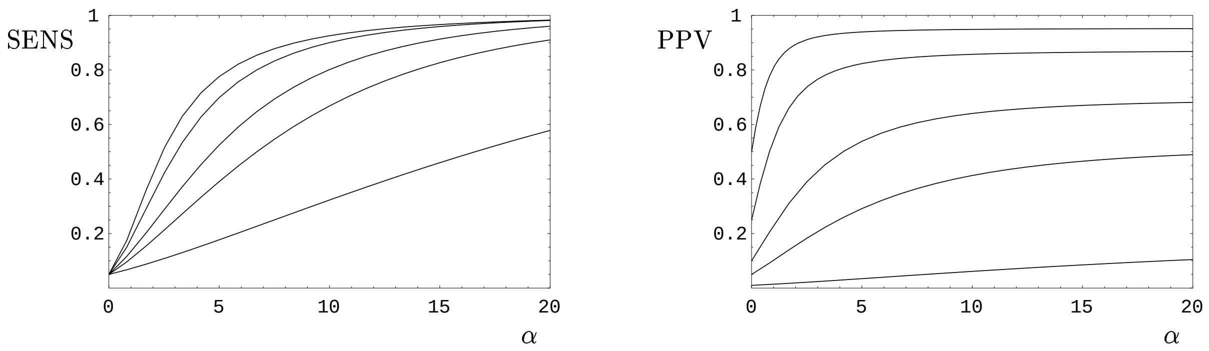

Figure 3. Variation of the perceptron performance with increasing size of the training set. As in the left panel of figure 2 the individual curves of both plots correspond to PREV = 0.01, 0.05, 0.10, 0.25 and 0.50 from bottom to top.

The graph on the left shows the sensitivity at fixed specificity (SPEC = 0.95). The sensitivity at $\alpha = 0$ is common to all curves. This is due to the fact, that the ROC curve for a test without predictive power is given by the line of identity; the value 5% just reflects the considered specificity level of 95%. The α dependence of the sensitivity looks very similar to the shape of the learning curves shown in figure 2.

The corresponding positive predictive values are displayed on the right panel; the values at $\alpha = 0$ coincide with the respective values of the prevalence. The increase in PPV becomes astonishingly slow, yet all curves tend to 1 as $\alpha \to \infty$.

Proceeding on a given prevalence the overlap R can be calculated for any value of α (cf. section 5). From this, the perceptron threshold γ which yields the desired specificity can be determined by using the ROC equations (10). Finally, by using (10) again, one obtains the sensitivity as a function of the normalized size $\alpha = P/N$ of the training set. This dependence is shown in figure 3. The corresponding positive predictive values, which are also presented in figure 3, can be obtained the same way.

From graphs like the ones in figure 3 one can easily read the required size of the training set for a desired test performance. Additionally such graphs may indicate regions, where further collection of training examples gives only neglectable improvements to the test performance. The latter applies especially with respect to the positive predictive values.

7. Discussion

We applied the concept of ROC curves to describe the performance of a perceptron which realizes a threshold classification. For this, we revisited the thoroughly studied scenario where the perceptron learns from examples classified by a rule, which is of the same architecture as the perceptron itself (see for instance [1, 5, 6, 7] and references therein).

Investigating the shape of ROC curves of a perceptron with fixed overlap R, we observed a pronounced dependence of sensitivity and specificity on the prevalence of the quality considered (cf. figure 1). This seems to contradict the frequently encountered statement that these quantities should be prevalence independent measures of validity, which is based on their definition by means of the cross table. It is important to realize that this predication refers to ROC curves obtained from differently composed validation sets for exactly the same classification problem. But even in this sense, the statement does not necessarily apply to realistic situations as discussed in [9].

By introducing the bias θ in the rule, we extended the analysis in [6] to situations where items of the two subgroups occur with different frequencies. Combining the results of this statistical-mechanics approach with standard measures of validity as defined in the context of biomedical statistics, we described analytically the influence of sample size and disease prevalence on the performance of a diagnostic test. It is hoped that this theoretical knowledge allows for more effective planning of clinical studies.

Our work is based on simple, yet rather general assumptions on the underlying data distribution and the classification scheme. Naturally, real world problems are more complicated in many respects. They are usually not completely learnable since the architecture of the rule is not known. In addition, classification tasks in practice will be affected by noise, i.e. the feature vectors can contain inaccurately measured values or the example classifications might be wrong themselves. The considered ideal training situation provides first insights and we expect the results to hold qualitatively in a wider range of settings. In particular, earlier studies of the perceptron have shown, that the presence of noise does not prohibit the success ($R \to 1$) of appropriate training schemes, in principle [6, 10, 11]. Nevertheless, further research should incorporate such more realistic situations, including non–isotropic data distributions and more complicated classification schemes which require, for instance, the use of multilayer networks.

Acknowledgments

We would like to thank E Domany, I Kanter, M Opper, and K Wegscheider for useful and stimulating discussions. This work is supported by the Deutsche Forschungsgemeinschaft, Me 799/3.

Appendix

The partition function corresponding to (11) reads

$$Z = \left(\prod_{j=1}^{N} \int \mathrm{d}J_j \right) \delta \left(\sum_{j=1}^{N} J_j^2 - 1 \right) \exp \left[-\beta \sum_{\mu} \Theta(-E^{\mu}) \right] \tag{A.1}$$

where

$$E^{\mu} = (\boldsymbol{J} \cdot \boldsymbol{S}^{\mu} - \gamma) \,\mathrm{sign} F^{\mu} \qquad \text{with} \qquad F^{\mu} = (\boldsymbol{B} \cdot \boldsymbol{S}^{\mu} - \theta).$$

The quenched free energy $-1/\beta \, \langle \ln Z \rangle$ can be performed using the identity

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\partial}{\partial n} \langle Z^n \rangle \,.$$

For integer n, Z^n is the partition function of an n–fold replicated system . Its average can be calculated by means of a saddle point integration and involves the order parameters

$$R^a = \boldsymbol{J}^a \cdot \boldsymbol{B} \quad \text{and} \quad q_{ab} = \boldsymbol{J}^a \cdot \boldsymbol{J}^b \quad \text{for} \quad a \neq b, \tag{A.2}$$

here $a, b = 1, \ldots, n$ denote the replica indices.

In analogy to [6] we assume replica symmetry, i.e. $R^a = R$ for all a and $q_{ab} = q$ for all $a \neq b$. Within this simplifying scheme it is straightforward to identify the solution in the limit $n \to 0$. We obtain, for $\beta \to \infty$, the quenched free energy

$$\begin{aligned}\Gamma &= \frac{\partial}{\partial n}\bigg|_{n=0} \frac{\langle Z^n \rangle}{N} \\ &= \frac{q - R^2}{2(1-q)} + \frac{1}{2}\ln(1-q) + \alpha \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \int_{-\infty}^{\infty} \frac{\mathrm{d}F}{\sqrt{2\pi}} \exp\left[-\frac{(F+\theta)^2}{2}\right] \\ &\quad \times \ln\Phi\left[\frac{x\sqrt{q-R^2} + |F|R - (\gamma - \theta R)\,\mathrm{sign}F}{\sqrt{1-q}}\right]\end{aligned} \tag{A.3}$$

Further we restrict the analysis to the case $\gamma = \theta$. As discussed in the text, the relation $q = R$ is satisfied in this case and one obtains the simplified free energy

$$\begin{aligned}\Gamma = \frac{1}{2}[R + \ln(1-R)] + \alpha\sqrt{\frac{1-R}{R}} \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{\sqrt{2\pi}} \left[\exp\left(t - \frac{\Theta}{\sqrt{R}}\right) + \exp\left(t + \frac{\Theta}{\sqrt{R}}\right)\right] \\ \times \Phi\left(t\sqrt{\frac{R}{1-R}}\right) \ln\Phi\left(t\sqrt{\frac{R}{1-R}}\right).\end{aligned} \tag{A.4}$$

References

[1] Hertz J A, Krogh A and Palmer 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[2] Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press)
[3] Akay M 1992 *Biol. Cybernetics* **67** 361
[4] Moses L E, Shapiro D and Littenberg B 1993 *Statistics in Medicine* **12** 1293
[5] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys* **65** 499
[6] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* (Singapore: World Scientific) p 3
[7] Opper M and Kinzel W 1991 *Physics of Neural Networks* (Berlin: Springer) p 149
[8] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* **A 45** 6056
[9] Brenner H and Gefeller O 1997 *Statistics in Medicine* **16** 981
[10] Biehl M, Riegler P and Stechert M 1995 *Phys. Rev.* **E 52** 4626
[11] Opper M and Haussler D 1991 *Proc. of the 4th Annual Workshop on Computational Learning Theory* (Santa Cruz: Morgan Kaufman) p 75