Chapter 6
Multiple Testing and Binary Classification

Abstract We describe connections between multiple testing and binary classification. Under certain sparsity assumptions, classical multiple tests controlling a type I error rate at a fixed level α can, at least asymptotically as the number of classification trials tends to infinity, achieve the optimal (Bayes) classification risk. Under non-sparsity, combinations of type I and type II error rates are discussed as appropriate proxies for the (weighted) misclassification risk, and we provide algorithms for binary classification which are based on multiple testing. The problem of feature selection for binary classification is addressed by the higher criticism criterion, a concept originally introduced for testing the global null hypothesis in a multiple test problem.

Binary classification denotes the problem of assigning random objects to one of exactly two classes. This problem is often addressed by statistical learning techniques, see, for instance, Hastie et al. (2009) and Vapnik (1998) for introductions. Binary classification and multiple testing are related statistical fields. The decision pattern of a multiple test for a family of m hypotheses has the same structure as the output of a binary classifier for m data points to be classified, namely, a vector in $\{0, 1\}^m$ indicating the m binary decisions. Moreover, in both problems typically realizations x_i, $1 \leq i \leq m$, of random vectors X_i with values in $\mathbb{R}^k$ build the basis for the decision rule (the multiple test or the classifier) which is thus chosen according to statistical criteria. In the testing context, x_i has the interpretation of a data sample (or the value of a sufficient statistic) for the i-th individual test, while x_i is referred to as the i-th feature vector to be classified in the classification terminology. On the other hand, usual loss functions for binary classification differ from the ones that are typically utilized in multiple testing.

Definition 6.1. Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ denote stochastically independent and identically distributed random tuples, where X_i takes values in $\mathbb{R}^k$ and Y_i is a binary indicator with values in $\{0, 1\}$, $1 \leq i \leq m$. Let the data-generating process be modeled by a (joint) probability measure $\mathbb{P}$, where some systematic relationship between Y_1 and X_1 is assumed. Namely, the random vectors $X_1, \ldots, X_m$ are assumed

T. Dickhaus, *Simultaneous Statistical Inference*,
DOI: 10.1007/978-3-642-45182-9_6,

to be continuously distributed with class-conditional cdfs given by $F_j(x) = \mathbb{P}(X_i \leq x | Y_i = j)$ for $x \in \mathbb{R}^k, j = 0, 1$ and $i = 1, \ldots, m$. Assume that the "labels" $Y_1, \ldots, Y_m$ can not be observed. Formally, we describe the classification task by the pairs of hypotheses $H_i : Y_i = 0$ versus $K_i : Y_i = 1, 1 \leq i \leq m$.

(a) For a given cost parameter $c \in (0, 1)$ and a rejection region $\Gamma \subset \mathbb{R}^k$, the Bayes risk associated with the action $a_i = \mathbf{1}_\Gamma(x_i)$ is given by

$$R^{(i)}_{\text{Bayes}}(\Gamma) = (1 - c)\mathbb{P}(X_i \in \Gamma, Y_i = 0) + c\mathbb{P}(X_i \notin \Gamma, Y_i = 1). \tag{6.1}$$

Under the additive risk assumption, this entails that the Bayes risk for all m classification tasks together is given by

$$\begin{aligned} R_{\text{Bayes}}(\Gamma) &= \sum_{i=1}^{m} R^{(i)}_{\text{Bayes}}(\Gamma) \\ &= (1 - c)\mathbb{E}[V_m] + c\mathbb{E}[T_m], \end{aligned} \tag{6.2}$$

where the multiple testing error quantities V_m and T_m are as in Table 1.1 and refer to a multiple test with fixed rejection region Γ for every marginal test.

(b) Let a data-dependent classification rule be given by a measurable random mapping $\hat{h}_m : \mathbb{R}^k \to \{0, 1\}$, where we use the observed data $X_1 = x_1, \ldots, X_m = x_m$ to construct the rule $\hat{h}_m$. Then, the transductive and the inductive risk of $\hat{h}_m$, respectively, are given by

$$R^{(T)}(\hat{h}_m) = m^{-1} \sum_{i=1}^{m} \mathbb{P}(\hat{h}_m(X_i) \neq Y_i), \tag{6.3}$$

$$R^{(I)}(\hat{h}_m) = \mathbb{P}(\hat{h}_m(X_{m+1}) \neq Y_{m+1}), \tag{6.4}$$

where the tuple $(X_{m+1}, Y_{m+1}) \sim (X_1, Y_1)$ is stochastically independent of all (X_i, Y_i) for $1 \leq i \leq m$.

Similarly as in the Neyman-Pearson fundamental lemma, a set of best rejection regions $\Gamma = \Gamma_c$ considered in part (a) of Definition 6.1 is given by (see, for instance, Sect. 5.3.3 in Berger 1985)

$$\Gamma_c = \left\{ x \in \mathbb{R}^k : \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \leq c \right\} = \left\{ x \in \mathbb{R}^k : \lambda(x) \geq \frac{\pi_0(1 - c)}{\pi_1 c} \right\}, \tag{6.5}$$

where f_j is the pdf (or likelihood function) corresponding to $F_j, j = 0, 1$, $\lambda(x) = f_1(x)/f_0(x)$ denotes the likelihood ratio, and $\pi_0 = 1 - \pi_1 = \mathbb{P}(Y_1 = 0)$. The interpretation of part (b) of Definition 6.1 is that $\mathbb{P}$ is typically unknown or only partially known in practice and that the data-dependent classifier $\hat{h}_m$ "learns" a

rejection region Γ from the observed data. Notice that (at least for non-extreme values of c) the classification risk measures introduced in Definition 6.1 do not imply a fixed bound on a multiple type I error rate like the FWER or the FDR, but are of different type in the sense that type I and type II errors are treated (more or less) symmetrically. Weighting of type I and type II errors (i. e., misclassifying a member of the "0"-class as "1" and vice versa) is possible by choosing c appropriately. On the other hand, the data themselves implicitly also induce a weighting, namely by the relative frequencies of the true, but unobserved labels ($m_0 = |\{1 \le i \le m : y_i = 0\}|$ and $m_1 = m - m_0$).

6.1 Binary Classification Under Sparsity

From the preceding discussion, it becomes clear that multiple tests controlling a type I error rate like the FWER or the FDR at a fixed significance level are in general not good classifiers, because they treat null hypotheses and alternatives asymmetrically in the underlying risk criterion. Remarkable exceptions are sparse cases where class probabilities are highly unbalanced. Under sparsity, multiple tests can, at least asymptotically ($m \to \infty$), achieve optimal classification risks. As noted by Neuvial and Roquain (2012), the optimal rejection region Γ_c in (6.5) simplifies to a threshold for the data point x_i itself if $k = 1$ and the likelihood ratio λ is increasing in its argument x. The label $\hat{y}_i = 1$ is chosen if x_i exceeds a certain value. If a model for F_0 is available, this rule can equivalently be formalized by deciding $\hat{y}_i = 1$ if the p-value $p_i(x) = 1 - F_0(x_i)$ falls below the corresponding threshold on the p-value scale. This connects the theory of binary classification with that of p-value based multiple hypotheses testing that we have considered in the previous chapters of the present work.

It seems that Abramovich et al. (2006) were the first to realize that the linear step-up test φ^{LSU} from Definition 5.6, which has originally been developed for FDR control under independence, has remarkable properties with respect to a broad range of risk measures under sparsity assumptions, meaning that m_1 is small. While Abramovich et al. (2006) considered the particular problem of estimation under ℓ_r loss in high dimensions by employing thresholding estimators, their findings have also been the basis for studying classification risk properties of φ^{LSU} under sparsity. Bogdan et al. (2011) defined the concept of "Asymptotic Bayes optimality under sparsity" (ABOS) in a normal scale mixture model.

Definition 6.2. (Bogdan et al. (2011)). Under the assumptions of Definition 6.1, let $k = 1$. For the distribution of the independent observables $X_i : 1 \le i \le m$, consider the Bayesian model

$$\begin{aligned} X_i|\mu_i &\sim \mathcal{N}(\mu_i, \sigma_\varepsilon^2), \\ \mu_i &\sim \pi_0 \mathcal{N}(0, \sigma_0^2) + \pi_1 \mathcal{N}(0, \sigma_0^2 + \tau^2), \end{aligned}$$

where the μ_i, $1 \le i \le m$, are stochastically independent and σ_0^2 may be equal to zero. Hence, marginally, $X_i \sim \pi_0 \mathcal{N}(0, \sigma^2) + \pi_1 \mathcal{N}(0, \sigma^2 + \tau^2)$, with $\sigma^2 = \sigma_\varepsilon^2 + \sigma_0^2$. Assuming σ^2 to be known, the optimal rejection region Γ_c from (6.5) is such that $\hat{y}_i = 1$ if and only if $x_i^2/\sigma^2 \ge K^2$, where the cutoff K^2 is given by

$$K^2 = (1 + 1/u)\{\log(v) + \log(1 + 1/u)\}, \tag{6.6}$$

with $u = (\tau/\sigma)^2$ and $v = u(\pi_0/\pi_1)^2\delta^2$, $\delta = (1 - c)/c$. Denote the Bayes risk $R_{\text{Bayes}}(\Gamma)$, evaluated at this best rejection region, by R_{opt}. In practice, K^2 can typically not be computed exactly, because π_0 and/or τ^2 may be unknown. For a given multiple test procedure φ operating on $x_1, \ldots, x_m$, let $R_{\text{Bayes}}(\varphi)$ denote the risk functional defined in (6.2), with V_m and T_m now referring to φ. Assume that the model is such that

$$\pi_1 = \pi_1(m) \to 0,\ u = u(m) \to \infty,\ v = v(m) \to \infty,\ \text{and}\ \log(v)/u \to C \in (0, \infty), \tag{6.7}$$

as $m \to \infty$ (where convergence or divergence, respectively, may be along a subsequence indexed by $t = 1, 2, \ldots$). Notice that the dependence of v on m may imply that c depends on m, too. Then, φ is called asymptotically Bayes optimal under sparsity (ABOS), if

$$\frac{R_{\text{Bayes}}(\varphi)}{R_{opt}} \to 1, \quad t \to \infty. \tag{6.8}$$

It is clear that, under the conditions given in (6.7), eventually (for large m) the type I error component of the Bayes risk will dominate the type II error component, due to sparsity. Consequently, it turns out that classical multiple tests which are targeted towards type I error control are ABOS in the sense of Definition 6.2, at least for particular parameter configurations.

Theorem 6.1 (Bogdan et al. (2011)). *Under the model assumptions from Definition 6.2, the following assertions hold true.*

(a) Consider the Bonferroni test $\varphi^{Bonf} = (\varphi_i^{Bonf} : 1 \le i \le m)$ (cf. Example 3.1) operating on $x_1, \ldots, x_m$, the FWER level $\alpha = \alpha(m)$ of which fulfills $\alpha(m) \to \alpha_\infty \in [0, 1)$ such that $\alpha(m)/(1 - \alpha(m)) \propto (\delta\sqrt{u})^{-1}$. Then, φ^{Bonf} is ABOS if $\pi_1(m) \propto m^{-1}$. The condition imposed on α_m means that the Bayesian FDR (see Efron and Tibshirani (2002)) of φ^{Bonf} is proportional to α_m.

(b) Consider the linear step-up test φ^{LSU} from Definition 5.6 operating on $x_1, \ldots, x_m$, the FDR level $\alpha = \alpha(m)$ of which fulfills $\alpha(m) \to \alpha_\infty \in [0, 1)$ such that $\alpha(m)/(1 - \alpha(m)) \propto (\delta\sqrt{u})^{-1}$. Then, φ^{LSU} is ABOS whenever $\pi_1(m) \to 0$ such that $m\pi_1(m) \to s \in (0, \infty]$ as $m \to \infty$. In this sense, φ^{LSU} adapts to the unknown degree of sparsity in the data.

Neuvial and Roquain (2012) generalized the findings of Bogdan et al. (2011) concerning φ^{LSU} to a broader class of distributions of X_1. Namely, they assumed that

the (conditional) distribution of X_1 given $Y_1 = 0$ belongs to the parametric family considered by Subbotin (1923).

Definition 6.3. For a given shape parameter $\zeta \geq 1$, the distribution with Lebesgue density f_ζ, given by

$$f_\zeta(x) = \exp(-|x|^\zeta/\zeta)\{2\Gamma(1/\zeta)\zeta^{1/\zeta-1}\}^{-1}, \quad x \in \mathbb{R}, \tag{6.9}$$

is called ζ-Subbotin distribution.

The family of ζ-Subbotin distributions is closely related to the family of generalized error distributions (GEDs), cf., e.g., Nelson (1991) and references therein. In fact, the Lebesgue density of the GED with shape parameter equal to ζ is a scaled version of the ζ-Subbotin density f_ζ. In case of $\zeta = 2$, both distributions coincide with the standard normal. The 1-Subbotin distribution is equal to the Laplace (or double-exponential) distribution, while the GED with shape parameter equal to 1 has the same shape, but lighter tails.

Theorem 6.2 (Neuvial and Roquain (2012)). *Assume that the (conditional) distribution of X_1 on $\mathbb{R}$, given $Y_1 = 0$, is the ζ-Subbotin distribution with Lebesgue density f_ζ as in (6.9) and that the (conditional) distribution of X_1 given $Y_1 = 1$ is a shifted or scaled ζ-Subbotin distribution with Lebesgue density given by $f_{shift}(x) = f_\zeta(x - \mu_m)$ or $f_{scaled}(x) = f_\zeta(x/\sigma_m)/\sigma_m$, where $(\mu_m)_{m\in\mathbb{N}}$ or $(\sigma_m)_{m\in\mathbb{N}}$, respectively, is a sequence of unknown parameters. For all $m \in \mathbb{N}$, assume that μ_m or σ_m, respectively, is such that the density of the (random) p-value p_i corresponding to X_i has under $Y_i = 1$ a continuously decreasing Lebesgue density f_m, fulfilling $f_m(0^+) > \tau_m > f_m(1^-)$, where*

$$\tau_m = \frac{\pi_0(m)}{\pi_1(m)} = m^\beta, \quad 0 < \beta \leq 1.$$

Denoting the cdf corresponding to the p-value density f_m by F_m, assume that there exist constants C_- and C_+ such that $0 < C_- \leq F_m(f_m^{-1}(\tau_m)) \leq C_+ < 1$. Let the FDR level $\alpha = \alpha_m$ in the definition of φ^{LSU} be chosen such that $\alpha_m \to 0$ and $\log(\alpha_m) = o((\log m)^\gamma)$ as $m \to \infty$, where $\gamma = 1 - 1/\zeta$ for $\zeta > 1$ in case of shift alternatives and $\gamma = 1$ for $\zeta \geq 1$ in case of scale alternatives. Then, φ^{LSU} is asymptotically optimal in the sense that it fulfills

$$R_m(\varphi^{LSU}) \sim R_m^{opt}, \quad m \to \infty. \tag{6.10}$$

In (6.10), R_m is either one of the risk measures introduced in (6.3) and (6.4), and R_m^{opt} is the corresponding risk of the Bayes-optimal classifier with respect to R_m (which thresholds p-values at the fixed cutoff $f_m^{-1}(\tau_m)$).

In addition, Neuvial and Roquain (2012) derived exact convergence rates at which the relative excess risk $(R_m(\varphi^{LSU}) - R_m^{opt})/R_m^{opt}$ vanishes as $m \to \infty$. As in Theorem 6.1, also under the assumptions of Theorem 6.2 it turns out that φ^{LSU}, regarded as

a classifier, is highly adaptive to the amount of sparsity in the data, because the assertion holds true for any $0 < \beta \leq 1$. However, the fine-tuning of the nominal FDR level $\alpha = \alpha_m$ is a bottleneck in practice. Both under the model considered in Theorem 6.1 and under that considered in Theorem 6.2, even the (asymptotically) optimal order of magnitude of $\alpha = \alpha_m$ depends on unknown model parameters.

Remark 6.1.

(a) The risk measures $R^{(T)}(\hat{h}_m)$ and $R^{(I)}(\hat{h}_m)$ from part (b) of Definition 6.1 are defined without a weighting by a cost parameter c. As argued by Neuvial and Roquain (2012), the results of Theorem 6.2 remain to hold true if such a weighting is considered in $R^{(T)}(\hat{h}_m)$ and $R^{(I)}(\hat{h}_m)$.
(b) The sparsity assumptions regarding $\pi_1 = \pi_1(m)$ in Theorems 6.1 and 6.2 are appropriate for signal detection problems, where a small amount of signals (corresponding to $Y_i = 1$) is assumed within a huge amount of data points.
(c) Some further analytical results on FDR-controlled classification can be found in the works by Scott et al. (2009) and Genovese and Wasserman (2004). Cohen and Sackrowitz (2005a, b) studied the classes of single-step, step-down and step-up multiple tests with respect to admissibility and Bayes optimality in the classification context. In particular, they showed that step-up tests like φ^{LSU} are in general inadmissible under additive loss, meaning that uniformly better (and feasible) classification procedures exist, in particular in non-sparse models.

6.2 Binary Classification in Non-Sparse Models

For applications in which the class probabilities π_0 and π_1 are assumed to be (roughly) balanced, as for instance in brain-computer interfacing research that we will consider in Chap. 12, the Bayes risk decomposition given in (6.2) suggests to study multiple tests that control a weighted average of type I and type II error rates. In this direction, Storey (2003) pointed out that among the sets considered in (6.5) there is also a rejection region that minimizes the weighted average of the pFDR and its type II analogue, the positive false non-discovery rate (pFNR). This means, for a given weight parameter $w \in (0, 1)$ it exists a constant $c(w)$ such that

$$\min_{\Gamma \subset \mathbb{R}^k} (A(w)) = (1 - w) \cdot \mathrm{pFDR}(\Gamma_{c(w)}) + w \cdot \mathrm{pFNR}(\Gamma_{c(w)}), \text{ where} \quad (6.11)$$

$$A(w) = (1 - w) \cdot \mathrm{pFDR}(\Gamma) + w \cdot \mathrm{pFNR}(\Gamma). \quad (6.12)$$

Under the distributional assumptions of Definition 6.1, $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ are given by

$$\mathrm{pFDR}(\Gamma) = \mathbb{P}(H_0|X_1 \in \Gamma) = \mathbb{E}\left[\frac{V_m}{R_m}|R_m > 0\right],$$
$$\mathrm{pFNR}(\Gamma) = \mathbb{P}(H_1|X_1 \notin \Gamma) = \mathbb{E}\left[\frac{T_m}{W_m}|W_m > 0\right],$$

where V_m, R_m, T_m and W_m are again as in Table 1.1 and refer to a multiple test with fixed rejection region Γ for every marginal test. A particularly convenient scalability property is that $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ do not depend on m, in contrast to $\mathbb{E}[V_m]$ and $\mathbb{E}[T_m]$.

In practice, it remains to determine or at least to approximate the optimal cost parameter $c(w)$. Several possible methods for this have been discussed in the literature. As typical in the statistical learning context, many methods rely on utilizing a training sample $((x_i^{\mathrm{train}}, y_i^{\mathrm{train}}))_{1\le i\le m_{\mathrm{train}}}$ with known labels, where these training data points are assumed to be generated independently of $((X_i, Y_i))_{1\le i\le m}$ from the distribution $\mathbb{P}$. To this end, it is useful to notice that $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ can be computed in terms of the densities f_0 and f_1 by

$$\mathrm{pFDR}(\Gamma) = \frac{\pi_0 I_0(\Gamma)}{\pi_0 I_0(\Gamma) + \pi_1 I_1(\Gamma)},\ \mathrm{pFNR}(\Gamma) = \frac{\pi_1[1 - I_1(\Gamma)]}{\pi_1[1 - I_1(\Gamma)] + \pi_0[1 - I_0(\Gamma)]} \tag{6.13}$$

with $I_j(\Gamma) = \int_\Gamma f_j(\mathbf{u})\lambda^k(d\mathbf{u})$, $j = 0, 1$. Representation (6.13) shows that the Bayes risk defined in (6.1) can be regarded as a local version of the risk functional $A(w)$ from (6.12). Based on these considerations, in Sect. 7 of Storey (2003) the following algorithm for approximating $c(w)$ is outlined.

Algorithm 6.1

1. *Utilizing training data* $((x_i^{train}, y_i^{train}))_{1\le i\le m_{train}}$, *estimate the pdfs* f_0 *and* f_1 *by* $\hat{f}_j, j = 0, 1$.
2. *Approximate the sets* Γ_c *for given* $c \in (0, 1)$ *by plugging* $\hat{f}_j$ *into* (6.5) *instead of* f_j, $j = 0, 1$. *The prior probability* π_0 *can either be chosen explicitly or also be estimated from the training data.*
3. *Estimate* $pFDR(\Gamma_c)$ *and* $pFNR(\Gamma_c)$ *by numerical integration in* (6.13) *with* f_j *replaced by* $\hat{f}_j$, $j = 0, 1$.
4. *Choose* $w \in (0, 1)$ *and minimize the numerical approximation of* $(1 - w) \cdot pFDR(\Gamma_c) + w \cdot pFNR(\Gamma_c)$ *with respect to* c.

This approach automatically also delivers an estimate of the optimal rejection region $\Gamma_{c(w)}$, see the second step of the algorithm.

Remark 6.2.

(a) Actually, Storey (2003) describes a slightly different approach, namely, to estimate f_0 from training data drawn from the zero class and to estimate the marginal density $f = \pi_0 f_0 + \pi_1 f_1$ from possibly unlabeled data. This relates the statistical

model from Definition 6.1 also to the statistical learning task of semi-supervised novelty detection as in Blanchard et al. (2010).

(b) In contrast to the methods discussed in Sect. 6.1, Algorithm 6.1 is not restricted to feature vector dimensionality $k = 1$. In cases with $k > 1$, estimation methods for multivariate densities are applicable. Excellent textbook references for non-parametric density estimation are Silverman (1986) and Härdle et al. (2004).

Dickhaus et al. (2013) demonstrated that Algorithm 6.1 also works for stationary, but non-trivially auto-correlated feature vectors, at least under weak dependency assumptions. This generalization is important for the classification of multivariate time series data, cf. Chap. 12.

A second plausible approach for approximation of $c(w)$ in (6.11) relies on direct estimation of the likelihood ratio λ, which avoids plug-in of estimated densities. In a series of papers (cf. Sugiyama et al. (2009) and references therein), a group of Japanese researchers developed methods for and discussed applications of such direct estimation of density ratios. In Sects. 2.8 and 4 of Sugiyama et al. (2009), especially the so-called `uLSIF` algorithm is propagated. Hence, the following alternative algorithm has been investigated by Dickhaus et al. (2013), too.

Algorithm 6.2

1. *Utilizing training data* $((x_i^{train}, y_i^{train}))_{1 \le i \le m_{train}}$, *estimate the density ratio* λ *by* $\hat{\lambda}$.
2. *Approximate the set* Γ_c *for given* $c \in (0, 1)$ *by plugging* $\hat{\lambda}$ *instead of* λ *into the right-hand side of* (6.5). *The prior probability* π_0 *can either be chosen explicitly or be estimated from the training data.*
3. *Estimate* $pFDR(\Gamma_c)$ *and* $pFNR(\Gamma_c)$ *by calculating the relative frequencies of events* $\{y_i^{train} = 0\}$ *in the training sub-dataset with* $x_i^{train} \in \Gamma_c$ *and* $\{y_i^{train} = 1\}$ *in the training sub-dataset with* $x_i^{train} \notin \Gamma_c$, *respectively.*
4. *Choose* $w \in (0, 1)$ *and minimize the approximation of* $(1 - w) \cdot pFDR(\Gamma_c) + w \cdot pFNR(\Gamma_c)$ *with respect to* c.

The general finding of Dickhaus et al. (2013) was that Algorithm 6.1 seems to be more time-consuming, but that it had slightly better classification performance than Algorithm 6.2, both on computer-simulated and on real multivariate time series data. In the first step of Algorithm 6.1, the authors employed fixed-width kernel density estimators with Gaussian kernels and empirically sphered data, while in the first step of Algorithm 6.2 the proposed `uLSIF` algorithm of Sugiyama et al. (2009) was used.

An interesting direction for future research would be to study the general class of multiple testing based cost functions of the form

$$(1 - w) g_1(\mathbb{P}^{(V,R)}) + w g_2(\mathbb{P}^{(T,W)}),$$

where g_1 and g_2 are given functionals, with respect to binary classification in non-sparse models.

6.3 Feature Selection for Binary Classification via Higher Criticism

In cases where the feature vector dimension k is larger than 1, the explicit determination of the optimal rejection region Γ_c in (6.5) requires multivariate techniques. The presumably most well-known case is that of Fisher discrimination, meaning that the densities f_0 and f_1 are those of multivariate normal distributions on $\mathbb{R}^k$, $k > 1$, with common covariance matrix Σ, but class-specific mean vectors μ_0 and μ_1 (say). In this case, the Bayes-optimal rejection region is given by

$$\Gamma_c = \left\{ x \in \mathbb{R}^k : \left[x - \frac{1}{2}(\mu_1 + \mu_0) \right]^\top \Sigma^{-1}(\mu_1 - \mu_0) \geq \log\left(\frac{\pi_0(1-c)}{\pi_1 c} \right) \right\}. \tag{6.14}$$

This classification rule has a simple structure, because it is linear in the data. Hence, it is easy to apply in practice, provided that the parameters μ_j, $j = 0, 1$, and Σ are known. In case of unknown parameters, one typically estimates them from a training sample (cf. the first steps in Algorithms 6.1 and 6.2), leading to the so-called linear discriminant analysis (LDA). However, this approach causes severe issues if $k > m_{\text{train}}$, because in such cases the empirical covariance matrix is not invertible. The latter situation often occurs in modern life sciences, where typically a large set of features is at hand. Motivated by this example, Donoho and Jin (2008) were concerned with the problem of feature selection for classification based on multiple testing. Notice that this has close similarities to the problem of model selection that we will treat in Chap. 7.

Theorem 6.3 (Central limit theorem for order statistics). *Let $U_{1:k}, \ldots, U_{k:k}$ denote the order statistics of k stochastically independent, identically UNI$[0,1]$-distributed random variables $U_1, \ldots, U_k$. Let $q \in (0,1)$ be such that $i/k - q = o(k^{-1/2})$ as $k \to \infty$ for some integer-valued sequence $(i = i(k) : k \in \mathbb{N})$. Then, for all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\sqrt{k} \frac{U_{i:k} - q}{\sqrt{q(1-q)}} \leq t \right) \to \Phi(t), \quad k \to \infty.$$

Proof. See, for instance, Chap. 4 of Reiss (1989). □

Loosely formulated, the assertion of Theorem 6.3 means that for given $1 \leq i \leq k$, where k is large, $U_{i:k}$ is approximately normally distributed with mean i/k and variance $(i/k(1-i/k))/k$. It seems that John Wilder Tukey was the first who suggested to apply this result to multiple test problems with k marginal p-values which are under the global hypothesis H_0 distributed as $U_1, \ldots, U_k$ in Theorem 6.3, see Donoho and Jin (2004) and references therein.

Definition 6.4. (Higher criticism). Let $p_{1:k}, \ldots, p_{k:k}$ denote ordered marginal p-values for a multiple test problem $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H}_k)$. Then, the higher criticism

(HC) objective at index $1 \leq i \leq k$ is given by

$$HC(i, p_{i:k}) = \sqrt{k} \frac{i/k - p_{i:k}}{\sqrt{p_{i:k}(1 - p_{i:k})}}. \tag{6.15}$$

Alternatively and asymptotically equivalently, one may use i/k instead of $p_{i:k}$ in the denominator of $HC(i, p_{i:k})$, see Donoho and Jin (2008). For a given tuning parameter $\lambda \in (0, 1)$, the HC test statistic is given by

$$HC_k^* = \max_{1 \leq i \leq \lambda k} HC(i, p_{i:k}). \tag{6.16}$$

Asymptotic ($k \to \infty$) distributional results concerning HC_k^* have been derived by Donoho and Jin (2004). These results allow for utilizing HC_k^* as a test statistic for the global hypothesis H_0 in $\mathscr{H}_k$, provided that the number k of hypotheses is large. For the specific task of feature selection (where the number k of features is large), Donoho and Jin (2008) proposed the following algorithm.

Algorithm 6.3 *Under the assumptions of Definition 6.1, assume that $k >> 1$ and that a training sample $((x_i^{train}, y_i^{train}))_{1 \leq i \leq m_{train}}$ as described before Algorithm 6.1 is at hand. Furthermore, assume that there are some features (corresponding to components of the vector X_1) which are actually uninformative for the classification task. Then, selection of the informative features can be performed as follows.*

1. *For every feature $1 \leq j \leq k$, construct a statistic $Z_j : \mathbb{R}^{m_{train}} \times \{0, 1\}^{m_{train}} \to \mathbb{R}$ such that $Z = (Z_1, \ldots, Z_k)^\top$ is an (at least asymptotically) Gaussian random vector with stochastically independent components and mean vector $\mu = (\mu_1, \ldots, \mu_k)^\top$, where $\mu_j = 0$ if and only if feature j is uninformative for the classification task.*
2. *For all $1 \leq j \leq k$, compute the p-value p_j corresponding to the two-sided Z-test of the hypothesis $H_j : \{\mu_j = 0\}$ based on Z_j.*
3. *With these p-values, evaluate $HC(j, p_{j:k})$ for all $1 \leq j \leq k$, as well as HC_k^*, see Definition 6.4. Denote the index yielding the maximum in the definition of HC_k^* by j^*.*
4. *Select those features j for which $|Z_j|$ exceeds $|Z_{j^*}|$.*

Under certain assumptions regarding the asymptotic ($k \to \infty$) order of magnitude of the (common) mean of those random variables Z_j for which feature j is informative and the proportion of informative features, Donoho and Jin (2008) demonstrated (and outlined a rigorous proof) that Algorithm 6.3 leads to asymptotically optimal error rate classifiers.

Acknowledgments Parts of Sect. 6.2 originated from joint work with the Berlin brain-computer interface group, in particular with Benjamin Blankertz and Frank C. Meinecke. I thank Gilles Blanchard, Etienne Roquain and Masashi Sugiyama for fruitful discussions.

References

Abramovich F, Benjamini Y, Donoho DL, Johnstone IM (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann Stat 34(2):584–653, doi:10.1214/009053606000000074.

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer-Verlag, Springer Series in Statistics. New York etc.

Blanchard G, Lee G, Scott C (2010) Semi-supervised novelty detection. Journal of Machine Learning Research 11:2973–3009

Bogdan M, Chakrabarti A, Frommlet F, Ghosh JK (2011) Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. Ann Stat 39(3):1551–1579, doi:10.1214/10-AOS869.

Cohen A, Sackrowitz HB (2005a) Characterization of Bayes procedures for multiple endpoint problems and inadmissibility of the step-up procedure. Ann Stat 33(1):145–158, doi:10.1214/009053604000000986.

Cohen A, Sackrowitz HB (2005b) Decision theory results for one-sided multiple comparison procedures. Ann Stat 33(1):126–144, doi:10.1214/009053604000000968.

Dickhaus T, Blankertz B, Meinecke FC (2013) Binary classification with pFDR-pFNR losses. Biom J 55(3):463–477, doi:10.1002/bimj.201200054.

Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat 32(3):962–994, doi:10.1214/009053604000000265.

Donoho D, Jin J (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proc Natl Acad Sci USA 105(39):14,790–14,795.

Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 23(1):70–86

Genovese C, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061, doi:10.1214/009053604000000283.

Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and semiparametric models. Springer, Springer Series in Statistics. Berlin

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Data mining, inference, and prediction. 2nd ed. Springer Series in Statistics. New York, NY: Springer., doi:10.1007.b94608

Nelson DB (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59(2):347–370, doi:10.2307/2938260.

Neuvial P, Roquain E (2012) On false discovery rate thresholding for classification under sparsity. Ann Stat 40(5):2572–2600

Reiss RD (1989) Approximate distributions of order statistics. Springer, With applications to nonparametric statistics. Springer Series in Statistics. New York etc.

Scott C, Bellala G, Willett R (2009) The false discovery rate for statistical pattern recognition. Electronic Journal of Statistics 3:651–677

Silverman B (1986) Density estimation for statistics and data analysis. Chapman and Hall, Monographs on Statistics and Applied Probability. London - New York

Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat 31(6):2013–2035

Subbotin MT (1923) On the law of frequency of errors. Matematicheskii Sbornik 31(2):296–301

Sugiyama M, Kanamori T, Suzuki T, Hido S, Sese J, Takeuchi I, Wang L (2009) A Density-ratio Framework for Statistical Data Processing. IPSJ Transactions on Computer Vision and Application 1:183–208

Vapnik VN (1998) Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Chichester: Wiley.