
Bayes Risk Consistency of Large Margin Binary Classification Methods - A Survey

Paramveer S. Dhillon

CIS, University of Pennsylvania, Philadelphia, PA, U.S.A
pasingh@seas.upenn.edu

Abstract

In this paper we survey the Bayes Risk Consistency of various Large Margin Binary Classifiers. Many classification algorithms minimize a tractable convex surrogate ϕ of the 0-1 loss function. For e.g. the SVM (Support Vector Machine) and AdaBoost, which minimize the hinge loss and the exponential loss respectively. By imposing some sort of regularization conditions, it is possible to demonstrate the Bayes-risk consistency of methods based on minimizing the convex surrogate of the intractable 0-1 loss function. We have surveyed four papers which present such results in the binary classification setting, namely [BJM04], [Jia00], [Zha04] and [LV04].

1 Introduction

We consider the binary classification problem in its standard probabilistic setting i.e. n points are drawn from a probability distribution on $\mathcal{X} \times \mathcal{Y}$, where $Y \in \{\pm 1\}$ is the response variable and $X \in \mathcal{X}$ are the predictors. The aim is to find a function $f : \mathcal{X} \mapsto \mathbb{R}$ that accurately predicts the binary response variable Y , using the covariate X , such that $R(f) = \mathbb{E}\ell(Yf(X))$, the risk of the thresholded function, is minimized. Here $\ell(z)$ is the indicator function of the event $z \leq 0$. Large margin classification methods use some convex loss function $\phi : \mathbb{R} \mapsto \mathbb{R}$, and find a function f from some class \mathcal{F} that minimizes the ϕ -risk, $R_\phi(f) = \mathbb{E}\phi(Yf(X))$, i.e. the expected loss evaluated at the margin $Yf(X)$. These methods minimize the empirical ϕ -risk, $\hat{R}_\phi(f)$ or a regularized version. Methods like the Support Vector Machines (SVM) and AdaBoost fall in this category. They minimize the Hinge loss or the Exponential loss respectively instead of their non-convex counterparts.

However, the aim of any classification method is to find a function f whose probability of misclassification $R(f)$ (i.e. the risk) is close to the minimum possible risk i.e. (Bayes Risk R^*). It becomes important to investigate the conditions which guarantee that if the ϕ -risk of f gets close to the optimal then the risk of f also approaches the Bayes Risk R^* . If this condition can be realized then we say that the classification method based on ϕ is *Bayes Consistent*.

Bayes risk consistency of various large margin binary classification methods that use surrogate losses has been extensively studied in literature in a variety of settings by ([BJM04], [Jia00], [LV04] and [Zha04]).

2 Basic Methodology used in proving Consistency Results

There are three key steps in proving the consistency results of this kind. Firstly, the proof should have a theorem which relates the excess risk $R(f) - R^*$ to the excess ϕ -risk, $R_\phi(f) - R_\phi^*$, where R^* is the Bayes Risk, which is infimum over all measurable f of $R(f)$ and $R_\phi^* = \inf_f R_\phi(f)$ is the analogous quantity for the ϕ -risk. In [BJM04] they introduce the concept of “classification calibration” to relate the two risks, and similarly in the other three papers also i.e. [Jia00], [LV04], [Zha04] the

authors relate these risks using explicit inequalities or otherwise. The second step involves showing that the functions used by the method are rich enough to approximate f_ϕ^* , which is the measurable function that minimizes the ϕ - risk. This involves showing that

$$\bigcup_{\lambda>0} \mathcal{F}_k(\mathcal{H}, \lambda), \bigcup_{\lambda>0} \mathcal{F}_b(\mathcal{G}, \lambda) \text{ and } \bigcup_{k>0} \mathcal{F}_b(k)$$

where \mathcal{H} is the RKHS (Reproducing Kernel Hilbert Space) of functions on \mathcal{X} and $\mathcal{G} \in \{\pm 1\}^{\mathcal{X}}$ has finite VC Dimension. The authors of the papers [Zha04], [LV04], [Jia00] respectively show that the above are sufficiently rich.

The third and the last step is to chose a sequence of suitably restricted subsets $\mathcal{F}_n \subseteq \mathcal{F}$, as a function of the sample size n , so that the ϕ - risk of the estimated $\hat{f}_n \in \mathcal{F}_n$ converges to its infimum i.e. $\inf_{f \in \mathcal{F}_n} R_\phi(f)$. The authors of [Jia00], [LV04] and [Zha04], define \mathcal{F}_n as the set of combinations of k_n functions from $\mathcal{G} \in \{\pm 1\}^{\mathcal{X}}$, or the set of combinations of functions from \mathcal{G} with the coefficient vector having one - norm no more than λ_n , or a ball of radius λ_n in an RKHS (Reproducing Kernel Hilbert Space) \mathcal{H} . The third step is a little involved in the [Jia00], since their result involves an algorithm that does not minimize an objective function involving the empirical risk.

Now, we consider the properties of the loss function ϕ that allow comparison theorems, and hence consistence results for “large margin” methods in general. The result in [Jia00] is for the exponential loss function $\phi(\alpha) = e^{-\alpha}$, and their proof uses a smoothness assumption on the join probability distribution that ensures that the optimal f_ϕ^* is continuous. Lugosi et. al. assume that ϕ is differentiable, strictly convex and monotonic. Zhang assumes that ϕ satisfies three conditions:

1. ϕ is convex
2. The minimal conditional ϕ - risk,

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)) \quad (1)$$

decreases polynomially with $|\frac{1}{2} - \eta|$.

3. For any $\eta \neq \frac{1}{2}$, any minimizer α^* of the conditional ϕ - risk, $\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$ has the same sign as $\eta - \frac{1}{2}$. Thus, a pointwise minimization of the conditional ϕ - risk leads to a function that gives the correct sign everywhere.

3 The Main Results

In this section we enumerate and discuss the main results of the four papers [BJM04], [Jia00], [LV04], [Zha04].

We can categorize the results in the papers, based on the loss function ϕ , the class of functions \mathcal{F} , and the algorithm used to minimize R_ϕ as follows:

1. The consistency result of [Zha04] applies to several loss functions, and concerns kernel methods, which chose a function f from the RKHS \mathcal{H} of functions on \mathcal{X} to minimize a regularized empirical ϕ - risk,

$$\hat{R}_\phi(f) + C\|f\|_{\mathcal{H}} \quad (2)$$

where $\|\cdot\|_{\mathcal{H}}$ is the Hilbert Space norm. It is indeed equivalent to choosing f from the function class

$$\mathcal{F}_k(\mathcal{H}, \lambda) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\} \quad (3)$$

so as to minimize $\hat{R}_\phi(f)$.

2. The result of [LV04] considers the exponential loss function

$$\phi(\alpha) = e^{-\alpha} \quad (4)$$

They chose the function class

$$\mathcal{F}_b(\mathcal{G}, \lambda) = \left\{ \sum_i \alpha_i g_i : \|\alpha\|_1 \leq \lambda, g_i \in \mathcal{G} \right\} \quad (5)$$

where $\mathcal{G} \in \{\pm 1\}^{\mathcal{X}}$ has finite VC Dimension, and an algorithm that minimizes the empirical ϕ - risk.

3. The authors of [Jia00] also consider the exponential loss function as above , the function class

$$\mathcal{F}_b(k) = \left\{ \sum_{i=1}^k \alpha_i g_i \text{ ; } g_i \in \mathcal{G} \right\} \quad (6)$$

and the AdaBoost algorithm, which chooses α_i, g_i sequentially, to greedily minimize the empirical ϕ - risk.

3.1 Classification Calibration

In the paper [BJM04], the authors talk about the concept of “classification claibration” in order to relate the excess risk to the excess ϕ - risk. Their methodology allows us to find quantitative relationship between the risk associated with ϕ and the risk associated with 0 - 1 loss. We will describe their methodology in some detail below.

They begin by defining the following functional transform of the loss function ϕ :

Definition 1: Given a loss function $\phi: \mathbb{R} \mapsto [0, \infty)$, define the function $\tilde{\psi}: [0, 1] \mapsto [0, \infty)$ by

$$\tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right) \quad (7)$$

where,

$$\begin{aligned} H(\eta) &= \inf_{\alpha \in \mathbb{R}} (\eta\phi(\alpha) + (1-\eta)(\phi(-\alpha))) \\ H^-(\eta) &= \inf_{\alpha: \alpha(2\eta-1) \leq 0} (\eta\phi(\alpha) + (1-\eta)(\phi(-\alpha))) \end{aligned} \quad (8)$$

The ψ - transform is defined to be the function $\psi: [0, 1] \mapsto [0, 1)$ that is convex closure of $\tilde{\psi}$.

Theorem 1: For any non - negative loss function ϕ , any measurable $f: \mathcal{X} \mapsto \mathbb{R}$ and any probability dsitribution on $\mathcal{X} \times \{\pm 1\}$,

$$\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^* \quad (9)$$

This theorem establishes a general quantitative relationship between the excess ϕ - risk and the excess risk. But, for this relationship to be useful, it needs to be shown that ψ has particular properties, namely that pointwise minimization of the conditional ϕ - risk leads to a function that gives the correct sign. We can express this condition in the following way:

Definition 2: ϕ is classification calibrated, if for any $\eta \neq \frac{1}{2}$

$$H^-(\eta) > H(\eta) \quad (10)$$

Equivalently, ϕ is classification calibrated if for any sequence (α_i) such that $\lim_{i \rightarrow \infty} \{\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\} = H(\eta)$, we have $\lim_{i \rightarrow \infty} \text{sign}(\alpha_i(\eta - 1/2)) = 1$. In particular, if the infimum is achieved at a minimizing value α^* , then this value must have the correct sign. Therefore, this condition is essentially an extension of Theorem 2.1 in [Zha04] and can be viewed as a form of Fisher consistency that is appropriate for classification. Thus we can conclude the following, which are infact important results in the [BJM04] paper.

Theorem The following conditions are equivalent as mentioned in [BJM04]:

1. ϕ is classification calibrated.
2. For any sequence (θ_i) in $[0, 1]$, $\psi(\theta_i) \mapsto 0$ iff $\theta_i \mapsto 0$.
3. For every sequence of measurable functions $f_i: \mathcal{X} \mapsto \mathbb{R}$ and every probability distribution P ,

$$R_\phi(f_i) \mapsto R_\phi^* \text{ implies } R(f_i) \mapsto R^* \quad (11)$$

Hence, we can see that we obtain a general comparison theorem under the weakest possible condition on the loss function ϕ . Also, it can be shown that for a given ϕ , the ψ - transform is optimal as the bound given in Theorem 1 cannot be improved in general, everywhere in its domain.

Also, the formulation of “classification calibration” does not assume that ϕ is convex. If instead we assume that ϕ is convex then the function $\tilde{\psi}$ in Definition 1 is necessarily closed and convex, and therefore the ψ - transform is specified directly via $\psi(\theta) = H^{-}(\frac{1+\theta}{2}) - H(\frac{1+\theta}{2})$. Also, if ϕ is convex, then it is possible to show that it is classification calibrated iff it is differentiable at 0 and $\phi'(0) < 0$.

The comparison theorem i.e. Theorem 1, and its analogous theorems in the other three papers [Jia00], [LV04], [Zha04], which relate the excess risk to the excess ϕ - risk, suggest a general framework for studying pattern classification methods that involve a surrogate loss function. Also, it is quite common to view the excess risk as a combination of an estimation term and an approximation term:

$$R(f) - R^* = (R(f) - \inf_{g \in \mathcal{F}} R(g)) + (\inf_{g \in \mathcal{F}} R(g) - R^*) \quad (12)$$

But, it turns out that the task of choosing a function with minimal risk over a class \mathcal{F} , is equivalent to the problem of minimizing empirical risk, which is computationally infeasible for typical classes of \mathcal{F} that we are interested in. For the function classes typically used by the algorithms in [Jia00], [LV04], [Zha04] i.e. boosting and kernel methods, the estimator term (the first term in the above equation) does not converge to zero for the minimizer of the empirical risk. On the other hand, the comparison theorems that we are considering in this survey suggest splitting the upper bound on excess risk into an estimation and an approximation term:

$$\psi(R(f) - R^*) \leq (R_\phi(f) - R_\phi^*) = (R_\phi(f) - \inf_{g \in \mathcal{F}} R_\phi(g)) + (\inf_{g \in \mathcal{F}} R_\phi(g) - R_\phi^*) \quad (13)$$

We can view the function ψ provided by the comparison theorem (“classification calibration”) as quantifying the penalty incurred by using a surrogate convex loss function ϕ in place of the 0 - 1 loss, and linking the excess risk to the approximation error and estimated error associated with ϕ - risk. In many cases it is possible to minimize the ϕ - risk efficiently over a convex class \mathcal{F} , and hence find an $f \in \mathcal{F}$ for which the upper bound on risk is near minimal. This holds despite the fact that finding an $f \in \mathcal{F}$ with near minimal risk is computationally infeasible.

3.2 Main Theorems of [LV04], [Jia00], [Zha04]

The main Theorems in [Zha04] that relate the excess risk to the excess ϕ - risk and then prove the Bayes Risk consistency are Definition 2.1 and Theorem 2.1 in their paper. They are stated below.

Definition: We define function $f_\phi^*(\eta) : [0, 1] \mapsto \mathbb{R}^*$ as:

$$f_\phi^*(\eta) = \arg \min_{f \in \mathbb{R}^*} Q(\eta, f),$$

$$Q^*(\eta) = \inf_{f \in \mathbb{R}} Q(\eta, f) = Q(\eta, f_\phi^*(\eta))$$

where $Q(\eta, f) = \eta\phi(f) + (1 - \eta)\phi(-f)$, $\phi(f)$ being the loss function. The Loss functions that the authors consider, and the corresponding values of $f_\phi^*(\eta)$ and $Q^*(\eta)$ are as follows:

- *Least Squares:* $f_\phi^*(\eta) = 2\eta - 1$; $Q^*(\eta) = 4\eta(1 - \eta)$.
- *Modified Least Squares:* $f_\phi^*(\eta) = 2\eta - 1$; $Q^*(\eta) = 4\eta(1 - \eta)$.
- *SVM:* $\text{sign}(2\eta - 1)$; $Q^*(\eta) = 1 - |2\eta - 1|$.
- *Exponential:* $f_\phi^*(\eta) = \frac{1}{2} \ln \frac{\eta}{1-\eta}$; $Q^*(\eta) = 2\sqrt{\eta(1 - \eta)}$
- *Logistic:* $f_\phi^*(\eta) = \frac{1}{2} \ln \frac{\eta}{1-\eta}$; $Q^*(\eta) = -\eta \ln(\eta) - (1 - \eta) \ln(1 - \eta)$

Next the authors provide the inequality that relates the excess risk to the excess ϕ - risk.

Theorem: Assume $f_\phi^*(\eta) > 0$ when $\eta > 0.5$. Assume there exists $c > 0$ and $s \geq 1$ such that for all $\eta \in [0, 1]$,

$$|0.5 - \eta|^s \leq c^s \Delta Q(\eta, 0), \quad (14)$$

then for any measurable function $f(x)$:

$$L(f(\cdot)) \leq L^* + 2c\Delta Q(f(\cdot))^{1/s}, \quad (15)$$

where L^* is the optimal Bayes error: $L^* = L(2\eta(\cdot) - 1)$, and $L(f(\cdot))$ is the classification error of a predictor f .

The above theorem implies that if we obtain ϕ by approximately minimizing

$$Q(f(\cdot)) = \mathbf{E}_{\mathbf{X}}[\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))] \quad (16)$$

where $\eta(x)$ denotes the conditional probability $P(Y = 1|X = x)$, and $\mathbf{E}_{\mathbf{X}}$ denotes the expectation over input data \mathbf{X} , so that $\Delta Q(f(\cdot))$ is small, then the classification error rate of $f(\cdot)$ is close to that of the Bayes error rate $L^* = L(f_{\phi}^*(\eta(\cdot)))$. In particular, if we can find a sequence of predictors $f_k(\cdot) \in C$ such that $\Delta Q(f_k(\cdot)) \mapsto 0$, then we are able to achieve classification error rate arbitrarily close to that of the Bayes error rate.

The authors of [LV04], also provide a similar Lemma (Lemma 5), that relates the excess risk to the excess ϕ risk. Before describing the lemma, we will need to describe some assumptions and a related theorem, from which the Lemma will follow.

Assumption : Let ϕ be a differentiable strictly convex, strictly increasing cost function such that $\phi(0) = 1$, $\lim_{x \mapsto -\infty} \phi(x) = 0$

Theorem : Assume that the cost function ϕ satisfies the above assumption and the distribution of (X, Y) and the class C are such that

$$\lim_{\lambda \mapsto \infty} \inf_{f \in \lambda \cdot \mathcal{F}} A(f) = A^*, \quad (17)$$

where $A^* = \inf A(f)$ over all measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$. Assume that C has a finite VC Dimension.

Let $\lambda_1, \lambda_2, \lambda_3, \dots$ be a sequence of positive numbers satisfying

$$\lambda_n \mapsto \infty \text{ and } \lambda_n \phi'(\lambda_n) \sqrt{\frac{\ln(n)}{n}} \mapsto 0 \text{ as } n \mapsto \infty \quad (18)$$

and define the estimator $f_n = \hat{f}_n^{\lambda_n} \in \mathcal{F}$. Then g_{f_n} is strongly Bayes Risk consistent, that is:

$$\lim_{n \mapsto \infty} L(g_{f_n}) = L^* \text{ almost surely.} \quad (19)$$

Now, we are in a position to state the main lemma of [LV04].

Lemma: Let ϕ be a cost function satisfying the above assumption. Let f_n be an arbitrary sequence of functions such that

$$\lim_{n \mapsto \infty} A(f_n) = A^* \quad (20)$$

Then the classifier

$$g_{f_n}(x) = \begin{cases} 1, & \text{if } f_n(x) > 0, \\ -1, & \text{otherwise} \end{cases} \quad (21)$$

has a probability of error converging to L^* , where L^* is the optimal Bayes error rate. The consistency result now follows from the above Lemma.

In [Jia00], the authors particularly consider the consistency of AdaBoost. The results of this paper are a bit involved, so first we will describe the setting that they use, then we will describe their main results.

Notation and the basic setting in [Jia00]: $(X_i, Z_i)_1^n$ and (\mathbf{X}, \mathbf{Z}) are i.i.d random quantities valued in $[0, 1]^d \times \{\pm 1\}$. \mathbf{H} is a base hypothesis space which is a set of functions $f: [0, 1]^d \mapsto \{\pm 1\}$. $C_n(F) = n^{-1} \sum_1^n e^{-Z_i F(X_i)}$ i.e. the AdaBoost cost function and $C_{\infty}(F) = E e^{-ZF(X)}$ (the population version). For each $n = 1, 2, 3, \dots, \infty$, F_n^t are the sequential fits. They are used to describe the sample version and the population version (for $n = \infty$) of the AdaBoost algorithm.

They use the following regularity conditions. Conditions 1, 2 are on the joint distribution of X, Y and 3, 4 and 5 are on the base hypothesis space H .

Condition 1: (Distribution of Predictor). The distribution of X is assumed to be absolutely continuous with respect to the Lebesgue Measure on $[0, 1]^d$.

Condition 2: (Continuity of Log Odds). The function $F_B(\cdot)$ is continuous on $[0, 1]^d$.

Condition 3: (Completeness of base hypothesis space). The linear span of H is complete on $L_2(P_X)$ on $[0, 1]^d$. That is, for any function g such that $\|g\|_{L_2(P_X)} \equiv \sqrt{\int_{[0,1]^d} g(x)^2 P_X dx} < \infty$ and for any $\epsilon > 0$, there exists a linear combination $\sum_{s=1}^m \alpha_s f_s(\cdot)$ for some $m \in 1, 2, \dots$ such that $f_s \in H$ and $\alpha_s \in \mathbb{R}$ and $\|g(x) - \sum_{s=1}^m \alpha_s f_s(x)\|_{L_2(P_X)} < \epsilon$.

Condition 4: (Finite VC Dimension of base hypothesis space). The VC dimension of H is finite, that is, $VC(H) < \infty$.

Condition 5: (Compactness of Base Hypothesis space) The base hypothesis space H is a compact set of ± 1 -valued functions on $[0, 1]^d$ in the $L_2 P(X)$ metric.

We can briefly summarize the above conditions and their need as follows:

1. Conditions 1 - 3 ensure that for any sequence F_∞^t of fits from the population version of AdaBoost, we can have $\lim_{t \rightarrow \infty} \|F_\infty^t - F_B\|_{L_2(P_X)} = 0$. This guarantees that at $n = \infty$ AdaBoost does the right thing and gives the optimal Bayes Prediction.
2. Conditions 4 - 5 are used to prove that the large - n case is close to the population version at finite time t . Condition 4 typically holds, but Condition 5 holds in situations when H is “continuously parametrized on a compact space”.
3. Condition 2 also implies that the coefficients of the population AdaBoost are well defined, that is $|\alpha_\infty^s| < \infty$ for all s .
4. The proof of the final theorem is non - constructive and we do not know what rate t_n can take. Presumably some t_n that increases to ∞ quite slowly will work. This is so, as $t_n \mapsto \infty$, the “approximation error” is related to $\|F_\infty^{t_n} - F_B\|_{L_2(P_X)}$, which goes to zero. On the other hand, if the growth of t_n is slow enough, the sample AdaBoost fit $F_\infty^{t_n}$ is sufficiently close to the population version $F_\infty^{t_n}$ for large n .

Now we are in a position to state the main theorem of this paper, which is as follows:

Theorem: (Process Consistency for AdaBoost). Under conditions 1 - 5, there exists a sequence $t = t_n$ for AdaBoost fits (sequential) F_n^t such that $\lim_{n \rightarrow \infty} E_S L_\infty(F_n^{t_n}) = L_\infty(F_B)$ and hence $\lim_{n \rightarrow \infty} \inf_t E_S L_\infty(F_n^t) = L_\infty(F_B)$ also.

4 Conclusions

The results in the papers [BJM04], [Jia00], [LV04], [Zha04] are really interesting and raise some very interesting questions. One interesting question raised especially by [Zha04] is that of convergence rates. Their paper makes a start in this direction for kernel methods. Few other people, namely Tsybakov (2001) have considered empirical risk minimization in pattern classification problems with low - noise, specifically where P_X i.e. the probability that $P(Y = 1|X)$ is nearly $\frac{1}{2}$ is small. They showed that the risk of the empirical minimizer converges to the minimal value surprisingly quickly in these cases. It turns out that, under Tsybakov’s low noise conditions, the relation between excess risk and excess ϕ - risk can be improved using the concept of “classification calibration” [BJM04], as we discuss earlier in this paper. In that case, if the loss function ϕ is uniformly convex and \mathbb{F} is convex, then the excess risk converges to its minimal.

We can conclude by saying that the problem of classification has been a fruitful domain in which to explore connections between statistical and computational science. Efficient algorithms can be designed to solve large - scale classification problems by exploiting tools from convex optimization and the statistical consequence of using these tools are beginning to be understood. The papers by [Jia00],[LV04],[Zha04] represent significant progress on the general problem of incorporating

considerations of computational complexity in statistical theory, providing hints of general tradeoffs between statistical accuracy and computational resources that are only beginning to be explored

References

- [BJM04] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Large margin classifiers: convex loss, low noise, and convergence rates. In *Advances in Neural Information Processing Systems*, 16, 2004.
- [Jia00] Wenxin Jiang. Process consistency for adaboost. *Annals of Statistics*, 32:13–29, 2000.
- [LV04] Gábor Lugosi and Nicolas Vayatis. On the bayes risk consistency of regularized boosting methods. *Annals of Statistics*, 32:30–55, 2004.
- [Zha04] Tong Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85, 2004.