

Classification with a reject option

Marten Wegkamp

Department of Statistics
Florida State University

Cambridge, 8 May 2008

Research is supported by NSF grant DMS 0706829

1

Bibliography

2

Introduction

- Binary classification model
- Reject option
- Notation and definitions
- Bayes rule with reject option

3

Plug-in rules

- Definition
- Theorem

4

Empirical Risk Minimizers




- Introduction
- Convex surrogate
- Notation and definitions
- Condition 1
- Structural risk minimization
- Oracle target
- Condition 2
- Oracle inequality
- Choice of the tuning parameter r_n

5




SVM classifier with reject option

- Hinge loss
- Generalized hinge loss
- Classification calibrated
- Excess risk comparison
- Margin condition
- Oracle inequality

Talk based on:

-  P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* (accepted)
-  R. Herbei and M. Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 4(4): 709–721, 2006.
-  M. Wegkamp. Lasso type classifiers with a reject option. *Electronic Journal of Statistics*, 1, 155–168 (2007).

Additional literature:

-  B. Tarigan and S. van de Geer. Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli*, 12(6): 1045–1076, 2006.
-  A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32: 135–166, 2004.
-  S. van de Geer. High dimensional generalized linear models and the Lasso. *Annals of Statistics* (in press).

Bibliography
Introduction
Plug-in rules
Empirical Risk Minimizers
SVM classifier with reject option

Binary classification model
Reject option
Notation and definitions
Bayes rule with reject option

INTRODUCTION

Binary classification model

Independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where

- $X_i \in \mathcal{X}$ (feature space)
- $Y_i \in \{-1, 1\}$ (labels)

Binary classification model

Independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where

- $X_i \in \mathcal{X}$ (feature space)
- $Y_i \in \{-1, 1\}$ (labels)

A discriminant function $f : \mathcal{X} \rightarrow \mathbb{R}$

Binary classification model

Independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where

- $X_i \in \mathcal{X}$ (feature space)
- $Y_i \in \{-1, 1\}$ (labels)

A discriminant function $f : \mathcal{X} \rightarrow \mathbb{R}$ produces a classifier

$$\text{sign}(f(x)).$$

Binary classification model

Independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where

- $X_i \in \mathcal{X}$ (feature space)
- $Y_i \in \{-1, 1\}$ (labels)

A discriminant function $f : \mathcal{X} \rightarrow \mathbb{R}$ produces a classifier

$$\text{sign}(f(x)).$$

It errs if margin $yf(x) < 0$.

Reject option

If the conditional probability

$$\eta(x) = \mathbb{P}\{Y = 1|X = x\}$$

is close to $1/2$, then we might just as well toss a coin to make a decision.

Reject option

If the conditional probability

$$\eta(x) = \mathbb{P}\{Y = 1|X = x\}$$

is close to $1/2$, then we might just as well toss a coin to make a decision.

This motivates the introduction of a *reject option* for classifiers, by allowing for a third decision, \mathbb{R} (*reject*), expressing doubt.

Although classifiers with a reject option are valuable in practice, few theoretical results are available in the statistical literature (Ripley 1996, Herbei and Wegkamp 2006, Wegkamp 2007, Bartlett and Wegkamp, forthcoming).

Although classifiers with a reject option are valuable in practice, few theoretical results are available in the statistical literature (Ripley 1996, Herbei and Wegkamp 2006, Wegkamp 2007, Bartlett and Wegkamp, forthcoming).

In the engineering community on the other hand this option is more common and empirically shown to effectively reduce the misclassification rate (Chow 1970, Fumera and Roli 2002, 2004, Fumera et al 2000, Golfarelli et al 1997, Hansen et al 1997).

We want to construct classifiers with a built-in reject option.

We want to construct classifiers with a built-in reject option.

We will study two types of classifiers

- Plug-in rules

We want to construct classifiers with a built-in reject option.

We will study two types of classifiers

- Plug-in rules (easy to compute)

We want to construct classifiers with a built-in reject option.

We will study two types of classifiers

- Plug-in rules (easy to compute)
- Empirical risk minimizers

We want to construct classifiers with a built-in reject option.

We will study two types of classifiers

- Plug-in rules (easy to compute)
- Empirical risk minimizers (especially useful for high dimensional data)

Notation and definitions

- τ threshold value

Notation and definitions

- τ threshold value
- Report $\text{sign}(f(x)) \in \{\pm 1\}$ if $|f(x)| > \tau$

Notation and definitions

- τ threshold value
- Report $\text{sign}(f(x)) \in \{\pm 1\}$ if $|f(x)| > \tau$
- Withhold decision and report \textcircled{R} if $|f(x)| \leq \tau$

Notation and definitions

- τ threshold value
- Report $\text{sign}(f(x)) \in \{\pm 1\}$ if $|f(x)| > \tau$
- Withhold decision and report \textcircled{R} if $|f(x)| \leq \tau$
- Cost of making a wrong decision is 1 and that of utilizing the reject option is $d > 0$

Notation and definitions

- τ threshold value
- Report $\text{sign}(f(x)) \in \{\pm 1\}$ if $|f(x)| > \tau$
- Withhold decision and report \textcircled{R} if $|f(x)| \leq \tau$
- Cost of making a wrong decision is 1 and that of utilizing the reject option is $d > 0$
- Appropriate risk
 $R(f) = d\mathbb{P}\{|Yf(X)| \leq \tau\} + \mathbb{P}\{Yf(X) < -\tau\}$, the appropriate risk.

We see that

$$R(f) = \mathbb{E} [\ell(Yf(X))]$$

for the discontinuous loss

$$\ell(z) = \begin{cases} 1 & \text{if } z < -\tau, \\ d & \text{if } |z| < \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Bayes rule with reject option

The optimal rule assigns -1 , $+1$ or \textcircled{R} depending on which of $\eta(x)$, $1 - \eta(x)$ or d is smallest. Its risk is $\mathbb{E}[\min\{d, \eta(X), 1 - \eta(X)\}]$ (Chow 1970).

Bayes rule with reject option

The optimal rule assigns -1 , $+1$ or \textcircled{R} depending on which of $\eta(x)$, $1 - \eta(x)$ or d is smallest. Its risk is $\mathbb{E}[\min\{d, \eta(X), 1 - \eta(X)\}]$ (Chow 1970).

According to this rule, we should never invoke the reject option if $d \geq 1/2$ and we should always reject if $d = 0$. For this reason we restrict ourselves to the cases $0 \leq d \leq 1/2$.

Bayes rule with reject option

The Bayes rule with reject option becomes

$$\begin{cases} -1 & \text{if } \eta(x) < d \\ +1 & \text{if } \eta(x) > 1 - d \\ \textcircled{R} & \text{otherwise.} \end{cases}$$

Bayes rule with reject option

The Bayes rule with reject option becomes

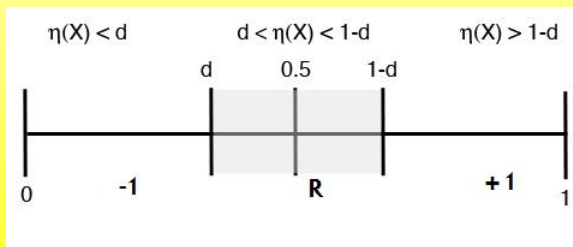
$$\begin{cases} -1 & \text{if } \eta(x) < d \\ +1 & \text{if } \eta(x) > 1 - d \\ \textcircled{R} & \text{otherwise.} \end{cases}$$

This rule corresponds to the discriminant function

$$f_0(x) = \frac{\tau}{1 - 2d} \{2\eta(x) - 1\}.$$

Indeed, $\text{sign}(f_0(x)) = \text{sign}(2\eta(x) - 1)$ and $|f_0(x)| \leq \delta \Leftrightarrow |2\eta(x) - 1| \leq 1 - 2d \Leftrightarrow d \leq \eta(x) \leq 1 - d$.

Bayes rule (with reject option)



Bayes rule with reject option

Remarks

- The case $d = \frac{1}{2}$ reduces to the classical situation *without* the reject option.

Bayes rule with reject option

Remarks

- The case $d = \frac{1}{2}$ reduces to the classical situation *without* the reject option.
- d can be viewed as an upper bound on the conditional probability of misclassification that is considered tolerable (under the hypothetical assumption that we use the Bayes classifier). That is, if $\min\{\eta(x), 1 - \eta(x)\} > d$, then we would prefer the reject option.

Plug-in rules

Plug-in rules

We consider the plug-in classification rule

$$\begin{cases} -1 & \text{if } \hat{\eta}(x) \leq d \\ +1 & \text{if } \hat{\eta}(x) > 1 - d \\ \mathbb{R} & \text{otherwise} \end{cases}$$

based on some estimate $\hat{\eta}$ of $\eta(x)$.

$R_0 = \mathbb{E} [\min\{d, \eta(X), 1 - \eta(X)\}]$ is the smallest risk (Bayes).

$R_0 = \mathbb{E} [\min\{d, \eta(X), 1 - \eta(X)\}]$ is the smallest risk (Bayes).

The difference $\Delta(\hat{f}) = R(\hat{f}) - R_0$ (excess risk) depends on the following two criteria:

$R_0 = \mathbb{E} [\min\{d, \eta(X), 1 - \eta(X)\}]$ is the smallest risk (Bayes).

The difference $\Delta(\hat{f}) = R(\hat{f}) - R_0$ (excess risk) depends on the following two criteria:

- How well does $\hat{\eta}(X)$ estimate $\eta(X)$?

$R_0 = \mathbb{E} [\min\{d, \eta(X), 1 - \eta(X)\}]$ is the smallest risk (Bayes).

The difference $\Delta(\hat{f}) = R(\hat{f}) - R_0$ (excess risk) depends on the following two criteria:

- How well does $\hat{\eta}(X)$ estimate $\eta(X)$?
- What is the behavior of $\eta(X)$ near d and $1 - d$?

Theorem

Theorem

Let $0 \leq d \leq 1/2$ and define for all $0 \leq \tau \leq 1$,

$$P(\tau) = \mathbb{P} \{ |d - \eta(X)| \leq \tau \} + \mathbb{P} \{ |1 - d - \eta(X)| \leq \tau \}.$$

Theorem

Theorem

Let $0 \leq d \leq 1/2$ and define for all $0 \leq \tau \leq 1$,

$$P(\tau) = \mathbb{P} \{ |d - \eta(X)| \leq \tau \} + \mathbb{P} \{ |1 - d - \eta(X)| \leq \tau \}.$$

We have

$$\mathbb{E} \Delta(\hat{f}) \leq \inf_{\tau \geq 0} \{ 2(1 - d) \mathbb{P} \{ |\eta(X) - \hat{\eta}(X)| > \tau \} + \tau P(\tau) \}.$$

Remark

Fast rates (faster than $n^{-1/2}$) can be achieved using plug-in estimates!

Empirical Risk Minimizers

Definition

We consider minimization of the empirical counterpart of the risk

$$R(f) = d\mathbb{P}\{|Yf(X)| \leq \tau\} + \mathbb{P}\{Yf(X) < -\tau\} = \mathbb{E}[\ell(Yf(X))]$$

over a set of discriminant functions $f : \mathcal{X} \rightarrow \mathbb{R}$.

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on the sample size n ,

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on the sample size n , the behavior of $\eta(x)$ around d and $1 - d$,

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on the sample size n , the behavior of $\eta(x)$ around d and $1 - d$, and the metric entropy of the class of considered discriminant functions.

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on the sample size n , the behavior of $\eta(x)$ around d and $1 - d$, and the metric entropy of the class of considered discriminant functions.

Despite its attractive theoretical properties, the naive empirical risk minimization method is often hard to implement.

Herbei and Wegkamp (2006) establish oracle inequalities for the excess risk of the ERM and show that the classification error depends on the sample size n , the behavior of $\eta(x)$ around d and $1 - d$, and the metric entropy of the class of considered discriminant functions.

Despite its attractive theoretical properties, the naive empirical risk minimization method is often hard to implement.

We propose minimization of a convex surrogate for the loss function.

Convex surrogate

The estimators

$$f_\lambda(x) = \sum_{i=1}^M \lambda_i f_i(x), \quad \lambda \in \mathbb{R}^M,$$

of $f_0(x)$ that we study are linear combinations of base functions f_j from a dictionary $F_M = \{f_1, \dots, f_M\}$.

Convex surrogate

The estimators

$$f_\lambda(x) = \sum_{i=1}^M \lambda_i f_i(x), \quad \lambda \in \mathbb{R}^M,$$

of $f_0(x)$ that we study are linear combinations of base functions f_j from a dictionary $F_M = \{f_1, \dots, f_M\}$.

We suggest regularized empirical risk minimization based using convex surrogate loss functions ϕ and a penalty term $\rho(\lambda) = 2r_n|\lambda|_1$ that is proportional to the ℓ_1 -norm $|\lambda|_1$ of the parameter λ .

The regularized empirical risk

$$\frac{1}{n} \sum_{i=1}^n \phi(Y_i f_\lambda(X_i)) + p(\lambda) \quad (1)$$

is then convex in λ and its minimization can be solved by a (tractable) convex program.

General setting

We consider a loss function $\phi : \mathbb{R} \rightarrow [0, \infty)$ that is Lipschitz,

$$|\phi(y) - \phi(y')| \leq C_\phi |y - y'|$$

with $C_\phi < \infty$ and based on this loss function, we define the risk functions

$$R_\phi(\lambda) = \mathbb{E}[\phi(Yf_\lambda(X))] \quad \text{and} \quad \hat{R}_\phi(\lambda) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i f_\lambda(X_i)).$$

Notation and definitions

Assumption: f_0 *also* minimizes the risk $\mathbb{E}[\phi(Yf(X))]$ over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$. The loss function ϕ is classification calibrated.

Notation and definitions

We measure the performance of our estimators in terms of the excess risk

$$\Delta_\phi(\lambda) = R_\phi(\lambda) - R_\phi(f_0).$$

Notation and definitions

We measure the performance of our estimators in terms of the excess risk

$$\Delta_\phi(\lambda) = R_\phi(\lambda) - R_\phi(f_0).$$

Based on the penalty

$$p(\lambda) = 2r_n|\lambda|_1 = 2r_n \sum_{i=1}^M |\lambda_i|$$

with r_n specified later, the penalized empirical risk minimizer $\hat{\lambda}$ satisfies

$$\hat{R}_\phi(\hat{\lambda}) + p(\hat{\lambda}) \leq \hat{R}_\phi(\lambda) + p(\lambda) \quad \text{for all } \lambda \in \mathbb{R}^M. \quad (2)$$

Notation and definitions

In particular, (2) ensures that for $\lambda_0 = (0, \dots, 0)$,

Notation and definitions

In particular, (2) ensures that for $\lambda_0 = (0, \dots, 0)$,

$$p(\hat{\lambda}) \leq \hat{R}_{\phi}(\hat{\lambda}) + p(\hat{\lambda}) \leq \hat{R}_{\phi}(\lambda_0) + p(\lambda_0) = \phi(0)$$

which in turn implies

Notation and definitions

In particular, (2) ensures that for $\lambda_0 = (0, \dots, 0)$,

$$p(\hat{\lambda}) \leq \hat{R}_\phi(\hat{\lambda}) + p(\hat{\lambda}) \leq \hat{R}_\phi(\lambda_0) + p(\lambda_0) = \phi(0)$$

which in turn implies $|\hat{\lambda}|_1 \leq \phi(0)/(2r_n)$.

Notation and definitions

In particular, (2) ensures that for $\lambda_0 = (0, \dots, 0)$,

$$p(\hat{\lambda}) \leq \hat{R}_{\phi}(\hat{\lambda}) + p(\hat{\lambda}) \leq \hat{R}_{\phi}(\lambda_0) + p(\lambda_0) = \phi(0)$$

which in turn implies $|\hat{\lambda}|_1 \leq \phi(0)/(2r_n)$.

This means that we effectively minimize the penalized empirical risk $\hat{R}_{\phi}(\lambda) + p(\lambda)$ over λ in the set

$$\Lambda_n = \left\{ \lambda \in \mathbb{R}^M : |\lambda|_1 \leq \phi(0)/(2r_n) \right\}.$$

We impose two conditions.

- Link between $\|f_\lambda - f_0\|$ and the excess risk $\Delta_\phi(\lambda)$.

We impose two conditions.

- Link between $\|f_\lambda - f_0\|$ and the excess risk $\Delta_\phi(\lambda)$.
This difficulty is absent in regression and density estimation context.

We impose two conditions.

- Link between $\|f_\lambda - f_0\|$ and the excess risk $\Delta_\phi(\lambda)$.
This difficulty is absent in regression and density estimation context.
- Local mutual coherence assumption on the Gram matrix.

Assumption 1

Given some finite measure μ on \mathcal{X} , set

$$\langle f, g \rangle = \int f(x)g(x) \mu(dx) \quad \text{and} \quad \|f\|^2 = \int f^2(x) \mu(dx).$$

The first condition imposes a link between the distance $\|f_\lambda - f_0\|$ and excess risk $\Delta_\phi(\lambda)$:

Assumption 1

Given some finite measure μ on \mathcal{X} , set

$$\langle f, g \rangle = \int f(x)g(x) \mu(dx) \quad \text{and} \quad \|f\|^2 = \int f^2(x) \mu(dx).$$

The first condition imposes a link between the distance $\|f_\lambda - f_0\|$ and excess risk $\Delta_\phi(\lambda)$:

Condition 1. *There exist $C_{\Delta, \mu} < \infty$ and $0 \leq \beta < 1$ such that, for all $\lambda \in \Lambda_n$,*

$$\|f_\lambda - f_0\| \leq C_{\Delta, \mu} \Delta_\phi^\beta(\lambda). \quad (3)$$

- In regression and density estimation problems, this condition trivially holds with $\beta = 1/2$ and $C_{\Delta, \mu} = 1$.

- In regression and density estimation problems, this condition trivially holds with $\beta = 1/2$ and $C_{\Delta, \mu} = 1$.
- This relation is more delicate to establish in classification problems. It depends on the behavior of the conditional probability $\eta(X)$ near d and $1 - d$.

Our goal is to estimate f_0 via linear combinations $f_\lambda(x)$ and to evaluate performance in terms of the excess risk $\Delta_\phi(\lambda)$.

Our goal is to estimate f_0 via linear combinations $f_\lambda(x)$ and to evaluate performance in terms of the excess risk $\Delta_\phi(\lambda)$.

For any $I = \{i_1, \dots, i_m\} \subseteq \{1, \dots, M\}$, we define the approximating parameter space

$$\Lambda(I) = \left\{ \lambda \in \mathbb{R}^M : \lambda_i = 0 \text{ for all } i \notin I \right\}$$

and let $\hat{\lambda}_I$ minimize $\hat{R}_\phi(\lambda)$ over $\Lambda(I)$.

An oracle that knows f_0 would be able to tell us in advance which approximating space $\Lambda(I)$ yields the smallest excess risk $\Delta_\phi(\hat{\lambda}_I)$.

An oracle that knows f_0 would be able to tell us in advance which approximating space $\Lambda(I)$ yields the smallest excess risk $\Delta_\phi(\hat{\lambda}_I)$.

However, f_0 is unknown so the best we can do is to mimic the behavior of the oracle.

An oracle that knows f_0 would be able to tell us in advance which approximating space $\Lambda(I)$ yields the smallest excess risk $\Delta_\phi(\hat{\lambda}_I)$.

However, f_0 is unknown so the best we can do is to mimic the behavior of the oracle. General theory for empirical risk minimization in the classification context (Boucheron, Bousquet and Lugosi; Herbei and Wegkamp) indicates that

$$\Delta_\phi(\hat{\lambda}_I) \lesssim \inf_{\lambda \in \Lambda(I)} \Delta_\phi(\lambda) + \left(\frac{|I|}{n} \right)^\rho.$$

An oracle that knows f_0 would be able to tell us in advance which approximating space $\Lambda(I)$ yields the smallest excess risk $\Delta_\phi(\hat{\lambda}_I)$.

However, f_0 is unknown so the best we can do is to mimic the behavior of the oracle. General theory for empirical risk minimization in the classification context (Boucheron, Bousquet and Lugosi; Herbei and Wegkamp) indicates that

$$\Delta_\phi(\hat{\lambda}_I) \lesssim \inf_{\lambda \in \Lambda(I)} \Delta_\phi(\lambda) + \left(\frac{|I|}{n} \right)^\rho.$$

Various choices are possible for the parameter ρ depending on the margin exponent $\alpha \geq 0$.

Our target of interest, the oracle vector $\lambda^* \in \Lambda_n$, depends on β .
Formally, we define it as follows:

Our target of interest, the oracle vector $\lambda^* \in \Lambda_n$, depends on β .
Formally, we define it as follows:

Definition Let $c_\mu = \min_{1 \leq i \leq M} \|f_i\|$ and let λ^* be the minimizer of

$$3\Delta_\phi(\lambda) + 2 \left(\frac{8C_{\Delta,\mu}}{c_\mu} \right)^{\frac{1}{1-\beta}} (r_n^2 |\lambda|_0)^{\frac{1}{2-2\beta}},$$

over $\lambda \in \Lambda_n$, where $|\lambda|_0 = \sum_{i=1}^M |\lambda_i|$ is the number of non-zero coefficients of the vector λ .

Thus λ^* balances the approximation error, as measured by the excess risk $\Delta_\phi(\lambda)$,

Thus λ^* balances the approximation error, as measured by the excess risk $\Delta_\phi(\lambda)$, and the complexity of the parameter set $\Lambda(I)$ to which λ^* belongs to, as measured by the regularization term $(r_n^2 |\lambda|_0)^{1/(2-2\beta)}$.

Thus λ^* balances the approximation error, as measured by the excess risk $\Delta_\phi(\lambda)$, and the complexity of the parameter set $\Lambda(I)$ to which λ^* belongs to, as measured by the regularization term $(r_n^2 |\lambda|_0)^{1/(2-2\beta)}$.

The constants 3 and $2(8C_{\Delta, \mu})^{1/(1-\beta)}$ can be changed: A decrease in the former will lead to a increase in the latter, and vice-versa. The constant c_μ can be avoided altogether if we take the penalty $\rho(\lambda) = 2r_n \sum_{i=1}^M \|f_i\| |\lambda_i|$, but in practice μ , and consequently $\|f_i\|$, is unknown. Surely we could plug in estimates for $\|f_i\|$.

Condition 2

Let

$$I^* = \{i : \lambda_i^* \neq 0\}$$

be the collection of non-zero coefficients of λ^* ,

Condition 2

Let

$$I^* = \{i : \lambda_i^* \neq 0\}$$

be the collection of non-zero coefficients of λ^* ,

$$|\lambda^*|_0 = \sum_{i=1}^M I_{\{\lambda_i^* \neq 0\}}$$

be the cardinality of I^* ,

Condition 2

Let

$$I^* = \{i : \lambda_i^* \neq 0\}$$

be the collection of non-zero coefficients of λ^* ,

$$|\lambda^*|_0 = \sum_{i=1}^M I_{\{\lambda_i^* \neq 0\}}$$

be the cardinality of I^* , and

$$\rho(i, j) = \frac{\langle f_i, f_j \rangle}{\|f_i\| \cdot \|f_j\|}$$

be the correlation between f_i and f_j .

Condition 2

Our second assumption requires that

$$\rho^* = \max_{i \in I^*} \max_{j \neq i} |\rho(i, j)| \quad (4)$$

is small:

Condition 2: Let $c_\mu = \min_{1 \leq j \leq M} \|f_j\|$ and assume that

$$12\rho^*|\lambda^*|_0 \leq c_\mu. \quad (5)$$

Condition 2

Our second assumption requires that

$$\rho^* = \max_{i \in I^*} \max_{j \neq i} |\rho(i, j)| \quad (4)$$

is small:

Condition 2: Let $c_\mu = \min_{1 \leq j \leq M} \|f_j\|$ and assume that

$$12\rho^*|\lambda^*|_0 \leq c_\mu. \quad (5)$$

This mainly states that the submatrix $(\langle f_i, f_j \rangle)_{i,j \in I^*}$ is positive definite and that the correlations $\rho(i, j)$ between elements f_i , $i \in I^*$, of this submatrix and outside elements f_j , $j \notin I^*$, are relatively small.

Instrumental in our argument is the random quantity

$$\hat{r} = \sup_{\lambda \in \Lambda_n} \frac{|(\hat{R}_\phi - R_\phi)(\lambda) - (\hat{R}_\phi - R_\phi)(\lambda^*)|}{|\lambda - \lambda^*|_1 + \varepsilon_n} \quad (6)$$

where we take $\varepsilon_n = \phi(0)/(nr_n)$.

Our first result states the oracle inequality. It holds true as long as the tuning parameter r_n in the penalty term exceeds \hat{r} .

Theorem

Assume that (3) and (5) hold. On the event $r_n > \hat{r}$,

$$\begin{aligned}
 \Delta_\phi(\hat{\lambda}) + r_n |\hat{\lambda} - \lambda^*|_1 &\leq 3\Delta_\phi(\lambda^*) + 2 \left(\frac{8C_{\Delta,\mu}}{c_\mu} \right)^{\frac{1}{1-\beta}} (r_n^2 |\lambda^*|_0)^{\frac{1}{2-2\beta}} \\
 &\quad + \frac{2\phi(0)}{n}.
 \end{aligned} \tag{7}$$

The ℓ_1 -penalized estimator adapts to both

- the unknown sparsity

The ℓ_1 -penalized estimator adapts to both

- the unknown sparsity
- and the margin condition on $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$

The ℓ_1 -penalized estimator adapts to both

- the unknown sparsity
- and the margin condition on $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$

This is a big deal – AIC/BIC type penalties cannot do this !!

Choice of the tuning parameter r_n

We discuss choices of the tuning parameter r_n that ensure that the probability of the event $\{r_n \geq \hat{r}\}$ is large.

Choice of the tuning parameter r_n

We discuss choices of the tuning parameter r_n that ensure that the probability of the event $\{r_n \geq \hat{r}\}$ is large.

The next lemma states that \hat{r} is sharply concentrated around its mean.

Lemma

Let $C_F = \max_{1 \leq j \leq M} \|f_j\|_\infty$. We have

$$0 \leq \hat{r} \leq 2C_\phi C_F$$

and, for all $\delta > 0$,

$$\mathbb{P} \{ \hat{r} - \mathbb{E}[\hat{r}] \geq \delta \} \leq \exp \left(-\frac{1}{2} \frac{n\delta^2}{C_\phi^2 C_F^2} \right)$$

The range of \hat{r} is important for implementation of the method: We suggest to find a good value for r_n based on cross validation and the grid can be taken on the interval $[0, 2C_\phi C_F]$.

The range of \hat{r} is important for implementation of the method: We suggest to find a good value for r_n based on cross validation and the grid can be taken on the interval $[0, 2C_\phi C_F]$.

The 2nd claim is important for theoretical considerations. It shows that we should take

$$r_n = \mathbb{E}[\hat{r}] + \sqrt{\frac{2 \log(1/\delta)}{n}} C_\phi C_F$$

for some $0 < \delta < 1$, since then

$$\mathbb{P}\{r_n \geq \hat{r}\} \geq 1 - \delta.$$

The expected value $\mathbb{E}[\hat{r}]$ is of order $\{\log(M \vee n)/n\}^{1/2}$ by the following lemma.

Lemma

Let J_n be the smallest integer such that $2^{J_n} \geq n$. Then, for all $M, n \geq 1$ and $0 < \delta < 1$

$$\mathbb{E}[\hat{r}] \leq \frac{7C_\phi C_F}{\sqrt{n}} \sqrt{2 \log 2(M \vee n)} + \frac{J_n C_\phi C_F}{2(M \vee n)^2}.$$

Consequently,

Corollary

Assume that (3) and (5) hold, and take

$$r_n \geq \frac{7C_\phi C_F}{\sqrt{n}} \sqrt{2 \log 2(M \vee n)} + \frac{J_n C_\phi C_F}{2(M \vee n)^2} + C_\phi C_F \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

Then oracle inequality (7) holds with probability at least $1 - \delta$.

SVM with reject option

Hinge loss

Idea:

Truncate the SVM classifier based on the hinge loss

$$\phi(\zeta) = \max(0, 1 - \zeta).$$

Set

$$r(\zeta) = \eta\phi(\zeta) + (1 - \eta)\phi(\zeta)$$

and a little algebra gives

$$r(\zeta) = \begin{cases} \eta - \eta\zeta & \text{if } \zeta \leq -1, \\ 1 + \zeta(1 - 2\eta) & \text{if } -1 < \zeta < 1, \\ (1 - \eta) + \zeta(1 - \eta) & \text{if } \zeta \geq 1 \end{cases}$$

Minimizing $r(\zeta)$ yields

$$\zeta = -\{\eta \leq 1/2\} + \{\eta > 1/2\},$$

so that minimizing

$$\mathbb{E} [(1 - Yf(X))_+] = \mathbb{E}[r(f(X))]$$

yields

$$f(x) = \text{sign}(2\eta(x) - 1).$$

Minimizing $r(\zeta)$ yields

$$\zeta = -\{\eta \leq 1/2\} + \{\eta > 1/2\},$$

so that minimizing

$$\mathbb{E}[(1 - Yf(X))_+] = \mathbb{E}[r(f(X))]$$

yields

$$f(x) = \text{sign}(2\eta(x) - 1).$$

Truncating the sign function does not yield the Bayes (with reject option) classification rule, for any threshold $\tau > 0$!

Generalized hinge loss

We consider the convex surrogate loss

$$\phi_d(z) = \begin{cases} 1 - az & \text{if } z < 0, \\ 1 - z & \text{if } 0 \leq z < 1, \\ 0 & \text{otherwise} \end{cases}$$

where $a = (1 - d)/d \geq 1$.

The function $r(\zeta)$ based on the new ϕ_d can be written as

$$r(\zeta) = \begin{cases} \eta - a\eta\zeta & \text{if } \zeta \leq -1, \\ 1 + \zeta(1 - (1 + a)\eta) & \text{if } -1 < \zeta \leq 0, \\ 1 + \zeta(-\eta + a(1 - \eta)) & \text{if } 0 < \zeta \leq 1, \\ (1 - \eta) + \zeta a(1 - \eta) & \text{if } \zeta > 1 \end{cases}$$

and is minimized by

$$\zeta_0 = \begin{cases} -1 & \text{if } \eta < 1/(1 + a), \\ 0 & \text{if } 1/(1 + a) \leq \eta \leq a/(1 + a), \\ 1 & \text{if } \eta > a/(1 + 1) \end{cases}$$

Classification calibrated

Proposition

The minimizer of the risk

$$R_{\phi_d}(f) = \mathbb{E}[\phi_d(Yf(X))]$$

over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$ is

$$f_0(x) = \begin{cases} -1 & \text{if } \eta(x) < d, \\ 0 & \text{if } d \leq \eta(x) \leq 1 - d, \\ +1 & \text{if } \eta(x) > 1 - d. \end{cases}$$

Classification calibrated

Proposition

The minimizer of the risk

$$R_{\phi_d}(f) = \mathbb{E} [\phi_d(Yf(X))]$$

over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$ is

$$f_0(x) = \begin{cases} -1 & \text{if } \eta(x) < d, \\ 0 & \text{if } d \leq \eta(x) \leq 1 - d, \\ +1 & \text{if } \eta(x) > 1 - d. \end{cases}$$

Furthermore, $dR_{\phi_d}(f_0) = R_0$.

Excess risks

The Bayes discriminant function f_0 minimizes both the risks $\mathbb{E}[\ell(Yf(X))]$ and $\mathbb{E}[\phi_d(Yf(X))]$ over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$.

Excess risks

The Bayes discriminant function f_0 minimizes both the risks $\mathbb{E}[\ell(Yf(X))]$ and $\mathbb{E}[\phi_d(Yf(X))]$ over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$.

We see that $\phi_d(z) \geq \ell(z)$ for all $z \in \mathbb{R}$ as long as $0 \leq \tau \leq 1 - d$.

Excess risks

The Bayes discriminant function f_0 minimizes both the risks $\mathbb{E}[\ell(Yf(X))]$ and $\mathbb{E}[\phi_d(Yf(X))]$ over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$.

We see that $\phi_d(z) \geq \ell(z)$ for all $z \in \mathbb{R}$ as long as $0 \leq \tau \leq 1 - d$.

A relation like this holds not only for the loss functions and hence the risks, but for the excess risks as well.

In particular, for all $d \leq \tau \leq 1 - d$, we have

$$\mathbb{E}[\ell(Yf(X))] - \mathbb{E}[\ell(Yf_0(X))] \leq \mathbb{E}[\phi_d(Yf(X))] - \mathbb{E}[\phi_d(Yf_0(X))].$$

In particular, for all $d \leq \tau \leq 1 - d$, we have

$$\mathbb{E}[\ell(Yf(X))] - \mathbb{E}[\ell(Yf_0(X))] \leq \mathbb{E}[\phi_d(Yf(X))] - \mathbb{E}[\phi_d(Yf_0(X))].$$

This is important since minimization of (1) produces oracle inequalities in terms of the ϕ_d -excess risk, not in terms of the original excess risk directly. The latter risk has a sound statistical interpretation.

Margin condition

Condition: *There exist $A \geq 1$ and $\alpha \geq 0$ such that for all $t > 0$,*

$$\mathbb{P} \{ |\eta(X) - d| \leq t \} \leq At^\alpha \quad \text{and} \quad \mathbb{P} \{ |\eta(X) - (1 - d)| \leq t \} \leq At^\alpha.$$

It can be shown that Condition 1 holds with

- $C_\phi = (1 - d)/d$

It can be shown that Condition 1 holds with

- $C_\phi = (1 - d)/d$
- μ defined by $\mu(B) = \int_B \eta(x)\{1 - \eta(x)\}P_X(dx)$,

It can be shown that Condition 1 holds with

- $C_\phi = (1 - d)/d$
- μ defined by $\mu(B) = \int_B \eta(x)\{1 - \eta(x)\}P_X(dx)$,
- $C_{\Delta,\mu}$ given by

$$C_{\Delta,\mu} = \left\{ 4A(2d)^\alpha \|f_\lambda - f_0\|_\infty^{2+\alpha} \right\}^{\frac{1}{2+2\alpha}},$$

and $\beta = \alpha/(2 + 2\alpha)$

Corollary

Let λ^* minimize

$$3\Delta_{\phi}(\lambda) + C\|f^* - f_0\|_{\infty}(r_n^2|\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $C = C(c_{\mu}, d, A, \alpha)$.

Corollary

Let λ^* minimize

$$3\Delta_\phi(\lambda) + C\|f^* - f_0\|_\infty(r_n^2|\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $C = C(c_\mu, d, A, \alpha)$. Assume that

- $\rho^*|\lambda^*|_0 \leq \frac{c_\mu}{16-c_\mu}$

Corollary

Let λ^* minimize

$$3\Delta_\phi(\lambda) + C\|f^* - f_0\|_\infty (r_n^2 |\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $C = C(c_\mu, d, A, \alpha)$. Assume that

- $\rho^* |\lambda^*|_0 \leq \frac{c_\mu}{16 - c_\mu}$
- $r_n^{\frac{\alpha}{1+\alpha}} |\lambda^*|_0 \leq c,$

Corollary

Let λ^* minimize

$$3\Delta_\phi(\lambda) + C\|f^* - f_0\|_\infty(r_n^2|\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $C = C(c_\mu, d, A, \alpha)$. Assume that

- $\rho^*|\lambda^*|_0 \leq \frac{c_\mu}{16-c_\mu}$
- $r_n^{\frac{\alpha}{1+\alpha}}|\lambda^*|_0 \leq c,$

then, for all choices $r_n = r_n(\delta)$ in (8),

Corollary

Let λ^* minimize

$$3\Delta_{\phi}(\lambda) + C\|f^* - f_0\|_{\infty}(r_n^2|\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $C = C(c_{\mu}, d, A, \alpha)$. Assume that

- $\rho^*|\lambda^*|_0 \leq \frac{c_{\mu}}{16-c_{\mu}}$
- $r_n^{\frac{\alpha}{1+\alpha}}|\lambda^*|_0 \leq c,$

then, for all choices $r_n = r_n(\delta)$ in (8), we have

$$\begin{aligned} & \Delta_{\phi_d}(\hat{\lambda}) + r_n|\hat{\lambda} - \lambda^*|_1 \\ & \leq 3\Delta_{\phi_d}(\lambda^*) + C\|f^* - f_0\|_{\infty}(r_n^2|\lambda^*|_0)^{\frac{1+\alpha}{2+\alpha}} + \frac{2\phi(0)}{n}. \end{aligned}$$

with probability at least $1 - \delta$.

Bibliography
Introduction
Plug-in rules
Empirical Risk Minimizers
SVM classifier with reject option

Hinge loss
Generalized hinge loss
Classification calibrated
Excess risk comparison
Margin condition
Oracle inequality

Thanks!