CUHK STAT6050: Statistical Learning Theory Spring 2022

Lecture 8: Binary classification

Fisher consistency, surrogate loss and excess risk bounds

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 Binary classification

We denote a vector of features as $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$, and a binary outcome (label) as $Y \in \{-1, 1\}$. Our goal is to find a binary decision function $\psi : \mathscr{X} \to \{-1, 1\}$ to predict a label given a new instance, and its performance is evaluated as Misclassification error (MCR):

$$\mathbb{P}\big(Y \neq \psi(\mathbf{X})\big) = \mathbb{E}\mathbf{1}\big(Y\psi(\mathbf{X}) \leq 0\big).$$

Yet, ψ is a binary outcome function, to facilitate computation, we can make decision by taking sign of a continuous function $f : \mathscr{X} \to \mathbb{R}$, that is, $\psi(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x}))$. Then, the MCR loss and its risk function can be rewritten as:

$$l(Yf(\mathbf{X})) = \mathbf{1}\big(Yf(\mathbf{X}) \leq 0\big), \quad R(f) = \mathbb{E}\big(l(Yf(\mathbf{X}))\big).$$

Remark 1.1. In binary classification, the loss function can be degenerated to a univariate function based on $Yf(\mathbf{X})$.

Bayes classifier

With the same idea, we first consider the Bayes classifier based on misclassification error.

Before proceed, we assume $\mathbb{P}\big(Y = 1|\mathbf{X} = \mathbf{x}\big) \neq 1/2$ for all $\mathbf{x} \in \mathscr{X}$ to simplify the notations. Note that when $\mathbb{P}\big(Y = 1|\mathbf{X} = \mathbf{x}\big) = 1/2$, the prediction of Y can be arbitrary.

Lemma 1.2 (Bayes classifier)**.** *f^* is a global minimizer of $R(f)$ if and only if*

$$\operatorname{sgn}\big(f^*(\mathbf{x})\big) = \operatorname{sgn}\big(\eta(\mathbf{x}) - 1/2\big),$$

where $\eta(\mathbf{x}) = \mathbb{P}\big(Y = 1|\mathbf{X} = \mathbf{x}\big)$, and $\operatorname{sgn}(u) = 1$, if $x \geq 0$, -1, otherwise.

Remark 1.3 (Identifiability). In classification, we usually do not consider the "identifiability" issue, it partly because that the true decision function is also not unique. On the other hand, we are only interested in the performance (MCR).

Remark 1.4 (Plug-in prediction). A simple method motivated by Bayes classifier is using plug-in estimator: $\widehat{\psi}(\mathbf{x}) = \operatorname{sgn}(\widehat{\eta}(\mathbf{x}) - 1/2)$, where $\widehat{\eta}(\cdot)$ is an estimator of $\eta(\cdot)$. **PS**: The relationship/difference between classification and regression.

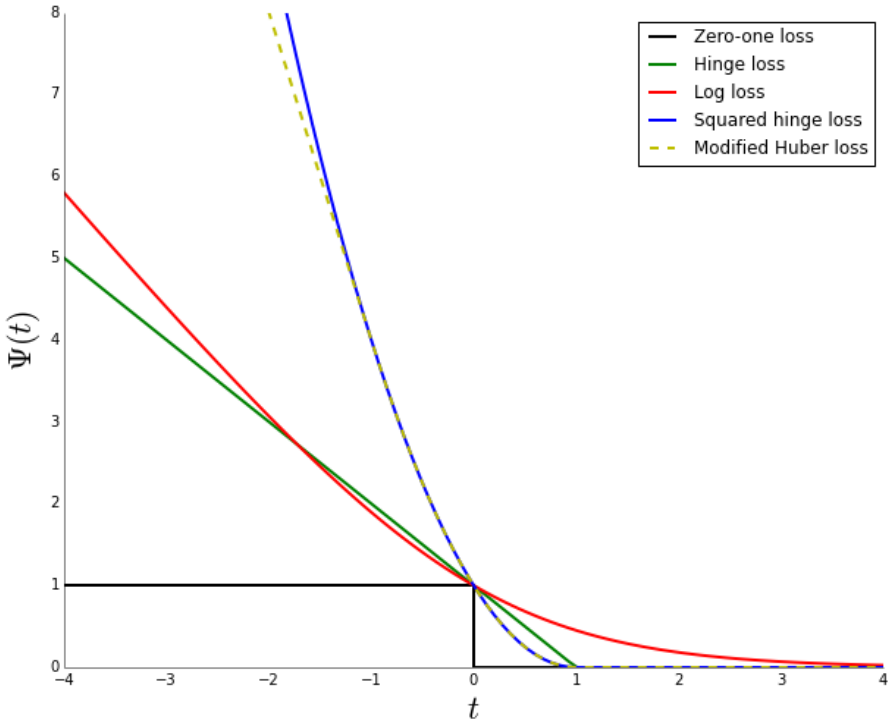

Figure 1: Plot for various surrogate losses from [Pedregosa, 2014].

Empirical risk minimization

Next, given training samples $(\mathbf{x}_i, y_i)_{i=1,\cdots,n}$, we give the ERM or R-ERM on binary classification based on MCR.

$$\min_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} l(Y_i f(\mathbf{x}_i)) + \lambda_n \|f\|_{\mathscr{F}}^2,$$

where $\mathscr{F}$ is a candidate function class, which can be RKHS, boosting, tree methods, or neural networks. However, note that the indicator in the proposed ERM is discontinuous, which is infeasible/difficult to handle in optimization. To facilitate the computation, we turn to replace the indicator function by a surrogate loss ϕ.

$$\widehat{f}_n = \arg\min_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(\mathbf{x}_i)) + \lambda_n \|f\|_{\mathscr{F}}^2, \tag{1}$$

Remark 1.5. The surrogate loss framework summarizes (also was inspired from) variety of classification methods: including logistic regression [Cox, 1972], AdaBoost [Freund and Schapire, 1997], and support vector machines (SVM; [Cortes and Vapnik, 1995]).

- Exponential loss: (i.e., AdaBoost)

$$\phi\big(Yf(\mathbf{X})\big) = \exp(-Yf(\mathbf{X})).$$

- Logistic loss: (i.e., logistic regression)

$$\phi\big(Yf(\mathbf{X})\big) = \log\big(1+\exp(-Yf(\mathbf{X}))\big).$$

- Square loss:

$$\phi\big(Yf(\mathbf{X})\big) = (1-Yf(\mathbf{X}))^2.$$

- Hinge loss: (i.e., SVM)

$$\phi\big(Yf(\mathbf{X})\big) = (1-Yf(\mathbf{X}))_+.$$

With the same manner, the risk function and excess risk based on a surrogate loss ϕ are defined as:

$$R_\phi(f) = \mathbb{E}\big(\phi(Yf(\mathbf{X}))\big), \quad \mathscr{E}_\phi(f) = R_\phi(f) - R_\phi(f_\phi^*),$$

where f_ϕ^* is a minimizer of $R_\phi(f)$.

A "good" surrogate loss

Note that $\widehat{f}_n$ from (1) is no longer a minimizer of ERM based on the evaluation loss $l(\cdot)$, but a replaced surrogate loss $\phi(\cdot)$. Therefore, the question we are interested here is if we still have the "nice" asymptotics of $\mathscr{E}(\widehat{f}_n)$. We turn to address the question by several steps.

- **Fisher consistency.** First, we consider the weakest possible condition on ϕ in population level. For every minimizer f_ϕ^* of $R_\phi(f)$, $f_\phi^*(\mathbf{x})$ should have the same sign as Bayes decision rule $\operatorname{sgn}(\eta(\mathbf{x})-1/2)$.

- **Relation between $\mathscr{E}(f)$ and $\mathscr{E}_\phi(f)$.** Since the $\widehat{f}_n$ is obtained by minimizing $\widehat{R}_{\phi,n}$, thus it is expected that $\mathscr{E}_\phi = o_P(1)$. It will be quite useful if we can obtain the relation between $\mathscr{E}(f)$ and $\mathscr{E}_\phi(f)$.

- **Convergence rates and probabilistic bounds.** The final aim is to give the convergence rate and probabilistic bound of $\mathscr{E}(\widehat{f}_n)$.

2 Fisher consistency

In this section, we give the formal definition of Fisher consistency (sometimes is referred as *classification calibration*).

Definition 2.1 (Fisher consistency [Lin, 2004])**.** A surrogate loss $\phi(\cdot)$ is Fisher consistency to $l(\cdot)$ if for any $f_\phi^* \in \arg\min_f R_\phi(f)$,

$$f_\phi^*(\mathbf{X}) > 0, \text{ if } \eta(\mathbf{X}) > 1/2, \quad f_\phi^*(\mathbf{X}) < 0, \text{ if } \eta(\mathbf{X}) < 1/2, \quad \text{almost surely.}$$

Lemma 2.2. *The following surrogate losses are all Fisher consistency, and their corresponding minimizers are given as:*

- ***Logistic loss.***
$$f_\phi^*(\mathbf{x}) = \sigma^{-1}(\eta(\mathbf{x})),$$
where $\sigma(u) = 1/(1+\exp(-u))$ *is a sigmoid function.*

- ***Exponential loss.***
$$f_\phi^*(\mathbf{x}) = \big(\log(\eta(\mathbf{x})) - \log(1-\eta(\mathbf{x}))\big)/2.$$

- ***Square loss.***
$$f_\phi^*(\mathbf{x}) = 2\eta(\mathbf{x}) - 1.$$

- ***Hinge loss.***
$$f_\phi^*(\mathbf{x}) = \operatorname{sgn}(\eta(\mathbf{x}) - 1/2).$$

From Lemma 2.2, we summarize the **point-wise minimization** procedure to verify Fisher consistency of ϕ.

- **Point-wise minimization.**

$$\mathbb{E}\phi\big(Y, f(\mathbf{X})\big) = \mathbb{E}_{\mathbf{X}}\mathbb{E}\Big(\phi\big(Yf(\mathbf{X})\big)\Big|\mathbf{X}\Big) = \mathbb{E}_{\mathbf{X}}\Big(\eta(\mathbf{X})\phi\big(f(\mathbf{X})\big) + (1-\eta(\mathbf{X}))\phi\big(-f(\mathbf{X})\big)\Big),$$

it suffices to consider the point-wise minimization:

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

where $H(\eta)$ is the optimal conditional ϕ-risk. Then, for any $\mathbf{x} \in \mathscr{X}$,

$$f_\phi^*(\mathbf{x}) = \arg\min_\alpha C_\eta(\alpha).$$

Theorem 2.3 (**Convex** Fisher consistent surrogate loss [Bartlett et al., 2006])**.** *Let* $\phi(\cdot)$ *be convex. Then* ϕ *is Fisher consistency if and only if it is differentiable at 0 and* $\phi'(0) < 0$.

3 Relation between $\mathscr{E}$ and $\mathscr{E}_\phi$

Next, we investigate the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$. First, we present some useful fact for $\mathscr{E}$.

Lemma 3.1. *For $R(f)$ and $\mathscr{E}(f)$ defined in Section 1, then*

$$R^* = R(f^*) = \mathbb{E}\Big(\min(\eta(\mathbf{X}), 1-\eta(\mathbf{X}))\Big) = \big|\eta(\mathbf{X}) - 1/2\big| + 1/2,$$
$$\mathscr{E}(f) = R(f) - R(f^*) = \mathbb{E}\Big(\mathbf{1}\big\{\operatorname{sgn}\big(f(\mathbf{X})\big) \neq \operatorname{sgn}\big(f^*(\mathbf{X})\big)\big\}\Big|2\eta(\mathbf{X}) - 1\Big|\Big).$$

To investigate the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$, we give the definition of the optimal disagreement risk $H^-(\eta)$ [Zhang, 2004]:

$$H^-(\eta) = \inf_{\alpha:\alpha(2\eta-1)\leq 0} C_\eta(\alpha).$$

Remark 3.2. Some important properties of $H(\eta)$ and $H^-(1-\eta)$.

- Fisher consistency is equivalent to $H^-(\eta) > H(\eta)$.
- $H(\eta) = H(1-\eta)$ and $H^-(\eta) = H^-(1-\eta)$.

Theorem 3.3 ([Zhang, 2004, Bartlett et al., 2006])**.** *For any nonnegative loss function ϕ, any measurable function $f : \mathscr{X} \to \mathbb{R}$ and any probability distribution on $\mathscr{X} \times \{-1, +1\}$,*

$$\psi\big(R(f) - R^*\big) \leq R_\phi(f) - R_\phi^*,$$

where $\psi : [0,1] \to [0,\infty)$ is defined as the Fenchel-Legendre biconjugate of $\widetilde{\psi}$, denoted as $\psi = \widetilde{\psi}$ and

$$\widetilde{\psi}(\theta) = H^-(\frac{1+\theta}{2}) - H(\frac{1+\theta}{2}).$$

Remark 3.4. If ψ is invertible on $[0,1]$, then

$$\mathscr{E}(f) = R(f) - R^* \leq \psi^{-1}\big(R_\phi(f) - R_\phi^*\big) = \psi^{-1}\big(\mathscr{E}_\phi(f)\big),$$

which tells the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$. More importantly, it yields that good performance in $\phi(\cdot)$ yields a good performance on $l(\cdot)$.

Theorem 3.5 (Excess risk bound for convex surrogate loss [Bartlett et al., 2006])**.** *If ϕ is convex and Fisher consistency, then*

$$\psi(u) = \phi(0) - H(\frac{1+u}{2}).$$

4 Examples

[Bartlett et al., 2006] illustrates some examples of surrogate losses in classification.

- Exponential loss: $\phi(u) = \exp(-u)$.
  - Fisher consistency (Theorem 2.3).
  - Excess risk bound (Theorem 3.5).
    - $H(\eta) = 2\sqrt{\eta(1-\eta)}$.
    - $\psi(u) = 1 - \sqrt{1-u^2}$, for $u \in [0,1]$.
    - $\psi^{-1}(u) = \sqrt{u(2-u)}$, for $u \in [0,1]$.
    - Excess risk bound:
$$\mathscr{E}(f) \leq \sqrt{2\mathscr{E}_\phi(f)}.$$
- Hinge loss: $\phi(u) = (1-u)_+$.
  - Fisher consistency (Theorem 2.3).
  - Excess risk bound (Theorem 3.5).
    - $H(\eta) = 2\min\{\eta, 1-\eta\}$.
    - $\psi(u) = |u| = u$, for $u \in [0,1]$.
    - $\psi^{-1}(u) = u$, for $u \in [0,1]$.
    - Excess risk bound:
$$\mathscr{E}(f) \leq \mathscr{E}_\phi(f).$$

Remark 4.1. We can show that: (i) for the squared loss:

$$\mathscr{E}(f) \leq \sqrt{\mathscr{E}_\phi(f)};$$

(ii) for the logistic loss:

$$\mathscr{E}(f) \leq \sqrt{2\mathscr{E}_\phi(f)}.$$

5 A loose excess risk upper bound for kernel SVM

According to the results in Sections 1 - 4, suppose we find an increasing function $\psi : [0,1] \to \mathbb{R}^+$, such that for any $f \in \mathscr{F}$:

$$\psi\big(\mathscr{E}(f)\big) \leq \mathscr{E}_\phi(f),$$

then, for $0 \leq \varepsilon_n \leq 1$, we have

$$\mathbb{P}\big(\mathscr{E}(\widehat{f}_n) \geq \varepsilon_n\big) = \mathbb{P}\Big(\psi\big(\mathscr{E}(\widehat{f}_n)\big) \geq \psi(\varepsilon_n)\Big) \leq \mathbb{P}\Big(\mathscr{E}_\phi(\widehat{f}_n) \geq \psi(\varepsilon_n)\Big). \tag{2}$$

Let $t_n = \psi(\varepsilon_n)$, it suffices to investigate the asymptotics of $\mathscr{E}_\phi(\widehat{f}_n)$.

R-ERM for kernel SVMs

For illustration, let's consider the kernel-based SVM, that is, SVM in RKHS:

$$\widehat{f_n} = \arg\min_{f \in \mathscr{H}_K} \frac{1}{n} \sum_{i=1}^{n} \big(1 - Y_i f(\mathbf{X}_i)\big)_+ + \lambda_n \|f\|^2_{\mathscr{H}_K}$$

Remark 5.1. For kernel SVM, we still have Representer theorem.

The reader may check Sections 4.5, 12.1-12.3 in [Hastie et al., 2001] to see the motivation and detailed computation of SVMs.

Remark 5.2. Truncation of $\widehat{f_n}(\mathbf{x})$ to $[-1,1]$ always yields a lower (or equal) loss.

On this ground, we assume $\|\widehat{f_n}\|_\infty$ is bounded by 1, otherwise we could consider the asymptotics of truncated estimator. Recall the decomposition:

$$\mathscr{E}_\phi(\widehat{f_n}) = R_\phi(\widehat{f_n}) - R_\phi(f^*) \le 2 \sup_{f \in \mathscr{F}} |\widehat{R}_{\phi,n}(f) - R_\phi(f)| + \mathbf{Approx}_\phi(\lambda_n),$$

where

$$\mathbf{Approx}_\phi(\lambda_n) = \inf_{f \in \mathscr{H}_n} R_\phi(f) - R_\phi(f^*) + \lambda_n \|f\|^2_{\mathscr{H}_n}.$$

Approximation error

In regression case, we know that the convergence rate of the approximation error is related to "smoothness" of the Bayes decision function. With the same manner, we have the following definition of "smoothness" in classification, which is so-called "geometric-noise" assumption.

Theorem 5.3 (Theorem 2.7 in [Steinwart and Scovel, 2007])**.** *Let $\mathscr{X} \subset \mathbb{R}^d$ be a compact domain, K be a Gaussian kernel with a hyperparameter σ_n, and the distribution of $(\mathbf{X},\mathbf{Y})$ satisfies the geometric-noise assumption (Definition 2.3 in [Steinwart and Scovel, 2007]) with geometric noise exponent $0 < \alpha < \infty$. Then, there exists a constant A_0, such that*

$$\mathbf{Approx}_\phi(\lambda_n) \le A_0 \lambda_n^{\frac{\alpha}{\alpha+1}},$$

provided by $\sigma_n = \lambda_n^{-\frac{1}{d(\alpha+1)}}$.

Estimation error

For any f and f' in $\mathscr{F} = \mathscr{H}_K$, we have

$$\big|\phi(Yf(\mathbf{X})) - \phi(Yf'(\mathbf{X}))\big| \le \big|f(\mathbf{X}) - f'(\mathbf{X})\big|,$$

Talagrand's contraction Lemma (Lemma 3.2 in Lecture 4) yields that,

$$\mathbb{E}\|\mathbf{Rad}_n(\phi \bullet f)\|_{\mathscr{H}_K} \le \mathbb{E}\|\mathbf{Rad}_n(f)\|_{\mathscr{H}_K} \le \lambda_n^{-1/2} K_0 \sqrt{\frac{1}{n}}.$$

Hyperparameter tuning

Then, by Corollary 3.1 in Lecture 6, if

$$\varepsilon_n \geq A_0 \lambda_n^{\frac{\alpha}{\alpha+1}} + \lambda_n^{-1/2} K_0 \sqrt{\frac{1}{n}} \geq \mathbf{Approx}_\phi(\lambda_n) + 8\mathbb{E}\|\mathbf{Rad}_n(\phi \bullet f)\|_{\mathscr{H}_K}.$$

Then,

$$\mathbb{P}\big(\mathscr{E}_\phi(\widehat{f}_n) \geq \varepsilon_n\big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{8(1+5\varepsilon_n/6)}\Big).$$

We can tune λ_n to improve the convergence rate:

$$\varepsilon_n^* = \inf_{\lambda_n} A_0 \lambda_n^{\frac{\alpha}{1+\alpha}} + cK_0(n\lambda_n)^{-1/2} = O(n^{-\alpha/(1+3\alpha)}),$$

obtained by $\lambda_n = O(n^{-(\alpha+1)/(1+3\alpha)})$. Therefore, the convergence rate is given as:

$$\mathscr{E}(\widehat{f}_n) = O_P(\varepsilon_n^*) = O_P(n^{-\frac{\alpha}{1+3\alpha}}).$$

Remark 5.4 (How to improve?). (i) The trivial function class $\widehat{f}_n \in \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \leq \lambda_n^{-1/2}\}$ can be significantly improved by "low-noise" and "geometric noise" assumptions. (ii) Local/Random Rademacher complexity.

References

[Bartlett et al., 2006] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Cox, 1972] Cox, D. R. (1972). The analysis of multivariate binary data. *Applied statistics*, pages 113–120.

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Lin, 2004] Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82.

[Pedregosa, 2014] Pedregosa, F. (2014). Surrogate loss functions in machine learning.

[Steinwart and Scovel, 2007] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607.

[Zhang, 2004] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.