Regularization of Kernel Methods by Decreasing the Bandwidth of the Gaussian Kernel

Jean-Philippe Vert (joint work with Régis Vert)
Jean-Philippe.Vert@ensmp.fr

Centre for Computational Biology
Ecole des Mines de Paris, ParisTech

Machine Learning Summer School, Taipei, June 28, 2006

Outline

Outline

1. Motivations
2. Main results
3. Proofs
   - Learning bound for the R_0 risk
   - From R_0 to Bayes excess risk
   - From R_0 excess risk to L_2 convergence
4. Conclusion

Gaussian kernel and RKHS

Definition

- The (normalized) Gaussian kernel with bandwidth $\sigma > 0$ on $\mathbb{R}^d \times \mathbb{R}^d$ is:

$$k_\sigma(x, x') = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right) .$$

- The Gaussian reproducing kernel Hilbert space (RKHS) consists of functions of the form:

$$f(x) = \sum_i \alpha_i k_\sigma(x_i, x) ,$$

with norm:

$$\| f \|^2_{\mathcal{H}_\sigma} = \sum_i \sum_j \alpha_i \alpha_j k_\sigma(x_i, x_j) .$$

Gaussian kernel and RKHS

Definition

- The (normalized) Gaussian kernel with bandwidth $\sigma > 0$ on $\mathbb{R}^d \times \mathbb{R}^d$ is:

$$k_\sigma(x, x') = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right) .$$

- The Gaussian reproducing kernel Hilbert space (RKHS) consists of functions of the form:

$$f(x) = \sum_i \alpha_i k_\sigma(x_i, x) ,$$

with norm:

$$\| f \|^2_{\mathcal{H}_\sigma} = \sum_i \sum_j \alpha_i \alpha_j k_\sigma(x_i, x_j) .$$

Gaussian RKHS

Properties

- For any f in $L_1(\mathbb{R}^d)$, its Fourier transform $\mathcal{F}[f] : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} e^{-i<x,\omega>} f(x) dx \,.$$

- The RKHS of the Gaussian kernel k_σ is:

$$\mathcal{H}_\sigma = \left\{ f \in \mathcal{C}_0(\mathbb{R}^d) \; : \; f \in L_1(\mathbb{R}^d) \text{ and } \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega < \infty \right\}$$

- For any $f \in \mathcal{H}_\sigma$ the RKHS norm of f is a smoothness functional:

$$\| f \|^2_{\mathcal{H}_\sigma} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega \,.$$

Learning in Gaussian RKHS

General setting

- Training set $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$.
- Loss function $L(y, \hat{y})$
- Learn a function $f : \mathbb{R}^d \to \mathbb{R}$ by solving for some regularization parameter $\lambda > 0$:

$$\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) + \lambda \| f \|_{\mathcal{H}_\sigma}^2 \right\} .$$

Pattern recognition

- $y \in \{-1, +1\}$
- $L(y, u) = \phi(yu)$ where ϕ is usually decreasing

Motivation 1: The effect of regularization

Overfitting

$$\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) + \frac{\lambda}{(2\pi)^d} \int_{\mathbb{R}^d} \left|\mathcal{F}\left[f\right](\omega)\right|^2 e^{\frac{\sigma^2 \parallel \omega \parallel^2}{2}} d\omega \right\} .$$

- Classical approach: Decrease λ
- Alternative approach: Decrease σ

Asymptotic behavior when $n \to \infty$

- Usually $\lambda \to 0$ (Tikhonov and Arsenin, 1977; Silverman, 1982) to obtain consistency
- $\lambda \to 0$ and $\sigma \to 0$ can lead to fast rates (e.g., Steinwart and Scovel, 2004)
- Can we get consistency with $\sigma \to 0$ only?

Motivation 1: The effect of regularization

Overfitting

$$\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \frac{\lambda}{(2\pi)^d} \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega \right\} .$$

- Classical approach: Decrease λ
- Alternative approach: Decrease σ

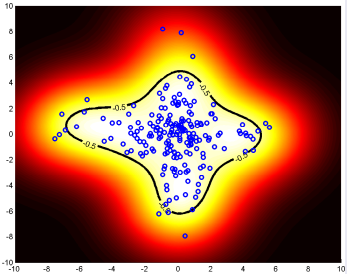
Asymptotic behavior when $n \to \infty$

- Usually $\lambda \to 0$ (Tikhonov and Arsenin, 1977; Silverman, 1982) to obtain consistency
- $\lambda \to 0$ and $\sigma \to 0$ can lead to fast rates (e.g., Steinwart and Scovel, 2004)
- Can we get consistency with $\sigma \to 0$ only?

Motivation 2: One-class SVM

Definition

$$\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\left(1 - f(x_i), 0\right) + \lambda \| f \|_{\mathcal{H}_\sigma}^2 \right\} .$$



Properties

- A popular method for outlier detection
- A particular case of learning in the Gaussian RKHS
- λ determines the ratio of outliers: should not decrease to zero as $n \to \infty$
- Can we get some consistency when $\sigma \to 0$ instead?

Outline

1. Motivations
2. Main results
3. Proofs
   - Learning bound for the R_0 risk
   - From R_0 to Bayes excess risk
   - From R_0 excess risk to L_2 convergence
4. Conclusion

Setting and notations

- $(X_i, Y_i)_{i=1,\dots,n}$ are i.i.d. $\sim P$ over $\mathbb{R}^d \times \{-1, 1\}$.
- Marginal $P(dx) = \rho(x)dx$.
- $\eta(x) : \mathbb{R}^d \to [0, 1]$ a measurable version of $P(Y = 1 \mid X)$.
- ϕ a convex function, Lipschitz, differentiable at 0 with $\phi'(0) < 0$.
- For any σ, we denote by $\hat{f}_\sigma$ the unique minimizer of the (strictly convex) problem:

$$\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi\left(Y_i f(X_i)\right) + \lambda \| f \|^2_{\mathcal{H}_\sigma} \right\} .$$

Intuitive behavior

Pointwise limit

- Law of large numbers for measurable f:

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(Y_i f(X_i)\right) \underset{n\to\infty}{\to} \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] .$$

- For $f \in \mathcal{H}_{\sigma_1}$:

$$\| f \|^2_{\mathcal{H}_\sigma} \underset{\sigma\to 0}{\to} \| f \|^2_{L_2}$$

Limit risk

This suggests to consider the following risk for measurable functions:

$$R_0\left(f\right) = \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] + \lambda \| f \|^2_{L_2} .$$

Intuitive behavior

Pointwise limit

- Law of large numbers for measurable f:

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(Y_i f(X_i)\right) \underset{n\to\infty}{\to} \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] .$$

- For $f \in \mathcal{H}_{\sigma_1}$:

$$\| f \|_{\mathcal{H}_\sigma}^2 \underset{\sigma\to 0}{\to} \| f \|_{L_2}^2$$

Limit risk

This suggests to consider the following risk for measurable functions:

$$R_0\left(f\right) = \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] + \lambda\| f \|_{L_2}^2 .$$

Main result: consistency

Theorem

- *If $\sigma = O\left(n^{-\frac{1}{d+\epsilon}}\right)$ for $\epsilon > 0$, then the procedure is consistent for the R_0 risk:*

$$R_0\left(\hat{f}_\sigma\right) \underset{n\to\infty}{\to} \inf_{f\in\mathcal{M}} R_0(f) \quad \textit{in probability.}$$

- *In that case, it is also Bayes consistent:*

$$R\left(\hat{f}_\sigma\right) \underset{n\to\infty}{\to} \inf_{f\in\mathcal{M}} R(f) \quad \textit{in probability,}$$

where R is the classification error $R(f) = P\left(Yf(X) < 0\right)$.

Main result: consistency

Theorem

- *If* $\sigma = O\left(n^{-\frac{1}{d+\epsilon}}\right)$ *for* $\epsilon > 0$, *then the procedure is consistent for the R_0 risk:*

$$R_0\left(\hat{f}_\sigma\right) \underset{n\to\infty}{\rightarrow} \inf_{f\in\mathcal{M}} R_0(f) \quad \textit{in probability.}$$

- *In that case, it is also Bayes consistent:*

$$R\left(\hat{f}_\sigma\right) \underset{n\to\infty}{\rightarrow} \inf_{f\in\mathcal{M}} R(f) \quad \textit{in probability,}$$

where R is the classification error $R(f) = P\left(Yf(X) < 0\right)$.

Main result: asymptotic shape

Theorem

- *The function* $f_0 : \mathbb{R}^d \to \mathbb{R}$ *defined for any* $x \in \mathbb{R}^d$ *by*

$$f_0(x) = \arg\min_{\alpha \in \mathbb{R}} \left\{ \rho(x) \left[\eta(x)\phi(\alpha) + (1 - \eta(x))\, \phi(-\alpha) \right] + \lambda\alpha^2 \right\}$$

is measurable and satisfies

$$R_0(f_0) = \inf_{f \in \mathcal{M}} R_0(f) \, .$$

- *Under the conditions of the previous theorem:*

$$\| \hat{f}_\sigma - f_0 \|_{L_2} \underset{n \to \infty}{\longrightarrow} 0 \quad \textit{in probability.}$$

Main result: asymptotic shape

Theorem

- *The function* $f_0 : \mathbb{R}^d \to \mathbb{R}$ *defined for any* $x \in \mathbb{R}^d$ *by*

$$f_0(x) = \arg\min_{\alpha \in \mathbb{R}} \left\{ \rho(x) \left[\eta(x)\phi(\alpha) + (1 - \eta(x))\, \phi(-\alpha) \right] + \lambda \alpha^2 \right\}$$

is measurable and satisfies

$$R_0(f_0) = \inf_{f \in \mathcal{M}} R_0(f)\ .$$

- *Under the conditions of the previous theorem:*

$$\| \hat{f}_\sigma - f_0 \|_{L_2} \underset{n \to \infty}{\longrightarrow} 0 \quad \textit{in probability.}$$

Application: two-class SVM

1-SVM

The L_2 limit of the SVM with hinge loss $\phi(u) = \max(1-u,0)$ is:

$$f_0(x) = \begin{cases} -1 & \text{if } \eta(x) \leq 1/2 - \lambda/\rho(x)\,, \\ (\eta(x)-1/2)\,\rho(x)/\lambda & \text{if } \eta(x) \in [1/2-\lambda/\rho(x), 1/2+\lambda/\rho(x)]\,, \\ 1 & \text{if } \eta(x) \geq 1/2 + \lambda/\rho(x)\,. \end{cases}$$

2-SVM

The L_2 limit of the SVM with square hinge loss $\phi(u) = \max(1-u,0)^2$ is:

$$f_0(x) = (2\eta(x)-1)\,\frac{\rho(x)}{\lambda+\rho(x)}$$

Application: two-class SVM

1-SVM

The L_2 limit of the SVM with hinge loss $\phi(u) = \max(1 - u, 0)$ is:

$$f_0(x) = \begin{cases} -1 & \text{if } \eta(x) \leq 1/2 - \lambda/\rho(x)\ , \\ (\eta(x) - 1/2)\,\rho(x)/\lambda & \text{if } \eta(x) \in [1/2 - \lambda/\rho(x), 1/2 + \lambda/\rho(x)]\ , \\ 1 & \text{if } \eta(x) \geq 1/2 + \lambda/\rho(x)\ . \end{cases}$$

2-SVM

The L_2 limit of the SVM with square hinge loss $\phi(u) = \max(1 - u, 0)^2$ is:

$$f_0(x) = (2\eta(x) - 1)\,\frac{\rho(x)}{\lambda + \rho(x)}$$

Application: one-class SVM

Limit

The L_2 limit of the one-class SVM with hinge loss is the density truncated to level 2λ and scaled:

$$f_0(x) = \begin{cases} \rho(x)/2\lambda & \text{if } \rho(x) \leq 2\lambda \,, \\ 1 & \text{otherwise.} \end{cases}$$

Corollary

One-class SVM thresholded at level $\mu/2\lambda$ is a consistent estimator (w.r.t. the excess-mass risk, cf Hartigan, 1987) of the density level set:

$$C_\mu = \left\{ x \in \mathbb{R}^d \; : \; \rho(x) \geq \mu \right\}$$

Outline

Overview

1. Learning bound for the R_0 risk: with a probability at least $1 - \epsilon$,

$$R_0\left(\hat{f}_\sigma\right) - \inf_{g \in \mathcal{M}} R_0(g) \leq C(\epsilon) \, .$$

2. From R_0 to Bayes excess risk: for any measurable function f,

$$R(f) - \inf_{g \in \mathcal{M}} R(g) \leq \psi\left(R_0\left(\hat{f}_\sigma\right) - \inf_{g \in \mathcal{M}} R_0(g)\right) \, .$$

3. From R_0 excess risk to L_2 convergence: for any measurable function f,

$$\| f - \hat{f}_\sigma \|_{L_2}^2 \leq \frac{1}{\lambda}\left[R_0\left(\hat{f}_\sigma\right) - \inf_{g \in \mathcal{M}} R_0(g)\right] \, .$$

Outline

Objectif

- Risks:

$$R_0(f) = \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] + \lambda \| f \|_{L_2}^2 ,$$

$$R_\sigma(f) = \mathbb{E}_P\left[\phi\left(Yf(X)\right)\right] + \lambda \| f \|_{\mathcal{H}_\sigma}^2 ,$$

$$\widehat{R}_\sigma(f) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(Y_i f(X_i)\right) + \lambda \| f \|_{\mathcal{H}_\sigma}^2 .$$

- Minimizers

$$R_0^* = R_0(f_0) = \min_{f \in \mathcal{M}} R_0(f)$$

$$R_\sigma^* = R_\sigma(f_\sigma) = \min_{f \in \mathcal{H}_\sigma} R_\sigma(f)$$

$$\widehat{R}_\sigma^* = \widehat{R}_\sigma\left(\hat{f}_\sigma\right) = \min_{f \in \mathcal{H}_\sigma} \widehat{R}_\sigma(f)$$

Decomposition of the excess R_0 risk

Decomposition

$$\begin{aligned} R_0\left(\hat{f}_\sigma\right) - R_0\left(f_0\right) = & \left[R_0\left(\hat{f}_\sigma\right) - R_\sigma\left(\hat{f}_\sigma\right)\right] \\ & + \left[R_\sigma\left(\hat{f}_\sigma\right) - R_\sigma^*\right] \\ & + \left[R_\sigma^* - R_\sigma\left(g\right)\right] \\ & + \left[R_\sigma\left(g\right) - R_0\left(g\right)\right] \\ & + \left[R_0\left(g\right) - R_0\left(f_0\right)\right] \end{aligned}$$

for any g in $\mathcal{H}_\sigma$.

Simplification

- $R_0\left(f\right) - R_\sigma\left(f\right) = \parallel f \parallel_{L_2}^2 - \parallel f \parallel_{\mathcal{H}_\sigma}^2 \leq 0$ for any $f \in \mathcal{H}_\sigma$.
- $R_\sigma^* - R_\sigma\left(g\right) \leq 0$ by definition of R_σ^*.

Upper bound on the R_0 risk

After simplification

$$
\begin{aligned}
R_0\left(\hat{f}_\sigma\right) - R_0\left(f_0\right) \leq & \left[R_\sigma\left(\hat{f}_\sigma\right) - R_\sigma^*\right] && \text{(estimation error)} \\
& + \| g \|_{\mathcal{H}_\sigma}^2 - \| g \|_{L_2}^2 && \text{(regularization error)} \\
& + \left[R_0\left(g\right) - R_0\left(f_0\right)\right] && \text{(approximation error)}
\end{aligned}
$$

for any g in $\mathcal{H}_\sigma$.

Choice of g

- g should be smooth (regularization error)
- g should be close to f_0 (approximation error)
- We choose $g = k_{\sigma_1} * f_0$, with $\sigma_1 \geq \sigma$

Upper bound on the R_0 risk

After simplification

$$
\begin{aligned}
R_0\left(\hat{f}_\sigma\right) - R_0\left(f_0\right) \leq & \left[R_\sigma\left(\hat{f}_\sigma\right) - R_\sigma^*\right] && \text{(estimation error)} \\
& + \| g \|^2_{\mathcal{H}_\sigma} - \| g \|^2_{L_2} && \text{(regularization error)} \\
& + \left[R_0\left(g\right) - R_0\left(f_0\right)\right] && \text{(approximation error)}
\end{aligned}
$$

for any g in $\mathcal{H}_\sigma$.

Choice of g

- g should be smooth (regularization error)
- g should be close to f_0 (approximation error)
- We choose $g = k_{\sigma_1} * f_0$, with $\sigma_1 \geq \sigma$

Estimation error bound

Concentration inequality

- Classical bounds of statistical learning theory
- Need an upper bound of the covering number of balls in the Gaussian RKHS (e.g., Steinwart and Scovel, 2004)
- Need a concentration inequality based on local Rademacher complexity (e.g., Bartlett et al., 2005)
- For any $x \geq 1, 0 < p < 2$ and $\delta > 0$, we have with probability at least $1 - e^x$:

$$R_\sigma\left(\hat{f}_\sigma\right) - R_\sigma^* \leq C_1 \left(\frac{1}{\sigma}\right)^{\frac{d[2+(2-p)(1+\delta)]}{2+p}} \left(\frac{1}{n}\right)^{\frac{2}{2+p}} + C_2 \left(\frac{1}{\sigma}\right)^d \frac{x}{n} .$$

Regularization error bound

Fourier representation of Gaussian RKHS

$$\| f \|^2_{\mathcal{H}_\sigma} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \mathcal{F}\left[f\right](\omega) \right|^2 e^{\frac{\sigma^2 \| \omega \|^2}{2}} d\omega \,.$$

Therefore, for any $0 < \sigma \leq \tau$, $\mathcal{H}_\tau \subset \mathcal{H}_\sigma \subset L_2(\mathbb{R}^d)$.

Lemma

- *For any* $\sigma > 0$ *and* $f \in L_1\left(\mathbb{R}^d\right) \cap L_2\left(\mathbb{R}^d\right)$*:*

$$k_\sigma * f \in \mathcal{H}_{\sqrt{2}\sigma} \quad \text{and} \quad \| k_\sigma * f \|_{\mathcal{H}_{\sqrt{2}\sigma}} = \| f \|_{L_2} \,.$$

- *For any* $0 < \sigma \leq \sqrt{2}\tau$ *and* $f \in L_1\left(\mathbb{R}^d\right) \cap L_2\left(\mathbb{R}^d\right)$*:*

$$k_\tau * f \in \mathcal{H}_\sigma \quad \text{and} \quad \| k_\tau * f \|^2_{\mathcal{H}_\sigma} - \| k_\tau * f \|^2_{L_2} \leq \frac{\sigma^2}{\tau^2} \| f \|^2_{L_2} \,.$$

Approximation error bound

Lemma

$$R_0\left(k_\sigma * f_0\right) - R_0\left(f_0\right) \leq \left(2\lambda \| f_0 \|_{L_\infty} + LM\right) \| k_\sigma * f_0 - f_0 \|_{L_1} \ ,$$

where L is the Lipschitz constant of ϕ and $M = \sup_{x \in \mathbb{R}^d} \rho(x)$. This shows that the approximation error converges to 0.

Quantitative bound

The modulus of continuity of f in the L_1-norm is:

$$\omega\left(f, \delta\right) = \sup_{0 \leq \| t \| \leq \delta} \| f\left(. + t\right) - f\left(.\right) \|_{L_1} \ .$$

For any $\sigma > 0$ the following holds:

$$\| k_\sigma * f_0 - f_0 \|_{L_1} \leq \left(1 + \sqrt{d}\right) \omega\left(f, \sigma\right) \ .$$

Approximation error bound

Lemma

$$R_0(k_\sigma * f_0) - R_0(f_0) \leq (2\lambda \| f_0 \|_{L_\infty} + LM) \| k_\sigma * f_0 - f_0 \|_{L_1} ,$$

where L is the Lipschitz constant of ϕ and $M = \sup_{x \in \mathbb{R}^d} \rho(x)$. This shows that the approximation error converges to 0*.*

Quantitative bound

The modulus of continuity of f in the L_1-norm is:

$$\omega(f, \delta) = \sup_{0 \leq \| t \| \leq \delta} \| f(. + t) - f(.) \|_{L_1} .$$

For any $\sigma > 0$ the following holds:

$$\| k_\sigma * f_0 - f_0 \|_{L_1} \leq \left(1 + \sqrt{d}\right) \omega(f, \sigma) .$$

Summary

Proof of R_0 consistency

Combining the 3 upper bounds on the estimation, regularization and approximation errors we obtain:

$$R_0\left(\hat{f}_\sigma\right) - R_0\left(f_0\right) \leq C_1 \left(\frac{1}{\sigma}\right)^{\frac{d[2+(2-p)(1+\delta)]}{2+p}} \left(\frac{1}{n}\right)^{\frac{2}{2+p}} + C_2 \left(\frac{1}{\sigma}\right)^{d} \frac{x}{n} + C_3 \frac{\sigma_1^2}{\sigma^2} + C_4 \omega\left(f_0, \sigma_1\right) .$$

Convergence to 0 is granted as soon as $\sigma = O\left(n^{-\frac{1}{d+\epsilon}}\right)$ and $\sigma_1 = o(\sigma)$. Terms can be balanced to obtain a bound that depends on the modulus of continuity of f_0.

Outline

Classification calibration

Definition (Bartlett et al., 2006)

For any $(\eta, \alpha) \in [0, 1] \times \mathbb{R}$, let

$$C_\eta(\alpha) = \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) \, .$$

The function ϕ is said to be classification-calibrated if for any $\eta \neq 1/2$,

$$\inf_{\alpha\in\mathbb{R}:\alpha(2\eta-1)\leq 0} C_\eta(\alpha) > \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha).$$

This condition ensures that for each point x, minimizing the conditional ϕ-risk provides a scalar of correct sign. We can then deduce the Bayes consistency of algorithms that minimize the ϕ risk instead of the classification error (Zhang, 2004; Lugosi and Vayatis, 2004; Bartlett et al., 2006).

R-classification calibration

Definition

We can rewrite the R_0-risk as:

$$R_0(f) = \int_{\mathbb{R}^d} \left\{ \left[\eta(x)\phi(f(x)) + (1 - \eta(x))\,\phi(-f(x)) \right] \rho(x) + \lambda f(x)^2 \right\} dx$$

Therefore, for any $(\eta, \rho, \alpha) \in [0, 1] \times (0, +\infty) \times \mathbb{R}$ let

$$C_{\eta,\rho}(\alpha) = C_\eta(\alpha) + \frac{\lambda \alpha^2}{\rho}\,.$$

We say that ϕ is R-classification calibrated if for any $\eta \neq 1/2$ and $\rho > 0$:

$$\inf_{\alpha \in \mathbb{R} : \alpha(2\eta - 1) \leq 0} C_{\eta,\rho}(\alpha) > \inf_{\alpha \in \mathbb{R}} C_{\eta,\rho}(\alpha).$$

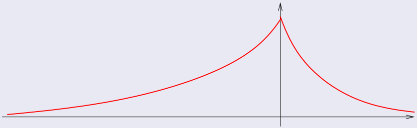
Some properties of calibration

Lemma

- $\phi(x)$ is R-calibrated iff $\phi(x) + tx^2$ is calibrated for all $t > 0$.
- Calibration (resp. R-calibration) does not imply R-calibration (resp. calibration).
- If ϕ is convex the it is calibrated iff it is R-calibrated iff it is differentiable at 0 and $\phi'(0) < 0$.

Calibrated but not R-calibrated

R-calibrated but not calibrated



Relating the R_0 risk to the classification error rate

Sketch

- When $\lambda = 0$ Bartlett et al. (2006) provide a control of the excess ϕ-risk by the excess classification error for classification calibrated functions.
- Following the same approach we obtain similar controls for the R_0 risk if ϕ is R-classification calibrated.

Outline

Minimum R_0 risk

Lemma

- For any $(\eta, \rho, \alpha) \in [0,1] \times [0, +\infty) \times \mathbb{R}$ let

$$G_{\eta,\rho}(\alpha) = \rho\left[\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right] + \lambda\alpha^2.$$

  i.e., for any $f \in \mathcal{M}$

$$R_0(f) = \int_{x\in\mathbb{R}^d} G_{\eta(x),\rho(x)}\left(f(x)\right) dx\ .$$

- If ϕ is convex then $G_{\eta,\rho}$ is strictly convex and admits a unique minimizer $\alpha(\eta, \rho)$.
- $f_0(x) = \alpha\left(\eta(x), \rho(x)\right)$ is measurable and minimizes R_0.

From R_0 risk to L_2 distance

Lemma

By strict convexity of $G_{\eta,\rho}$ we obtain, for all (η, ρ, α):

$$G_{\eta,\rho}(\alpha) - G_{\eta,\rho}\left(\alpha(\eta,\rho)\right) \geq \lambda \left(\alpha - \alpha(\eta,\rho)\right)^2 \; .$$

Conclusion

By integration we obtain:

$$R_0(f) - R_0\left(f_0\right) \geq \lambda \| f - f_0 \|_{L_2} \; .$$

Outline

Conclusion

Conclusion

- Consistency for the R_0 risk is obtained by decreasing the bandwidth of the Gaussian kernel
- The limit function in the L_2 sense is the minimizer of the R_0 risk, given explicitly and uniquely defined for convex ϕ.
- R_0-consistency ensures Bayes consistency for pattern recognition.
- One-class SVM is a consistent density level set estimator
- The convergence speed obtained are not optimal

Reference

R. Vert and J-P. Vert, Consistency and convergence rates of one-class SVMs and related algorithms, J. Mach. Learn. Res. 7:817-854, 2006.