Approximation Theoretical Questions for SVMs

Ingo Steinwart
LA-UR 07-7056

October 20, 2007

Informal Description of the Learning Goal

- X space of input samples
  Y space of labels, usually $Y \subset \mathbb{R}$.
- Already observed samples

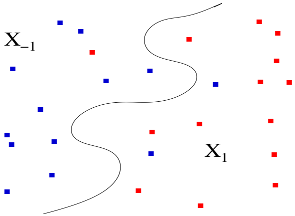
$$T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$$

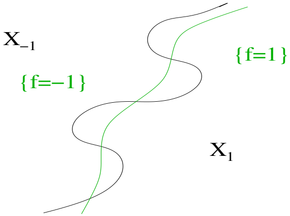
- **Goal:**
  With the help of T find a function $f : X \to \mathbb{R}$ which predicts label y for new, unseen x.

Illustration: Binary Classification

- **Problem:**
  The set X is devided into two *unknown* classes X_{-1} and X_1.
- **Goal:**
  Find approximately the classes X_{-1} and X_1.

Illlustration:





Left: Negative (blue) and positive (red) samples.
Right: Behaviour of a decision function (green) $f : X \to Y$.

Formal Definition of Statistical Learning

- **Basic Assumptions:**
  - P is an *unknown* probability measure on $X \times Y$.
  - $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ sampled from P^n.
  - Future (x, y) will also be sampled from P.
  - $L : Y \times \mathbb{R} \to [0, \infty]$ *loss function* that measures cost $L(y, t)$ of predicting y by t.
- **Goal:**
  Find a function $f_T : X \to \mathbb{R}$ with small *risk*

$$\mathcal{R}_{L,P}(f_T) := \int_{X \times Y} L\big(y, f_T(x)\big)\, dP(x, y) \ .$$

- **Interpretation:**
  Average future cost of predicting by f_T should be small.

Questions in Statistical Learning I

- **Bayes risk:**

$$\mathcal{R}^*_{L,P} := \inf\{\, \mathcal{R}_{L,P}(f) \mid f : X \to \mathbb{R} \text{ measurable} \,\} .$$

A function attaining this minimum is denoted by $f^*_{L,P}$.

- **Learning method:**
Assigns to every training set T a predictor $f_T : X \to \mathbb{R}$.

- **Consistency:**
A learning method is called *universally consistent* if

$$\mathcal{R}_{L,P}(f_T) \to \mathcal{R}^*_{L,P} \qquad \text{in probability} \qquad (1)$$

for $n \to \infty$ and *every* probability measure P on $X \times Y$.

- **Good news:**
Many learning methods are universally consistent.
First result: Stone (1977), AoS

Questions in Statistical Learning II

- **Rates:**
  Does there exist a learning method and a convergence rate $a_n \searrow 0$ such that

$$\mathbb{E}_{T\sim P^n}\mathcal{R}_{L,P}(f_T) - \mathcal{R}^*_{L,P} \;\leq\; C_P\, a_n\,, \qquad\qquad n \geq 1,$$

  for *every* probability measure P on $X \times Y$.

- **Bad news:** (Devroye, 1982, IEEE TPAMI)
  No! (if $|Y| \geq 2$, $|X| = \infty$, and L "non-trivial")
- **Good news:**
  Yes, if one makes some "mild?!" assumptions on P.
  Too many results in this direction to mention them.

Reproducing Kernel Hilbert Spaces I

- $k : X \times X \to \mathbb{R}$ is a **kernel**
  :⇔ there exist a Hilbert space H and a map $\Phi : X \to H$ with
  $$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \qquad \text{for all } x, x' \in X.$$
  ⇔ all $(k(x_i, x_j))_{i,j=1}^n$ are symmetric and positive semi-definite.
- **RKHS** of k: the "smallest" such H that consists of functions.
  - *"Construction":* Take the "completion" of
    $$\Big\{ \sum_{i=1}^n \alpha_i k(x_i, .) : n \in \mathbb{N}, \alpha_1, \dots, \alpha_n \in \mathbb{R}, x_1, \dots, x_n \in X \Big\}$$
    equipped with the dot product
    $$\Big\langle \sum_{i=1}^n \alpha_i k(x_i, .), \sum_{j=1}^m \beta_j k(\hat{x}_j, .) \Big\rangle := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, \hat{x}_j) .$$
  - **Feature map:** $\Phi : x \mapsto k(x, .)$.

Reproducing Kernel Hilbert Spaces II

- **Polynomial Kernels:**
  For $a \geq 0$ and $m \in \mathbb{N}$ let
  $$k(x,x') := (\langle x, x' \rangle + a)^m, \qquad x, x' \in \mathbb{R}^d .$$
- **Gaussian RBF kernels:**
  For $\sigma > 0$ let
  $$k_\sigma(x,x') := \exp(-\sigma^2 \|x - x'\|_2^2), \qquad x, x' \in \mathbb{R}^d .$$
  The parameter $1/\sigma$ is called *width*.
- **Denseness of Gaussian RKHSs:**
  The RKHS H_σ of k_σ is dense in $L_p(\mu)$ for all $p \in [1, \infty)$ and all probability measures μ on $\mathbb{R}^d$.

Support Vector Machines I

- **Support vector machines (SVMs)** solve the problem

$$f_{T,\lambda} = \arg\min_{f \in H} \; \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \; , \tag{2}$$

  - H is a RKHS,
  - $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$ is a training set,
  - $\lambda > 0$ is a *free* regularization parameter,
  - $L : Y \times \mathbb{R} \to [0, \infty)$ is a *convex* loss, e.g.
    - *hinge loss:* $L(y, t) := \max\{0, 1 - yt\}$
    - *least squares loss:* $L(y, t) := (y - t)^2$.
- **Representer Theorem:**
  The *unique* solution is of the form $f_{T,\lambda} = \sum_{i=1}^n \alpha_i k(x_i, .)$.
  Minimization actually takes place over $\{\alpha_1, \ldots, \alpha_n\}$.

Support Vector Machines II

Questions:

- Universally consistent?
- Learning rates?
- Efficient algorithms?
- Performance on real world problems?
- Additional properties?

An Oracle Inequality: Assumptions

Assumptions and notations:

- $L(y, 0) \leq 1$ for all $y \in Y$.
- $L(y, .) : \mathbb{R} \to [0, \infty)$ convex and has a minimum in $[-1, 1]$.
- $\breve{t} := \max\{-1, \min\{1, t\}\}$.
- L is locally Lipschitz:

$$\left|L(y,t) - L(y,t')\right| \leq |t - t'|, \qquad y \in Y, t, t' \in [-1, 1].$$

  This yields

$$L(y, \breve{t}) \leq \left|L(y, \breve{t}) - L(y, 0)\right| + L(y, 0) \leq 2$$

- Variance bound:
  $\exists \vartheta \in [0, 1]$ and $V \geq 2$ $\forall f : X \to \mathbb{R}$:

$$\mathbb{E}_P\big(L \circ \breve{f} - L \circ f^*_{L,P}\big)^2 \leq V \cdot \big(\mathbb{E}_P(L \circ \breve{f} - L \circ f^*_{L,P})\big)^\vartheta$$

Entropy Numbers

Let $S : E \to F$ be a bounded linear operator and $n \geq 1$. The *n-th* (dyadic) entropy number of S is defined by

$$e_n(S) := \inf\Big\{\varepsilon > 0 : \exists x_1, \ldots, x_{2^{n-1}} : SB_E \subset \bigcup_{i=1}^{2^{n-1}} (x_i + \varepsilon B_F)\Big\}.$$

An Oracle Inequality

Oracle Inequality (slightly simplified)

- H separable RKHS of measurable kernel with $\|k\|_\infty \leq 1$.
- Entropy assumption: $\exists p \in (0,1)$ and $a \geq 1$:

$$\mathbb{E}_{T_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(T_X)) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad i,\, n \geq 1.$$

- Fix an $f_0 \in H$ and a $B_0 \geq 1$ such that $\|L \circ f_0\|_\infty \leq B_0$,

Then there exists a constant $K > 0$ such that with probability P^n not less than $1 - e^{-\tau}$ we have

$$\begin{aligned}\mathcal{R}_{L,P}(\check{f}_{T,\lambda}) - \mathcal{R}^*_{L,P} \;&\leq\; 9\big(\lambda\|f_0\|_H^2 + \mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}\big) \\ &\quad + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\Big(\frac{72V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{30B_0\tau}{n}\end{aligned}$$

A Simplification

Consider the approximation error function

$$A(\lambda) := \min_{f \in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \Big)$$

and the (unique) minimizer $f_{P,\lambda}$.

$\Longrightarrow$ For $f_0 := f_{P,\lambda}$ we can choose $B_0 = 1 + 2\sqrt{\frac{A(\lambda)}{\lambda}}$

Refined Oracle inequality

$$\mathcal{R}_{L,P}(\check{f}_{T,\lambda}) - \mathcal{R}^*_{L,P} \quad \leq \quad 9A(\lambda) + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{60\tau}{n}\sqrt{\frac{A(\lambda)}{\lambda}} + 3\Big(\frac{72V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{30\tau}{n}$$

Consistency

Assumptions:

- $L(y,t) \leq c(1+t^q)$ for all $y \in Y$ and $t \in R$.
- H is dense in $L_q(P_X)$

$\Longrightarrow$ $A(\lambda) \to 0$ for $\lambda \to 0$.

$\Longrightarrow$ SVM is consistent whenever we chose λ_n such that

$$\begin{aligned} \lambda_n &\to 0 \\ \sup n\lambda_n &< \infty . \end{aligned}$$

Learning Rates

Assumptions:

- There exists constants $c \geq 1$ and $\beta \in (0, 1]$ such that

$$A(\lambda) \leq c\lambda^\beta, \qquad \lambda \geq 0.$$

  Note: $\beta = 1 \implies f^*_{L,P} \in H$.
- L is Lipschitz continuous (e.g. hinge loss).

$\implies$ Choosing $\lambda_n \sim n^{-\alpha}$ we obtain a polynomial learning rate.
Zhou et al. (2005?)
Some calculations show that the best learning rate we can obtain is

$$n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p-\vartheta+\vartheta p)+p}\}}.$$

It is achieved by

$$\lambda_n \sim n^{-\min\{\frac{2}{\beta+1}, \frac{1}{\beta(2-p-\vartheta+\vartheta p)+p}\}}.$$

Adaptivity

For $T = ((x_1, y_1), \ldots, (x_n, y_n))$ define $m := \lfloor n/2 \rfloor + 1$ and

$$\begin{aligned} T_1 &:= ((x_1, y_1), \ldots, (x_m, y_m)) \\ T_2 &:= ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)) \,. \end{aligned}$$

- Split T into T_1 and T_2.
- Fix an n^{-2} net Λ_n of $(0, 1]$.
- **Training:** Use T_1 to find $f_{T_1,\lambda}$, $\lambda \in \Lambda_n$.
- **Validation:** Use T_2 to determine a $\lambda_{T_2} \in \Lambda_n$ that satisifes

$$\mathcal{R}_{L,T_2}(\breve{f}_{T_1,\lambda_{T_2}}) = \min_{\lambda \in \Lambda_n} \mathcal{R}_{L,T_2}(\breve{f}_{T_1,\lambda}) \,.$$

$\Longrightarrow$ This yields a consistent learning method with learning rate

$$n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p-\vartheta+\vartheta p)+p}\}} \,.$$

Discussion

- The oracle inequality can be generalized to regularized risk minimizers.
- The presented oracle inequality yields fastest known rates in many cases.
- In some cases these rates are known to be optimal in a min-max sense.
- Oracle inequalities can be used to design adaptive strategies that learn fast without knowing key parameters of P.

Question: Which distributions can be learned fast?

Discussion II

Observations:

- Data often lies in high dimensional spaces, but not uniformly.
- Regression: target is often smooth (but not always).
- Classification: How much do classes "overlap"?

Observations

The relation between RKHS H and distribution P is descibed by two quantities:

- The constants a and p in

$$\mathbb{E}_{T_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(T_X)) \;\leq\; a\, i^{-\frac{1}{2p}}, \qquad i,\, n \geq 1.$$

- The approximation error function

$$A(\lambda) := \min_{f \in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^* \Big)$$

Task:
Find realistic assumptions on P such that both quantities are small for commonly used kernels.

Entropy Numbers

Consider the integral operator $T_k : L_2(P_X) \to L_2(P_X)$ defined by

$$T_k f(x) := \int_X k(x, x') f(x')\, P_X(dx')$$

Question:
What is the relation between the EW's of T_k and

$$\mathbb{E}_{T_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(T_X)) \ ?$$

Question:
What is the behaviour if $X \subset \mathbb{R}^d$ but P_X is not absolutely continuous with respect to the Lebesgue measure?

Approximation Error Function

- For the least squares loss we have

$$A(\lambda) = \inf_{f \in H} \lambda \|f\|_H^2 + \|f - f^*_{L,P}\|^2_{L_2(P_X)}.$$

- For Lipschitz continuous losses we have have

$$A(\lambda) \leq \inf_{f \in H} \lambda \|f\|_H^2 + \|f - f^*_{L,P}\|_{L_1(P_X)}.$$

Smale & Zhou 03:
In both cases the behaviour of $A(\lambda)$ for $\lambda \to 0$ can be characterized by the K-functional of the pair $(H, L_p(P_X))$.

Questions:
What happens if $X \subset \mathbb{R}^d$ but P_X is not absolutely continuous with respect to the Lebesgue measure?

Introduction

Observations:

- The Gaussian kernel is successfully used in many applications.
- It has a parameter σ that is almost never fixed a-priori.

Question:

How does σ influence the learning rates?

Appproximation Quantities for Gaussian Kernels

- For H_σ being the Gaussian kernel with width σ the entropy assumption is of the form

$$\mathbb{E}_{T_X \sim P_X^n} e_i(\mathrm{id} : H_\sigma \to L_2(T_X)) \leq a_\sigma \, i^{-\frac{1}{2p}}, \qquad i,\, n \geq 1.$$

- The approximation error function also depends on σ:

$$A_\sigma(\lambda) = \inf_{f \in H_\sigma} \lambda \|f\|_{H_\sigma}^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*.$$

Oracle Inequality

Oracle inequality using Gaussian kernels

$$\begin{aligned}\mathcal{R}_{L,P}(\check{f}_{T,\lambda}) - \mathcal{R}^*_{L,P} \quad &\leq \quad 9\big(\lambda\|f_0\|^2_{H_\sigma} + \mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}\big) \\ &\quad + K\left(\frac{a_\sigma^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{30B_0(\sigma)\tau}{n}\end{aligned}$$

Usually σ becomes larger with the sample size. **Task:**
Find estimates that are good in σ and i (or λ), simultaneously.

An Estimate for the Entropy Numbers

Theorem: (S. & Scovel, AoS 2007)
Let $X \subset \mathbb{R}^d$ be compact. Then for all $\varepsilon > 0$ and $0 < p < 1$ there exists a constant $c_{\varepsilon,p} \geq 1$ such that

$$\mathbb{E}_{T_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(T_X)) \leq c_{\varepsilon,p}\, \sigma^{\frac{(1-p)(1+\varepsilon)d}{2p}}\, i^{-\frac{1}{2p}}$$

This estimate does not consider properties of P_X.

Questions:
How good is this estimate?
For which P_X can this be significantly improved?

Distance to the Decision Boundary

- $\eta(x) := P(y = 1|x)$.
- $X_{-1} := \{\eta < 1/2\}$ and $X_1 := \{\eta > 1/2\}$.
- For $x \in X \subset \mathbb{R}^d$ we define

$$\Delta(x) := \begin{cases} d(x, X_1), & \text{if } x \in X_{-1}, \\ d(x, X_{-1}), & \text{if } x \in X_1, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where $d(x, A)$ denotes the distance between x and A.

Interpretation:
$\Delta(x)$ measures the distance of x to the "decision boundary".

Margin Exponents

- **Margin exponent:**
  $\exists c \geq 1$ and $\alpha > 0$ such that

$$P_X(\Delta(x) \leq t) \leq c t^\alpha\,, \qquad\qquad t > 0.$$

  Example:
  - $X \subset \mathbb{R}^d$ compact, positive volume.
  - P_X uniform distribution.
  - Decision boundary linear or circle.

  $\Longrightarrow \alpha = 1.$
  This remains true under transformations
- **Margin-noise exponent:**
  $\exists c \geq 1$ and $\beta > 0$ such that

$$|2\eta - 1|\, P_X(\Delta(x) \leq t) \leq c t^\alpha\,, \qquad\qquad t > 0.$$

Interpretation:

A Bound on the Approximation Error

Theorem: (S. & Scovel, AoS 2007)

- $X \subset \mathbb{R}^d$ compact.
- P distribution on $X \times \{-1, 1\}$ with margin-noise exponent β.
- L hinge loss.

$\exists\ c_{d,\tau} > 0$ and $\tilde{c}_{d,\beta} > 0\ \forall\ \sigma > 0$ and $\lambda > 0\ \exists\ f^* \in H_\sigma$ satisfying $\|f^*\|_\infty \leq 1$ and

$$\lambda \|f^*\|^2_{H_\sigma} + \mathcal{R}_{L,P}(f^*) - \mathcal{R}^*_{L,P} \leq c_{d,\tau} \lambda \sigma^d + \tilde{c}_{d,\beta}\, c\, \sigma^{-\beta}\,.$$

Remarks:

- Not optimal in λ.
- How can this be improved?
- Better dependence on dimension?!

Conclusion

- Oracle inequalities can be used to design adaptive SVMs.
- For which distributions do such adaptive SVMs learn fast?
  - Bounds for entropy numbers
  - Bounds for approximation error