Advanced Statistical Learning Theory

Olivier Bousquet
Pertinence
32, rue des Jeûneurs
F-75002 Paris, France
olivier.bousquet@pertinence.com

Machine Learning Summer School, September 2004

Roadmap

- Lecture 1: Union bounds and PAC Bayesian techniques
- Lecture 2: Variance and Local Rademacher Averages
- Lecture 3: Loss Functions
- Lecture 4: Applications to SVM

Lecture 1

Union Bounds and PAC-Bayesian Techniques

- Binary classification problem
- Union bound with a prior
- Randomized Classification
- Refined union bounds

Probabilistic Model

We consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$.
Here: classification case $\mathcal{Y} = \{-1, 1\}$.
Assumption: The pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to P (unknown).
Data: We observe a sequence of n i.i.d. pairs (X_i, Y_i) sampled according to P.
Goal: construct a function $g : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X, i.e. with low risk

$$R(g) = P(g(X) \neq Y) = \mathbb{E}\left[1_{[g(X) \neq Y]}\right]$$

Probabilistic Model

Issues

- P is unknown so that we cannot directly measure the risk
- Can only measure the agreement on the data
- Empirical Risk

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} 1_{[g(X_i) \neq Y_i]}$$

Bounds (1)

A learning algorithm

- Takes as input the data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Produces a function g_n

Can we estimate the risk of g_n ?

$\Rightarrow$ random quantity (depends on the data).

$\Rightarrow$ need probabilistic bounds

Bounds (2)

- Error bounds

$$R(g_n) \leq R_n(g_n) + B$$

  $\Rightarrow$ Estimation from an empirical quantity

- Relative error bounds
  - $\star$ Best in a class

$$R(g_n) \leq R(g^*) + B$$

  - $\star$ Bayes risk

$$R(g_n) \leq R^* + B$$

  $\Rightarrow$ Theoretical guarantees

Notation

Important: to simplify writing we use the notation:

- $Z = (X, Y)$
- $\mathcal{G}$: hypothesis class, g function from $\mathcal{X}$ to $\mathbb{R}$
- $\mathcal{F}$: loss class or centered loss class, f function from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$

$$f(z) = f((x, y)) = \ell(g(x), y) \quad \text{or} \quad \ell(g(x), y) - \ell(g^*(x), y)$$

Simplest case $\ell(g(x), y) = 1_{[g(x) \neq y]}$

- $R(g) = Pf := \mathbb{E}[f(X, Y)]$, $R_n(g) = P_n f := \frac{1}{n} \sum_{i=1}^n f(Z_i)$

Take Home Messages

- Two ingredients of bounds: deviations and union bound
- Optimal union bound with metric structure of the function space
- Can introduce a prior into the union bound
- PAC-Bayesian technique: improves the bound when averaged

Deviations

Hoeffding's inequality

for each fixed $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$Pf - P_n f \leq C\sqrt{\frac{\log \frac{1}{\delta}}{n}}\,. \tag{1}$$

Finite union bound

For a finite set of functions $\mathcal{F}$ with probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq C\sqrt{\frac{\log|\mathcal{F}| + \log\frac{1}{\delta}}{n}}\,. \tag{2}$$

- $\log|\mathcal{F}|$ is analogue to a variance
- extra variability from the unknown choice
- measures the size of the class

Weighted union bound

Introduce a probability distribution π over $\mathcal{F}$: with probability at least $1-\delta$,

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq C\sqrt{\frac{\log 1/\pi(f) + \log \frac{1}{\delta}}{n}}\,. \tag{3}$$

- the bound depends on the actual function f being considered
- capacity term could be small if π appropriate
- However, π has to be chosen before seeing the data

Comments

- π is just a technical prior
- allows to distribute the cost of not knowing f beforehand
- if one is lucky, the bound looks like Hoeffding
- goal: guess how likely each function is to be chosen

Randomized Classifiers

Given $\mathcal{G}$ a class of functions

- Deterministic: picks a function g_n and always use it to predict
- Randomized
  - ⋆ construct a distribution ρ_n over $\mathcal{G}$
  - ⋆ for each instance to classify, pick $g \sim \rho_n$
- Error is averaged over ρ_n

$$R(\rho_n) = \rho_n P f$$

$$R_n(\rho_n) = \rho_n P_n f$$

Union Bound (1)

Let π be a (fixed) distribution over $\mathcal{F}$.

- Recall the refined union bound

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq \sqrt{\frac{\log \frac{1}{\pi(f)} + \log \frac{1}{\delta}}{2n}}$$

- Take expectation with respect to ρ_n

$$\rho_n P f - \rho_n P_n f \leq \rho_n \sqrt{\frac{\log \frac{1}{\pi(f)} + \log \frac{1}{\delta}}{2n}}$$

Union Bound (2)

$$\begin{aligned}
\rho_n Pf - \rho_n P_n f &\leq \rho_n \sqrt{\left(-\log \pi(f) + \log \tfrac{1}{\delta}\right)/(2n)} \\
&\leq \sqrt{\left(-\rho_n \log \pi(f) + \log \tfrac{1}{\delta}\right)/(2n)} \\
&\leq \sqrt{\left(K(\rho_n, \pi) + H(\rho_n) + \log \tfrac{1}{\delta}\right)/(2n)}
\end{aligned}$$

- $K(\rho_n, \pi) = \int \rho_n(f) \log \frac{\rho_n(f)}{\pi(f)}\, df$ Kullback-Leibler divergence
- $H(\rho_n) = \int \rho_n(f) \log \rho_n(f)\, df$ Entropy

PAC-Bayesian Refinement

- It is possible to improve the previous bound.
- With probability at least $1-\delta$,

$$\rho_n P f - \rho_n P_n f \leq \sqrt{\frac{K(\rho_n, \pi) + \log 4n + \log \frac{1}{\delta}}{2n-1}}$$

- Good if ρ_n is spread (i.e. large entropy)
- Not interesting if $\rho_n = \delta_{f_n}$

Proof (1)

- Variational formulation of entropy: for any T

$$\rho T(f) \leq \log \pi e^{T(f)} + K(\rho, \pi)$$

- Apply it to $\lambda(Pf - P_n f)^2$

$$\lambda \rho_n (Pf - P_n f)^2 \leq \log \pi e^{\lambda (Pf - P_n f)^2} + K(\rho_n, \pi)$$

- Markov's inequality: with probability $1 - \delta$,

$$\lambda \rho_n (Pf - P_n f)^2 \leq \log \mathbb{E}\left[\pi e^{\lambda (Pf - P_n f)^2}\right] + K(\rho_n, \pi) + \log \tfrac{1}{\delta}$$

Proof (2)

- Fubini

$$\mathbb{E}\left[\pi e^{\lambda(Pf-P_nf)^2}\right] = \pi\mathbb{E}\left[e^{\lambda(Pf-P_nf)^2}\right]$$

- Modified Chernoff bound

$$\mathbb{E}\left[e^{(2n-1)(Pf-P_nf)^2}\right] \leq 4n$$

- Putting together ($\lambda = 2n-1$)

$$(2n-1)\rho_n(Pf-P_nf)^2 \leq K(\rho_n,\pi) + \log 4n + \log\tfrac{1}{\delta}$$

- Jensen $(2n-1)(\rho_n(Pf-P_nf))^2 \leq (2n-1)\rho_n(Pf-P_nf)^2$

Other refinements

- Symmetrization
- Transductive priors
- Rademacher averages
- Chaining
- Generic chaining

Symmetrization

When functions have range in $\{0,1\}$, introduce a ghost sample $Z'_1, \ldots, Z'_n$. Then the set $S_n = \{f(Z_1), \ldots, f(Z_n), f(Z'_1), \ldots, f(Z'_n) : f \in \mathcal{F}\}$ is finite. With probability at least $1-\delta$, $\forall f \in \mathcal{F}$

$$Pf - P_n f \leq C\sqrt{\frac{\log \mathbb{E}|S_n| + \log \frac{1}{\delta}}{n}}\,. \tag{4}$$

- Finite union bound applies to infinite case
- computing $\mathbb{E}|S_n|$ impossible in general
- need combinatorial parameters (e.g. VC dimension)

Transductive priors

If one defines a function $\Pi : \mathcal{Z}^{2n} \to \mathcal{M}_1^+(\mathcal{F})$ which is *exchangeable*, with probability at least $1-\delta$ (over the random choice of a double sample), for all $f \in \mathcal{F}$,

$$P_n' f - P_n f \leq C \sqrt{\frac{\log 1/\Pi(Z_1, \ldots, Z_n, Z_1', \ldots, Z_n')(f) + \log \frac{1}{\delta}}{n}}$$

- Allows the prior to depend on the (double) sample
- Can be useful when there exists a data-independent upper bound

Rademacher averages

No Union Bound

Recall that with probability at least $1-\delta$, for all $f \in \mathcal{F}$

$$Pf - P_n f \leq C\left(\frac{1}{n}\mathbb{E}_n\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^n \sigma_i f(Z_i) + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$$

- No union bound used at this stage, only deviations
- Union bound needed to upper bound the r.h.s.
- Finite case : $\sqrt{\log|\mathcal{F}|/n}$

Chaining

Global Metric Structure

Consider finite covers of the set of function at different scales. Construct a chain of functions that approximate a given function more and more closely. With probability at least $1-\delta$, for all $f \in \mathcal{F}$

$$Pf - P_n f \leq C \left(\frac{1}{\sqrt{n}} \mathbb{E}_n \int_0^\infty \sqrt{\log N(\mathcal{F}, \epsilon, d_n)} d\epsilon + \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

with d_n empirical L_2 metric

Generic chaining

Local Metric Structure

Let $r > 0$ and $(\mathcal{A}_j)_{j \geq 1}$ be partitions of $\mathcal{F}$ of diameter r^{-j} w.r.t. the distance d_n such that $\mathcal{A}_{j+1}$ refines $\mathcal{A}_j$. Previous integral replaced by

$$\inf_{\forall j, \pi^{(j)} \in \mathcal{M}_1^+(\mathcal{F})} \sup_{f \in \mathcal{F}} \sum_{j=1}^{\infty} r^{-j} \sqrt{\log[1/\pi^{(j)} A_j(f)]}$$

- Better adaptation to the local structure of the space
- Equivalent to the Rademacher average (up to log)

Take Home Messages

- Two ingredients of bounds: deviations and union bound $\Rightarrow$ next lecture improves the deviations
- Optimal union bound with metric structure of the function space $\Rightarrow$ generic chaining
- Can introduce a prior into the union bound $\Rightarrow$ best prior depends on the algorithm
- PAC-Bayesian technique: improves the bound when averaged $\Rightarrow$ can be combined with generic chaining

Lecture 2

Variance and Local Rademacher Averages

- Relative error bounds
- Noise conditions
- Localized Rademacher averages

Take Home Messages

- Deviations depend on the variance
- No noise means better rate of convergence
- Noise can be related to variance
- Rademacher averages can be improved with variance

Binomial tails

- $P_n f \sim B(p,n)$ binomial distribution $p = Pf$
- $\mathbb{P}\left[Pf - P_n f \geq t\right] = \sum_{k=0}^{\lfloor n(p-t) \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$
- Can be upper bounded
  - ⋆ Exponential $\left(\frac{1-p}{1-p-t}\right)^{n(1-p-t)} \left(\frac{p}{p+t}\right)^{n(p+t)}$
  - ⋆ Bennett $e^{-\frac{np}{1-p}((1-t/p)\log(1-t/p)+t/p)}$
  - ⋆ Bernstein $e^{-\frac{nt^2}{2p(1-p)+2t/3}}$
  - ⋆ Hoeffding e^{-2nt^2}

Tail behavior

- For small deviations, Gaussian behavior $\approx \exp(-nt^2/2p(1-p))$
  $\Rightarrow$ Gaussian with variance $p(1-p)$

- For large deviations, Poisson behavior $\approx \exp(-3nt/2)$
  $\Rightarrow$ Tails heavier than Gaussian

- Can upper bound with a Gaussian with large (maximum) variance $\exp(-2nt^2)$

Illustration (1)

Maximum variance ($p = 0.5$)

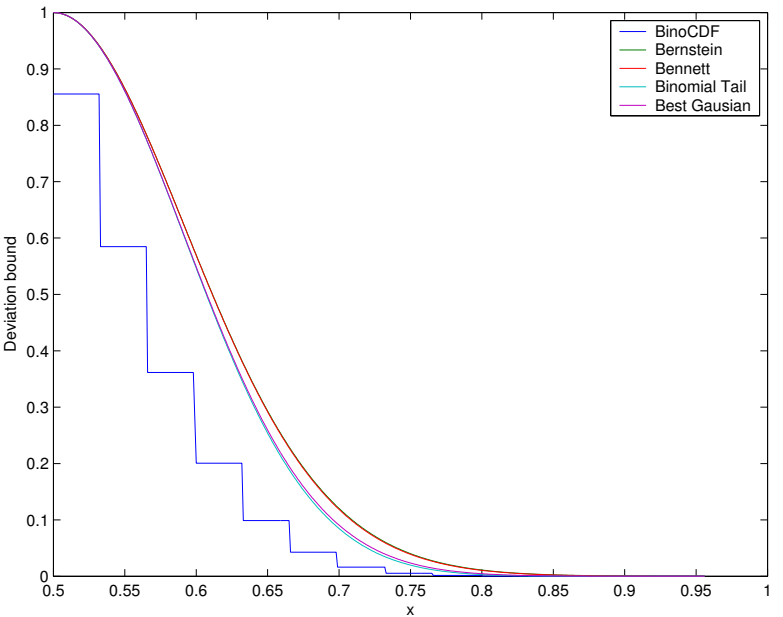
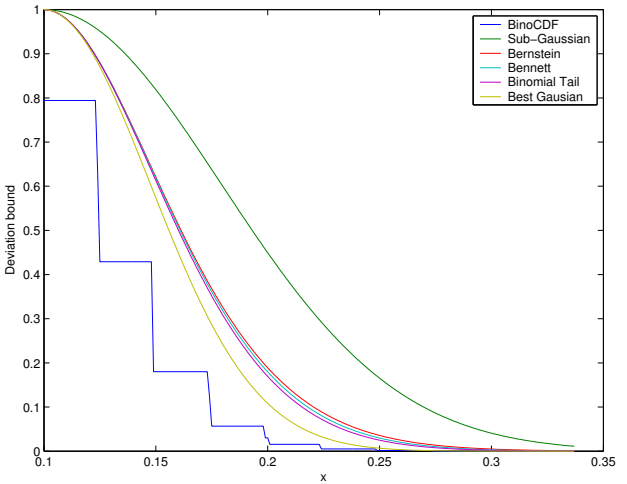
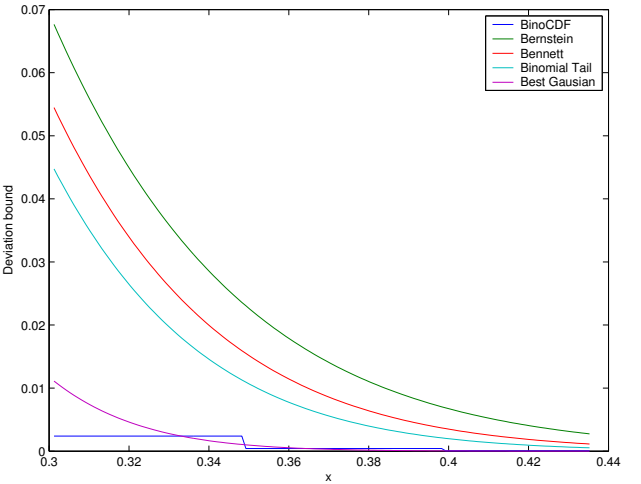


Illustration (2)

Small variance ($p = 0.1$)





Taking the variance into account (1)

- Each function $f \in \mathcal{F}$ has a different variance $Pf(1-Pf) \leq Pf$.
- For each $f \in \mathcal{F}$, by Bernstein's inequality

$$Pf \leq P_n f + \sqrt{\frac{2Pf \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}$$

- The Gaussian part dominates (for Pf not too small, or n large enough), it depends on Pf

$\Rightarrow$ Better bound when Pf is small

Taking the variance into account (2)

- Square root trick:

$$x \leq A\sqrt{x} + B \Rightarrow x \leq A^2 + B + \sqrt{B}A \leq 2A^2 + 2B$$

- Consequence

$$Pf \leq 2P_n f + C\frac{\log \frac{1}{\delta}}{n}\,.$$

$\Rightarrow$ Better bound when $P_n f$ is small

Normalization

- Previous approach was to upper bound

$$\sup_{f \in \mathcal{F}} Pf - P_n f$$

  The supremum is reached at functions with large variance. Those are not the interesting ones
- Here ($f \in \{0,1\}$), $\mathrm{Var}\left[f\right] \leq Pf^2 = Pf$
- Focus of learning: functions with small error Pf (hence small variance)
- Large variance $\Rightarrow$ large risk

Normalization

- The idea is to normalize functions by their variance
- After normalization, fluctuations are more "uniform"

$$\sup_{f \in \mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}$$

All functions on the same scale

$\Rightarrow$ The normalized supremum takes the learning method into account.

Relative deviations

Vapnik-Chervonenkis 1974

For $\delta > 0$ with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}$$

Consequence

From the square root trick we get

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + 2\sqrt{P_n f \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}} + 4\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}$$

Proof sketch

1. Symmetrization

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\frac{Pf-P_nf}{\sqrt{Pf}}\geq t\right]\leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}\frac{P'_nf-P_nf}{\sqrt{(P_nf+P'_nf)/2}}\geq t\right]$$

2. Randomization

$$\cdots=2\mathbb{E}\left[\mathbb{P}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{\frac{1}{n}\sum_{i=1}^n\sigma_i(f(Z'_i)-f(Z_i))}{\sqrt{(P_nf+P'_nf)/2}}\geq t\right]\right]$$

3. Tail bound

Zero noise

Ideal situation :

- g_n empirical risk minimizer
- Bayes classifier in the class $\mathcal{G}$
- $R^* = 0$ (no noise)

In that case

- $R_n(g_n) = 0$

$\Rightarrow \; R(g_n) = O(\frac{d \log n}{n})$.

Interpolating between rates ?

- Rates are not correctly estimated by this inequality
- Consequence of relative error bounds

$$\begin{aligned} Pf_n \quad \leq \quad & Pf^* + 2\sqrt{Pf^* \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}} \\ & + 4 \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n} \end{aligned}$$

- The quantity which is small is not Pf^* but $Pf_n - Pf^*$
- But relative error bounds do not apply to differences

Definitions

- $\eta(x) = \mathbb{E}[Y|X=x] = 2\mathbb{P}[Y=1|X=x] - 1$ is the regression function
- $t(x) = \operatorname{sgn}\eta(x)$ is the target function or Bayes classifier (Bayes risk $R^* = \mathbb{E}[n(X)]$)
- in the deterministic case $Y = t(X)$ ($\mathbb{P}[Y=1|X] \in \{0,1\}$)
- in general, noise level

$$\begin{aligned} n(x) &= \min(\mathbb{P}[Y=1|X=x], 1-\mathbb{P}[Y=1|X=x]) \\ &= (1-\eta(x))/2 \end{aligned}$$

Approximation/Estimation

- Bayes risk

$$R^* = \inf_g R(g) \, .$$

  Best risk a deterministic function can have (risk of the target function, or Bayes classifier).

- Decomposition: $R(g^*) = \inf_{g \in \mathcal{G}} R(g)$

$$R(g_n) - R^* = \underbrace{R(g) - R^*}_{\text{Approximation}} + \underbrace{R(g_n) - R(g^*)}_{\text{Estimation}}$$

- Only the estimation error is random (i.e. depends on the data).

Intermediate noise

Instead of assuming that $|\eta(x)| = 1$ (i.e. $n(x) = 0$), the deterministic case, one can assume that n is well-behaved.
Two kinds of assumptions

- n not too close to $1/2$

- n not often too close to $1/2$

Massart Condition

- For some $c > 0$, assume

$$|\eta(X)| > \frac{1}{c} \text{ almost surely}$$

- There is no region where the decision is completely random

- Noise bounded away from $1/2$

Tsybakov Condition

Let $\alpha \in [0,1]$, equivalent conditions

$$(1) \quad \exists c > 0,\ \forall g \in \{-1,1\}^{\mathcal{X}},$$
$$\mathbb{P}\left[g(X)\eta(X) \leq 0\right] \leq c(R(g) - R^*)^{\alpha}$$

$$(2) \quad \exists c > 0,\ \forall A \subset \mathcal{X},\ \int_A dP(x) \leq c(\int_A |\eta(x)| dP(x))^{\alpha}$$

$$(3) \quad \exists B > 0,\ \forall t \geq 0,\ \mathbb{P}\left[|\eta(X)| \leq t\right] \leq B t^{\frac{\alpha}{1-\alpha}}$$

Equivalence

- $(1) \Leftrightarrow (2)$ Recall $R(g) - R^* = \mathbb{E}\left[|\eta(X)| 1_{[g\eta \leq 0]}\right]$. For each function g, there exists a set A such that $1_{[A]} = 1_{[g\eta \leq 0]}$
- $(2) \Rightarrow (3)$ Let $A = \{x : |\eta(x)| \leq t\}$

$$\begin{aligned} \mathbb{P}\left[|\eta| \leq t\right] = \int_A dP(x) &\leq c\left(\int_A |\eta(x)| dP(x)\right)^\alpha \\ &\leq ct^\alpha \left(\int_A dP(x)\right)^\alpha \end{aligned}$$

$$\Rightarrow \mathbb{P}\left[|\eta| \leq t\right] \leq c^{\frac{1}{1-\alpha}} t^{\frac{\alpha}{1-\alpha}}$$

- $(3) \Rightarrow (1)$

$$
\begin{aligned}
R(g) - R^* &= \mathbb{E}\left[|\eta(X)|\, 1_{[g\eta \le 0]}\right] \\
&\ge t\mathbb{E}\left[1_{[g\eta \le 0]} 1_{[|\eta| > t]}\right] \\
&= t\mathbb{P}\left[|\eta| > t\right] - t\mathbb{E}\left[1_{[g\eta > 0]} 1_{[|\eta| > t]}\right] \\
&\ge t(1 - Bt^{\frac{\alpha}{1-\alpha}}) - t\mathbb{P}\left[g\eta > 0\right] = t(\mathbb{P}\left[g\eta \le 0\right] - Bt^{\frac{\alpha}{1-\alpha}})
\end{aligned}
$$

Take $t = \left(\frac{(1-\alpha)\mathbb{P}[g\eta \le 0]}{B}\right)^{(1-\alpha)/\alpha}$

$$
\Rightarrow \mathbb{P}\left[g\eta \le 0\right] \le \frac{B^{1-\alpha}}{(1-\alpha)^{(1-\alpha)}\alpha^{\alpha}} (R(g) - R^*)^{\alpha}
$$

Remarks

- α is in $[0,1]$ because

$$R(g) - R^* = \mathbb{E}\left[|\eta(X)| 1_{[g\eta \leq 0]}\right] \leq \mathbb{E}\left[1_{[g\eta \leq 0]}\right]$$

- $\alpha = 0$ no condition

- $\alpha = 1$ gives Massart's condition

Consequences

- Under Massart's condition

$$\mathbb{E}\left[\left(1_{[g(X)\neq Y]} - 1_{[t(X)\neq Y]}\right)^2\right] \leq c(R(g) - R^*)$$

- Under Tsybakov's condition

$$\mathbb{E}\left[\left(1_{[g(X)\neq Y]} - 1_{[t(X)\neq Y]}\right)^2\right] \leq c(R(g) - R^*)^\alpha$$

Relative loss class

- $\mathcal{F}$ is the loss class associated to $\mathcal{G}$
- The relative loss class is defined as

$$\tilde{\mathcal{F}} = \{f - f^* : f \in \mathcal{F}\}$$

- It satisfies

$$Pf^2 \leq c(Pf)^\alpha$$

Finite case

- Union bound on $\tilde{\mathcal{F}}$ with Bernstein's inequality would give

$$Pf_n - Pf^* \leq P_n f_n - P_n f^* + \sqrt{\frac{8c(Pf_n - Pf^*)^\alpha \log \frac{N}{\delta}}{n}} + \frac{4 \log \frac{N}{\delta}}{3n}$$

- Consequence when $f^* \in \mathcal{F}$ (but $R^* > 0$)

$$Pf_n - Pf^* \leq C \left(\frac{\log \frac{N}{\delta}}{n} \right)^{\frac{1}{2-\alpha}}$$

always better than $n^{-1/2}$ for $\alpha > 0$

Local Rademacher average

- Definition

$$\mathcal{R}(\mathcal{F}, r) = \mathbb{E}\left[\sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n f\right]$$

- Allows to generalize the previous result
- Computes the capacity of a small ball in $\mathcal{F}$ (functions with small variance)
- Under noise conditions, small variance implies small error

Sub-root functions

Definition

A function $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ is sub-root if

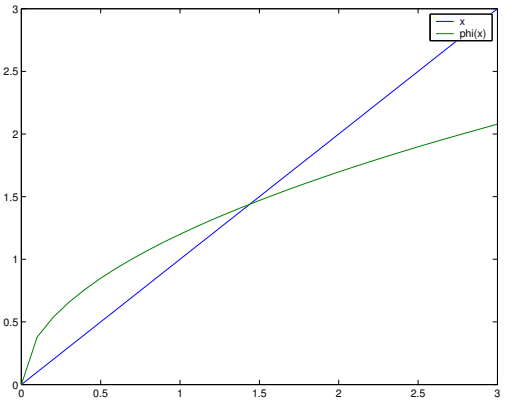
- ψ is non-decreasing
- ψ is non negative
- $\psi(r)/\sqrt{r}$ is non-increasing

Sub-root functions

Properties

A sub-root function

- is continuous
- has a unique fixed point $\psi(r^*) = r^*$



Star hull

- Definition

$$\star\mathcal{F} = \{\alpha f : f \in \mathcal{F},\ \alpha \in [0,1]\}$$

- Properties

$$\mathcal{R}_n(\star\mathcal{F}, r) \text{ is sub-root}$$

- Entropy of $\star\mathcal{F}$ is not much bigger than entropy of $\mathcal{F}$

Result

- r^* fixed point of $\mathcal{R}(\star\mathcal{F}, r)$
- Bounded functions

$$Pf - P_n f \leq C\left(\sqrt{r^* \mathsf{Var}\left[f\right]} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right)$$

- Consequence for variance related to expectation ($\mathsf{Var}\left[f\right] \leq c(Pf)^\beta$)

$$Pf \leq C\left(P_n f + (r^*)^{\frac{1}{2-\beta}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right)$$

Consequences

- For VC classes $\mathcal{R}(\mathcal{F}, r) \leq C\sqrt{\frac{rh}{n}}$ hence $r^* \leq C\frac{h}{n}$
- Rate of convergence of $P_n f$ to Pf in $O(1/\sqrt{n})$
- But rate of convergence of Pf_n to Pf^* is $O(1/n^{1/(2-\alpha)})$

Only condition is $t \in \mathcal{G}$ but can be removed by SRM/Model selection

Proof sketch (1)

- Talagrand's inequality

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + c\sqrt{\sup_{f\in\mathcal{F}} \mathrm{Var}\left[f\right]/n} + c'/n$$

- Peeling of the class

$$\mathcal{F}_k = \{f : \mathrm{Var}\left[f\right] \in [x^k, x^{k+1})\}$$

Proof sketch (2)

- Application

$$\sup_{f\in\mathcal{F}_k} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}_k} Pf - P_n f\right] + c\sqrt{x\mathsf{Var}\left[f\right]/n} + c'/n$$

- Symmetrization

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq 2\mathcal{R}(\mathcal{F}, x\mathsf{Var}\left[f\right]) + c\sqrt{x\mathsf{Var}\left[f\right]/n} + c'/n$$

Proof sketch (3)

- We need to 'solve' this inequality. Things are simple if $\mathcal{R}$ behave like a square root, hence the sub-root property

$$Pf - P_n f \leq 2\sqrt{r^* \mathrm{Var}\left[f\right]} + c\sqrt{x \mathrm{Var}\left[f\right]/n} + c'/n$$

- Variance-expectation

$$\mathrm{Var}\left[f\right] \leq c(Pf)^\alpha$$

  Solve in Pf

Data-dependent version

- As in the global case, one can use data-dependent local Rademcher averages

$$\mathcal{R}_n(\mathcal{F}, r) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F} : Pf^2 \leq r} R_n f \right]$$

- Using concentration one can also get

$$Pf \leq C \left(P_n f + (r_n^*)^{\frac{1}{2-\alpha}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

where r_n^* is the fixed point of a sub-root upper bound of $\mathcal{R}_n(\mathcal{F}, r)$

Discussion

- Improved rates under low noise conditions
- Interpolation in the rates
- Capacity measure seems 'local',
- but depends on all the functions,
- after appropriate rescaling: each $f \in \mathcal{F}$ is considered at scale r/Pf^2

Take Home Messages

- Deviations depend on the variance
- No noise means better rate of convergence
- Noise can be related to variance $\Rightarrow$ noise can be quantified
- Rademacher averages can be improved with variance $\Rightarrow$ localized

Lecture 3

Loss Functions

- Properties
- Consistency
- Examples
- Losses and noise

Motivation (1)

- ERM: minimize $\sum_{i=1}^{n} 1_{[g(X_i) \neq Y_i]}$ in a set $\mathcal{G}$

$\Rightarrow$ Computationally hard

$\Rightarrow$ Smoothing

  - $\star$ Replace binary by real-valued functions
  - $\star$ Introduce smooth loss function

$$\sum_{i=1}^{n} \ell(g(X_i), Y_i)$$

Motivation (2)

- Hyperplanes in infinite dimension have
  - ⋆ infinite VC-dimension
  - ⋆ but finite scale-sensitive dimension (to be defined later)

$\Rightarrow$ It is good to have a scale

$\Rightarrow$ This scale can be used to give a confidence (i.e. estimate the density)

- However, losses do not need to be related to densities
- Can get bounds in terms of margin error instead of empirical error (smoother $\longrightarrow$ easier to optimize for model selection)

Take Home Messages

- Convex losses for computational convenience
- No effect asymptotically
- Influence on the rate of convergence
- Classification or regression losses

Margin

- It is convenient to work with (symmetry of $+1$ and -1)

$$\ell(g(x), y) = \phi(yg(x))$$

- $yg(x)$ is the margin of g at (x, y)
- Loss

$$L(g) = \mathbb{E}\left[\phi(Yg(X))\right], \quad L_n(g) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i g(X_i))$$

- Loss class $\mathcal{F} = \{f : (x, y) \mapsto \phi(yg(x)) : g \in \mathcal{G}\}$

Minimizing the loss

- Decomposition of $L(g)$

$$\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[(1+\eta(X))\phi(g(X))+(1-\eta(X))\phi(-g(X))|X\right]\right]$$

- Minimization for each x

$$H(\eta)=\inf_{\alpha\in\mathbb{R}}\left((1+\eta)\phi(\alpha)/2+(1-\eta)\phi(-\alpha)/2\right)$$

- $L^*:=\inf_g L(g)=\mathbb{E}\left[H(\eta(X))\right]$

Classification-calibrated

- A minimal requirement is that the minimizer in $H(\eta)$ has the correct sign (that of the target t or that of η).

- Definition
  ϕ is classification-calibrated if, for any $\eta \neq 0$

$$\inf_{\alpha:\alpha\eta\leq 0}(1+\eta)\phi(\alpha)+(1-\eta)\phi(-\alpha) > \inf_{\alpha\in\mathbb{R}}(1+\eta)\phi(\alpha)+(1-\eta)\phi(-\alpha)$$

- This means the infimum is achieved for an α of the correct sign (and not for an α of the wrong sign, except possibly for $\eta = 0$).

Consequences (1)

Results due to (Jordan, Bartlett and McAuliffe 2003)

- ϕ is classification-calibrated iff for all sequences g_i and every probability distribution P,

$$L(g_i) \to L^* \;\Rightarrow\; R(g_i) \to R^*$$

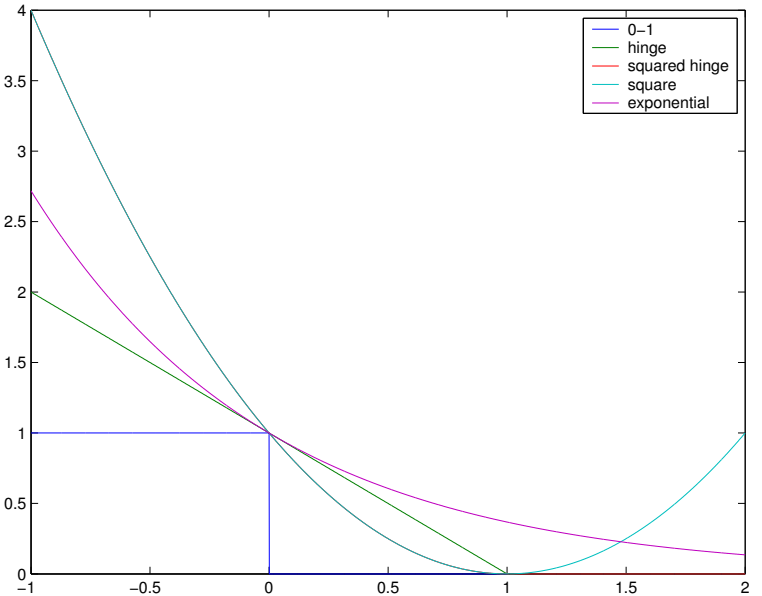
- When ϕ is convex (convenient for optimization) ϕ is classification-calibrated iff it is differentiable at 0 and $\phi'(0) < 0$

Consequences (2)

- Let $H^-(\eta) = \inf_{\alpha:\alpha\eta\leq 0} ((1+\eta)\phi(\alpha)/2 + (1-\eta)\phi(-\alpha)/2)$
- Let $\psi(\eta)$ be the largest convex function below $H^-(\eta) - H(\eta)$
- One has

$$\psi(R(g) - R^*) \leq L(g) - L^*$$

Examples (1)


0–1
hinge
squared hinge
square
exponential
4
3.5
3
2.5
2
1.5
1
0.5
0
−1
−0.5
0
0.5
1
1.5
2

Examples (2)

- Hinge loss

$$\phi(x) = \max(0, 1 - x), \ \psi(x) = x$$

- Squared hinge loss

$$\phi(x) = \max(0, 1 - x)^2, \ \psi(x) = x^2$$

- Square loss

$$\phi(x) = (1 - x)^2, \ \psi(x) = x^2$$

- Exponential

$$\phi(x) = \exp(-x), \ \psi(x) = 1 - \sqrt{1 - x^2}$$

Low noise conditions

- Relationship can be improved under low noise conditions
- Under Tsybakov's condition with exponent α and constant c,

$$c(R(g) - R^*)^\alpha \psi((R(g) - R^*)^{1-\alpha}/2c) \leq L(g) - L^*$$

- Hinge loss (no improvement)

$$R(g) - R^* \leq L(g) - L^*$$

- Square loss or squared hinge loss

$$R(g) - R^* \leq (4c(L(g) - L^*))^{\frac{1}{2-\alpha}}$$

Estimation error

- Recall that Tsybakov condition implies $Pf^2 \leq c(Pf)^\alpha$ for the relative loss class (with $0-1$ loss)

- What happens for the relative loss class associated to ϕ ?

- Two possibilities
  - ⋆ Strictly convex loss (can modify the metric on $\mathbb{R}$)
  - ⋆ Piecewise linear

Strictly convex losses

- Noise behavior controlled by modulus of convexity
- Result

$$\delta\left(\frac{\sqrt{Pf^2}}{K}\right) \leq Pf/2$$

  with K Lipschitz constant of ϕ and δ modulus of convexity of $L(g)$ with respect to $\|f-g\|_{L_2(P)}$

- Not related to noise exponent

Piecewise linear losses

- Noise behavior related to noise exponent

- Result for hinge loss

$$Pf^2 \leq CPf^\alpha$$

  if initial class $\mathcal{G}$ is uniformly bounded

Estimation error

- With bounded and Lipschitz loss with convexity exponent γ, for a convex class $\mathcal{G}$,

$$L(g) - L(g^*) \leq C \left((r^*)^{\frac{2}{\gamma}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

- Under Tsybakov's condition for the hinge loss (and general $\mathcal{G}$) $Pf^2 \leq CPf^\alpha$

$$L(g) - L(g^*) \leq C \left((r^*)^{\frac{1}{2-\alpha}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

Examples

Under Tsybakov's condition

- Hinge loss

$$R(g) - R^* \leq L(g^*) - L^* + C\left((r^*)^{\frac{1}{2-\alpha}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right)$$

- Squared hinge loss or square loss $\delta(x) = cx^2$, $Pf^2 \leq CPf$

$$R(g) - R^* \leq C\left(L(g^*) - L^* + C'(r^* + \frac{\log\frac{1}{\delta} + \log\log n}{n})\right)^{\frac{1}{2-\alpha}}$$

Classification vs Regression losses

- Consider a classification-calibrated function ϕ
- It is a classification loss if $L(t) = L^*$
- otherwise it is a regression loss

Classification vs Regression losses

- Square, squared hinge, exponential losses
  - ⋆ Noise enters relationship between risk and loss
  - ⋆ Modulus of convexity enters in estimation error
- Hinge loss
  - ⋆ Direct relationship between risk and loss
  - ⋆ Noise enters in estimation error

$\Rightarrow$ Approximation term not affected by noise in second case

$\Rightarrow$ Real value does not bring probability information in second case

Take Home Messages

- Convex losses for computational convenience
- No effect asymptotically $\Rightarrow$ Classification calibrated property
- Influence on the rate of convergence $\Rightarrow$ approximation or estimation, related to noise level
- Classification or regression losses $\Rightarrow$ depends on what you want to estimate

Lecture 4

SVM

- Computational aspects
- Capacity Control
- Universality
- Special case of RBF kernel

Take Home Messages

- Smooth parametrization
- Regularization
- RBF: universal, flexible, locally preserving

Formulation (1)

- Soft margin

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

- Convex objective function and convex constraints
- Unique solution
- Efficient procedures to find it

$\longrightarrow$ Is it the right criterion ?

Formulation (2)

- Soft margin

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- Optimal value of ξ_i

$$\xi_i^* = \max(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b))$$

- Substitute above to get

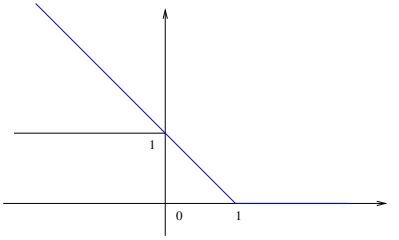
$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b))$$

Regularization

General form of regularization problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c(y_i f(x_i)) + \lambda \|f\|^2$$

$\longrightarrow$ Capacity control by regularization with convex cost



Loss Function

$$\phi(Yf(X)) = \max(0, 1 - Yf(X))$$

- Convex, non-increasing, upper bounds $1_{[Yf(X)\leq 0]}$
- Classification-calibrated
- Classification type ($L^* = L(t)$)

$$R(g) - R^* \leq L(g) - L^*$$

Regularization

Choosing a kernel corresponds to

- Choose a sequence (a_k)
- Set

$$\|f\|^2 := \sum_{k \geq 0} a_k \int |f^{(k)}|^2 dx$$

$\Rightarrow$ penalization of high order derivatives (high frequencies)

$\Rightarrow$ enforce smoothness of the solution

Capacity: VC dimension

- The VC dimension of the set of hyperplanes is $d+1$ in $\mathbb{R}^d$.
  Dimension of feature space ?
  ∞ for RBF kernel
- w choosen in the span of the data ($w = \sum \alpha_i y_i \mathbf{x}_i$)
  The span of the data has dimension m for RBF kernel ($k(.,x_i)$ linearly independent)
- The VC bound does not give any information

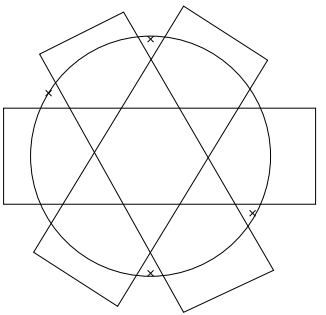
$$\sqrt{\frac{h}{n}} = 1$$

$\Rightarrow$ Need to take the margin into account

Capacity: Shattering dimension

Hyperplanes with Margin

If $\|x\| \leq R$,
$vc(\text{hyperplanes with margin } \rho, 1) \leq R^2/\rho^2$



Margin

- The shattering dimension is related to the margin
- Maximizing the margin means minimizing the shattering dimension
- Small shattering dimension $\Rightarrow$ good control of the risk

$\Rightarrow$ this control is automatic (no need to choose the margin beforehand)

$\Rightarrow$ but requires tuning of regularization parameter

Capacity: Rademacher Averages (1)

- Consider hyperplanes with $\|w\| \leq M$
- Rademacher average

$$\frac{M}{n\sqrt{2}}\sqrt{\sum_{i=1}^{n} k(x_i, x_i)} \leq \mathcal{R}_n \leq \frac{M}{n}\sqrt{\sum_{i=1}^{n} k(x_i, x_i)}$$

- Trace of the Gram matrix
- Notice that $\mathcal{R}_n \leq \sqrt{R^2/(n^2\rho^2)}$

Rademacher Averages (2)

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\|w\|\leq M} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \left\langle w, \delta_{x_i}\right\rangle\right] \\
&\quad = \quad \mathbb{E}\left[\sup_{\|w\|\leq M} \left\langle w, \frac{1}{n}\sum_{i=1}^{n}\sigma_i\delta_{x_i}\right\rangle\right] \\
&\quad \leq \quad \mathbb{E}\left[\sup_{\|w\|\leq M} \|w\| \left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\delta_{x_i}\right\|\right] \\
&\quad = \quad \frac{M}{n}\mathbb{E}\left[\sqrt{\left\langle \sum_{i=1}^{n}\sigma_i\delta_{x_i}, \sum_{i=1}^{n}\sigma_i\delta_{x_i}\right\rangle}\right]
\end{aligned}
$$

Rademacher Averages (3)

$$\begin{aligned}
&\frac{M}{n}\mathbb{E}\left[\sqrt{\left\langle\sum_{i=1}^n \sigma_i\delta_{x_i}, \sum_{i=1}^n \sigma_i\delta_{x_i}\right\rangle}\right] \\
&\quad\leq \quad \frac{M}{n}\sqrt{\mathbb{E}\left[\left\langle\sum_{i=1}^n \sigma_i\delta_{x_i}, \sum_{i=1}^n \sigma_i\delta_{x_i}\right\rangle\right]} \\
&\quad= \quad \frac{M}{n}\sqrt{\mathbb{E}\left[\sum_{i,j} \sigma_i\sigma_j \left\langle\delta_{x_i}, \delta_{x_j}\right\rangle\right]} \\
&\quad= \quad \frac{M}{n}\sqrt{\sum_{i=1}^n k(x_i, x_i)}
\end{aligned}$$

Improved rates – Noise condition

- Under Massart's condition ($|\eta| > \eta_0$), with $\|g\|_\infty \leq M$

$$\mathbb{E}\left[\left(\phi(Yg(X)) - \phi(Yt(X))\right)^2\right] \leq (M-1+2/\eta_0)(L(g)-L^*)\,.$$

$\longrightarrow$ If noise is nice, variance linearly related to expectation

$\longrightarrow$ Estimation error of order r^* (of the class $\mathcal{G}$)

Improved rates – Capacity (1)

- r_n^* related to decay of eigenvalues of the Gram matrix

$$r_n^* \leq \frac{c}{n} \min_{d \in \mathbb{N}} \left(d + \sqrt{\sum_{j>d} \lambda_j} \right)$$

- Note that $d = 0$ gives the trace bound
- r_n^* always better than the trace bound (equality when λ_i constant)

Improved rates – Capacity (2)

Example: exponential decay

- $\lambda_i = e^{-\alpha i}$
- Global Rademacher of order $\frac{1}{\sqrt{n}}$
- r_n^* of order

$$\frac{\log n}{n}$$

Kernel

Why is it good to use kernels ?

- Gaussian kernel (RBF)

$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

- σ is the width of the kernel

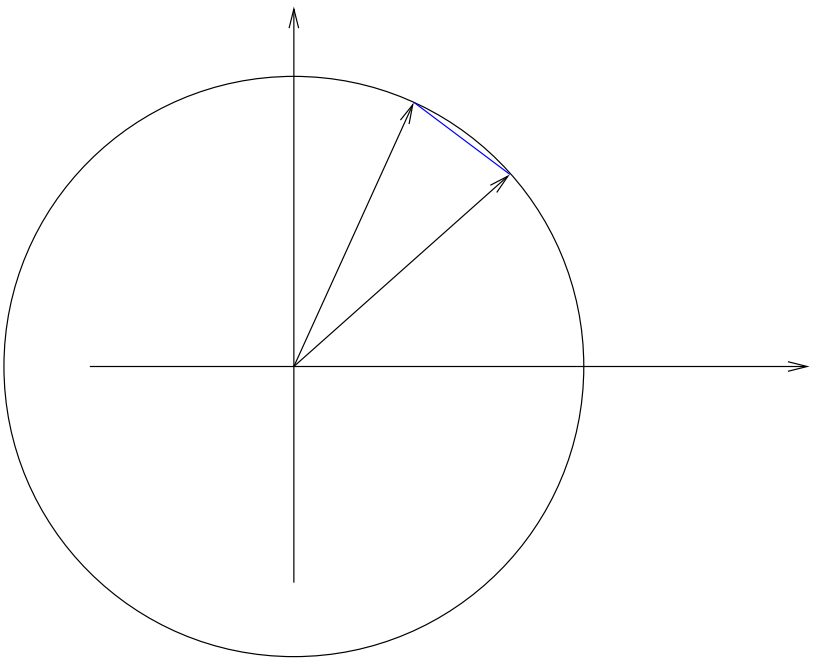
$\longrightarrow$ What is the geometry of the feature space ?

RBF

Geometry

- Norms
$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = e^0 = 1$$
$\longrightarrow$ sphere of radius 1
- Angles
$$\cos(\widehat{\Phi(x), \Phi(y)}) = \left\langle \frac{\Phi(x)}{\|\Phi(x)\|}, \frac{\Phi(y)}{\|\Phi(y)\|} \right\rangle = e^{-\|x-y\|^2/2\sigma^2} \geq 0$$
$\longrightarrow$ Angles less than 90 degrees
- $\Phi(x) = k(x, .) \geq 0$
$\longrightarrow$ positive quadrant

RBF



RBF

Differential Geometry

- Flat Riemannian metric

  $\longrightarrow$ 'distance' along the sphere is equal to distance in input space

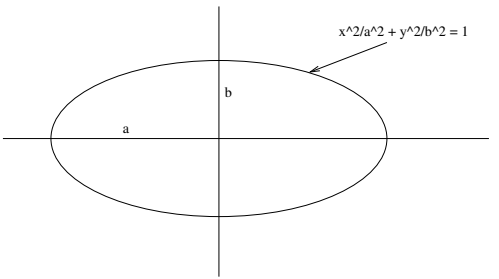
- Distances are contracted

  $\longrightarrow$ 'shortcuts' by getting outside the sphere

RBF

Geometry of the span

Ellipsoid



- $K = (k(x_i, x_j))$ Gram matrix
- Eigenvalues $\lambda_1, \ldots, \lambda_m$
- Data points mapped to ellispoid with lengths $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_m}$

RBF

Universality

- Consider the set of functions

$$\mathcal{H} = \mathrm{span}\{k(x, \cdot) : \ x \in \mathcal{X}\}$$

- $\mathcal{H}$ is dense in $C(\mathcal{X})$

$\longrightarrow$ Any continuous function can be approximated (in the $\|\|_\infty$ norm) by functions in $\mathcal{H}$

$\Rightarrow$ with enough data one can construct any function

RBF

Eigenvalues

- Exponentially decreasing
- Fourier domain: exponential penalization of derivatives
- Enforces smoothness with respect to the Lebesgue measure in input space

RBF

Induced Distance and Flexibility

- $\sigma \to 0$
  1-nearest neighbor in input space
  Each point in a separate dimension, everything orthogonal

- $\sigma \to \infty$
  linear classifier in input space
  All points very close on the sphere, initial geometry

- Tuning σ allows to try all possible intermediate combinations

RBF

Ideas

- Works well if the Euclidean distance is good
- Works well if decision boundary is smooth
- Adapt smoothness via σ
- Universal

Choosing the Kernel

- Major issue of current research
- Prior knowledge (e.g. invariances, distance)
- Cross-validation (limited to 1-2 parameters)
- Bound (better with convex class)

$\Rightarrow$ Lots of open questions...

Take Home Messages

- Smooth parametrization $\Rightarrow$ regularization and smoothness parameters
- Regularization $\Rightarrow$ soft capacity control
- RBF: universal, flexible, locally preserving $\Rightarrow$ trust the structure locally and do sensible things globally