Remarks on Statistical Learning Theory

Olivier Bousquet
Department of Empirical Inference
Max Planck Institute of Biological Cybernetics
olivier.bousquet@tuebingen.mpg.de

Machine Learning Summer School, August 2003

Learning Theory: some informal thoughts

- Error bars vs. error bounds

- What is a good bound ?

- What is the best approach ?

$\Rightarrow$ This is a personal view, do not trust me too much !

Disclaimer

When you see this sign

□

this means:

- Strong claim
- No formal proof
- Personal opinion
- You may disagree

Possible error estimates

- Empirical error (sample S)

$$R_S(g_S)$$

- Holdout error (T independent sample)

$$R_T(g_S)$$

- Cross-validation error

$$\frac{1}{m}\sum_{i=1}^{m} R_{S_i}(g_{S\setminus i})$$

- Leave-one-out error

$$\frac{1}{n}\sum_{i=1}^{n} R_{Z_i}(g_{S\setminus Z_i})$$

$\Rightarrow$ Picture

Bias and variance

- Variance of empirical error can be controlled (bounds)
- But favorably biased
- Leave-one-out error almost unbiased

$$\mathbb{E}\left[R_{loo}(g_n)\right] = \mathbb{E}\left[R(g_{n-1})\right]$$

- But hard to control the variance

What to prefer ?

- Depends on what you want to do
- Bounds give you guarantees
- Unbiased estimates may be good in practice
- Bounds tell you what is important (e.g. margin)

Error bars and error bounds

- Error bar = variance estimate

- How to use variance ? Chebyshev

$$\mathbb{P}\left[X - \mathbb{E}\left[X\right] \geq t\right] \leq \frac{\mathrm{Var}\left[X\right]}{t^2}$$

Inversion

$$X \leq \mathbb{E}\left[X\right] + \sqrt{\frac{\mathrm{Var}\left[X\right]}{\delta}}$$

- Exponential bounds yield (Gaussian case)

$$X \leq \mathbb{E}[X] + \sqrt{\mathsf{Var}[X] \log \frac{1}{\delta}}$$

- Numerically the difference may be small but conceptually it matters (exponential means control of all the moments)

Error bars and error bounds

Frequentist interpretation

- Bayesian approach:
  - ⋆ Pick a target (according to prior)
  - ⋆ Pick a sample (according to distribution)
  - ⋆ Label the sample

  ⇒ Error bars hold for most repeats of the above
- SLT approach
  - ⋆ Target is fixed
  - ⋆ Pick a sample

  ⇒ Error bounds hold for most samples

Error bars and error bounds

Frequentist interpretation ⇒ □For a given problem, error bars don't say anything

- Variance instead of full distribution
- Correct only if the prior is correct
- No way to test its correctness, only one experiment is allowed

⇒ Use them if you want but be aware of their (lack of) meaning !

What is a good bound ?

- Classification error between 0 and $1/2$

- Most theoretical bounds are useless (value $>> 1$)

- How to make them non-trivial ?

$\Rightarrow$ Here trivial does not mean easy but larger than 1

What is a good bound ?

- Depends on what you want to do with it
- Three levels of usage □

1. Quantitative
2. Model selection
3. Qualitative

First level

Obstacles

- Behavior of the error is complex
- Used techniques sharp in the asymptotic regime
- More precise techniques may exist but are much more messy
- Small bounds are unreadable

First level

Obstacles

- Behavior of the error is complex
- Used techniques sharp in the asymptotic regime
- More precise techniques may exist but are much more messy
- Small bounds are unreadable

$\Rightarrow$ ☐Hopeless ! use CV

Second level

Model selection

- Typical bounds behavior (picture)

- What matters is the location of the minimum

Second level

Model selection

- Typical bounds behavior (picture)

- What matters is the location of the minimum

$\Rightarrow$ ☐Little hope ! use CV if possible

Third level

Qualitative

- Use the quantities appearing in the bound to get new algorithms

- Does not give the best choice of the parameters

- But gives some robustness

- Avoid a posteriori justifications !

Third level

Qualitative

- Use the quantities appearing in the bound to get new algorithms

- Does not give the best choice of the parameters

- But gives some robustness

- Avoid a posteriori justifications !

$\Rightarrow$ □Very reasonable !

Third level

Example

- Large margin *correlated* to low error

- Hence one can maximize the margin

Wrong approach

- Large margin means low VC dimension

- Hence one should maximize the margin

Why a posteriori justifications are wrong ?

- Given a class of functions $\mathcal{F}$
- Define a (non-negative) functional $\Omega(f)$
- Obviously if $x \leq y$

$$\{\Omega(f) \leq x\} \subset \{\Omega(f) \leq y\}$$

- Hence $VC\{\Omega(f) \leq x\}$ is a non-decreasing function of x !

$\Rightarrow$ Algorithm should minimize $\Omega(f)$!

$\Rightarrow$ Arbitrary ! Same as choosing p in the refined union bound !

What is a good bound ?

- Forget about the value
- Try to capture meaningful behavior
- Do not put quantities in by hand
- Find what is responsible for deviations and how it influences them

What is the best approach ?

- Kernel methods

- Gaussian processes

- MDL

What is the best approach ?

- Kernel methods

- Gaussian processes

- MDL

$\Rightarrow$ □Slight differences but overall the same (fit + complexity)

What is the best approach ?

Do we have theoretical guarantees ?

- Kernel methods: theory justifies margin and high dimension, not kernels !

- GP: no theory but could be put in the same framework

- MDL: short means few possibilities, easy bounds !

What is the best approach ?

$\Rightarrow$ Depends on the nature of your prior knowledge

- Similarity measure ? Try kernels

- Nice coding scheme ? Try MDL

- Covariance intuition ? Use GP

Overall it is a matter of taste, flexibility and computational constraints.

What is learning theory for ?

- Bounds: if correctly used, OK, but just one aspect

- Try to formalize other learning settings

- NEEDED: New ways to encode prior knowledge

[Vapnik] Nothing is more practical than a good theory