

GI01/4C55: Supervised Learning

5. Learning Theory

October 31, 2005

Massimiliano Pontil

Today's plan

- Model selection problem
- Uniform error bounds
- Structural risk minimization
- VC–dimension
- Bias/variance trade-off

Bibliography: These lecture notes are available at:

<http://www.cs.ucl.ac.uk/staff/M.Pontil/courses/index-SL05.htm>

Lecture notes are in small part based on §7 of Hastie, Tibshirani & Friedman

Supervised learning model (I) – review

$P(\mathbf{x}, y)$: **fixed but unknown** probability density (defines the learning environment)

Expected error:

$$\mathcal{E}(f) := \mathbf{E}[V(y, f(\mathbf{x}))] = \int V(y, f(\mathbf{x})) dP(\mathbf{x}, y)$$

where $V : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a **loss function**

- Our goal is to minimize $\mathcal{E}$
- Optimal solution: $f^* := \operatorname{argmin}_f \mathcal{E}(f)$
- We cannot compute f^* because P is unknown

Supervised learning model (II)

We have encountered different loss functions

- square loss: $V(y, z) = (y - z)^2$
- misclassification loss (0-1 loss): $V(y, z) = 1$ if $y \neq z$ and zero otherwise (here $y, z \in \{c_1, \dots, c_K\}$)
- logistic regression: $V(y, z) = y \log(1 + e^{-z}) + (1 - y) \log(1 + e^z)$

We'll see more interesting loss functions later in the course when we speak about support vector machines

Supervised learning model (III)

$P(\mathbf{x}, y)$ is unknown $\Rightarrow$ cannot compute f^*

We are only given an i.i.d. sample from P :

$$S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$$

We approximate (replace) the expected error $\mathcal{E}(f)$ with the **empirical error**

$$\mathcal{E}_{\text{emp}}(f) := \frac{1}{m} \sum_{i=1}^m V(y_i, f(\mathbf{x}_i))$$

Supervised learning model (IV)

If we minimize $\mathcal{E}_{\text{emp}}$ over a sufficiently large set of functions, we can always find a function f with zero empirical error!

But $\mathcal{E}(f)$ may be far away from zero! (**overfitting**)

We introduce a restrictive class of functions $\mathcal{H}$ (**hypothesis space**) and minimize $\mathcal{E}_{\text{emp}}$ within $\mathcal{H}$. That is, our learning algorithm is:

$$f_S = \operatorname{argmin}_{f \in \mathcal{H}} \mathcal{E}_{\text{emp}}(f)$$

Linear regression: $\mathcal{H} = \{f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$

Regularization

We minimize the penalized (regularized) empirical error

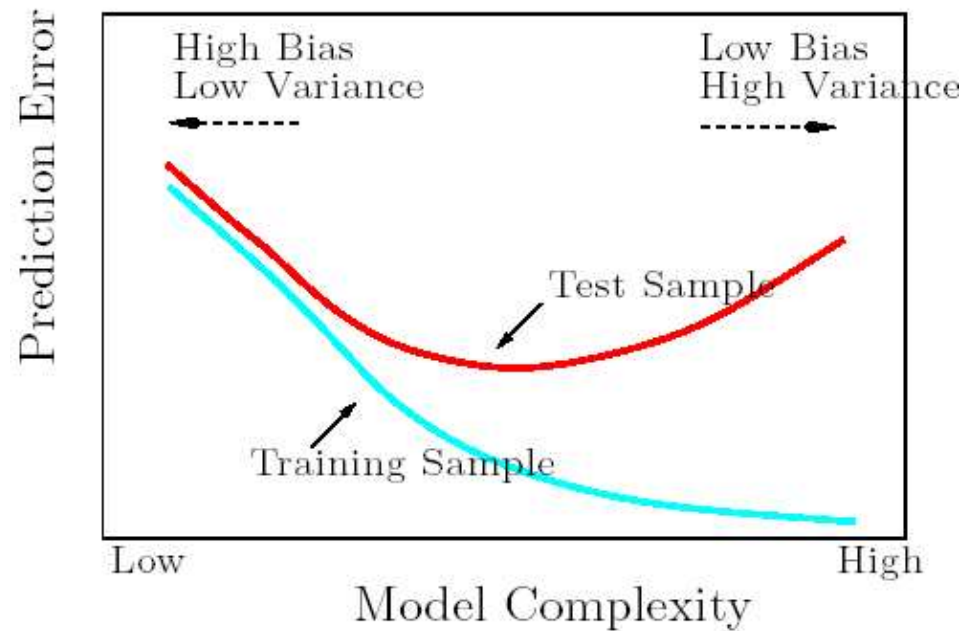
$$E_\lambda(f) := \frac{1}{m} \sum_{i=1}^m V(y_i, f(\mathbf{x}_i)) + \lambda R(f), \quad \lambda > 0$$

Ridge regression: $V(y, z) = (y - z)^2$, $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, $R(f) = \|\mathbf{w}\|^2$

This is similar to empirical error minimization in the hypothesis space $\mathcal{H}_A = \{f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \|\mathbf{w}\| \leq A\}$ for some $A > 0$

(this connection can be made formal: given $A > 0$ there is $\lambda(A) > 0$ such that ridge regression gives the same solution as empirical error minimization in $\mathcal{H}_A$ and vice versa, given λ there is $A(\lambda)$ such that...)

Model complexity and overfitting



- If we pick the best model (learning algorithm) by minimizing the training error we overfit the data
- We wish to estimate the expected (test) error

Model selection and assessment

- **Model selection:** aims at estimating the performance of different models (learning algorithms) in order to choose the (approximately) best one, for example:
 - best hypothesis space among many possible ones
 - best λ in ridge regression
 - best k in k -NN etc.
- **Model assessment:** having chosen a final model, we wish to estimate its expected error (aka generalization or prediction error) on new data

Model selection and assessment (cont.)

If we have a large set of examples, a natural approach is to split the data in three parts: training, validation and test set

1. Use the training set to fit the models (train different learning algorithms on it)
2. Use the validation set for model selection. So the best model is the one which minimizes the validation error
3. Use the test set for assessment of the expected error of the best model above

Typically, we keep most of the data for training (say $1/2$ for training, $1/4$ for validation and testing)

Generalization error bound

- Often, we have only few examples. Can we choose a model and assess its expected error directly (without splitting the training data)?
- **Learning theory** studies conditions which ensure the **predictivity** of a learning algorithm:
 - The expected error is close to the empirical error
 - The expected error decreases when the number of samples increases
- We discuss a central approach in the theory which allows us to **relate the training error to the generalization error**

Hoeffding inequality

Let ξ be a random variable with mean $\mu = \mathbf{E}[\xi]$ and taking values in the interval $[a, b]$

Let $\xi_1, \dots, \xi_m$ be an i.i.d. sample of ξ and define the empirical mean $\bar{\xi}(m) = \frac{1}{m} \sum_{i=1}^m \xi_i$

Then for every $\epsilon > 0$ we have that

$$\text{Prob}(\bar{\xi}(m) - \mu > \epsilon) \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

and

$$\text{Prob}(\mu - \bar{\xi}(m) > \epsilon) \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

Note: for a proof, see §8.2 Devroye, Györfi and Lugosi

Simplest case: $\mathcal{H} = \{f\}$

We shall apply Hoeffding's inequality to the random variable $\xi = V(y, f(\mathbf{x}))$. For simplicity, assume that $V(y, f(\mathbf{x})) \in [0, 1]$ (one can always reduce to this case anyway)

Hoeffding's inequality gives

$$\text{Prob}(\mathcal{E}(f) - \mathcal{E}_{\text{emp}}(f) > \epsilon) \leq \exp(-2m\epsilon^2)$$

This implies that with probability (confidence) at least $1 - \delta$ (think of δ as a small positive number)

$$\mathcal{E}(f) \leq \mathcal{E}_{\text{emp}}(f) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

We call $\mathcal{D}(f) := \mathcal{E}(f) - \mathcal{E}_{\text{emp}}(f)$ the deviation error of f

Finite hypothesis space

Now suppose that $\mathcal{H} = \{f^{(1)}, \dots, f^{(N)}\}$

For each fixed function in $\mathcal{H}$, Hoeffding's inequality holds true

However, now we are not interested in the deviation error of a fixed function but in the deviation error of the **minimizer** of $\mathcal{E}_{\text{emp}}$

$$f_S \equiv f^{(n_*)} := \operatorname{argmin}_{n=1}^N \sum_{i=1}^m V(y_i, f^{(n)}(\mathbf{x}_i))$$

What can we say about the probability of the following event?

$$\{\mathcal{D}(f^{(n_*)}) > \epsilon\}$$

Since $f^{(n_*)}$ can be any function in $\mathcal{H}$ we need a **uniform bound** over $\mathcal{H}$!

Union bound

Recall the following fundamental property of probability: for every set $A_1, A_2, \dots, A_n$ of events, we have that

$$P(A_1 \cup A_2 \cup \dots \cup A_N) \leq \sum_{n=1}^N P(A_n)$$

So, if we let $A_n = \{|\mathcal{E}(f^{(n)}) - \mathcal{E}_{\text{emp}}(f^{(n)})| \geq \epsilon\}$ we conclude that

$$A_1 \cup A_2 \cup \dots \cup A_N = \left\{ \max_{n=1}^N |\mathcal{E}(f^{(n)}) - \mathcal{E}_{\text{emp}}(f^{(n)})| \geq \epsilon \right\}$$

Thus, we have

$$\text{Prob} \left\{ \max_{n=1}^N \left\{ \mathcal{E}(f^{(n)}) - \mathcal{E}_{\text{emp}}(f^{(n)}) \right\} \geq \epsilon \right\} \leq N \exp(-2m\epsilon^2)$$

Uniform bound

$$\text{Prob} \left\{ \max_{n=1}^N \left\{ \mathcal{E}(f^{(n)}) - \mathcal{E}_{\text{emp}}(f^{(n)}) \right\} \geq \epsilon \right\} \leq N \exp(-2m\epsilon^2)$$

This is also called a **uniform** bound over $\mathcal{H}$ (it holds for every $f \in \mathcal{H}$). The bound implies that

$$\text{Prob} \{ \mathcal{E}(f_S) - \mathcal{E}_{\text{emp}}(f_S) \geq \epsilon \} \leq N \exp(-2m\epsilon^2)$$

which is equivalent to say that with confidence at least $1 - \delta$

$$\mathcal{E}(f_S) \leq \mathcal{E}_{\text{emp}}(f_S) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2m}}$$

As N increases more examples are required in order to avoid overfitting!

Sample complexity bound

$$\text{Prob} \{ \mathcal{E}(f_S) - \mathcal{E}_{\text{emp}}(f_S) \geq \epsilon \} \leq |\mathcal{H}| \exp(-2m\epsilon^2)$$

How many examples are needed to avoid overfitting?

Sample complexity: minimum number m of examples that we need in order to ensure that the deviation error of f_S will be less than ϵ with probability at least $1 - \delta$

The sample complexity depends on $\mathcal{H}$, ϵ and δ . In our case:

$$m(\epsilon, \mathcal{H}, \delta) = \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\epsilon^2}$$

Structural risk minimization

This is a model selection approach to choosing a hypothesis space within a nested family of spaces: $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_Q$
(For simplicity assume each set $\mathcal{H}_q$ is finite)

Let $f_{S,q}$ be the minimizer of $\mathcal{E}_{\text{emp}}$ in $\mathcal{H}_q$. For each fixed q we have with confidence at least $1 - \delta$

$$\mathcal{E}(f_{S,q}) \leq \mathcal{E}_{\text{emp}}(f_{S,q}) + \sqrt{\frac{\log |\mathcal{H}_q| + \log \frac{1}{\delta}}{2m}}$$

The **structural risk minimization** chooses the hypothesis space $\mathcal{H}_{q^*}$ which minimizes the r.h.s. of this inequality

Structural risk minimization (cont.)

For each fixed q we have

$$\mathcal{E}(f_{S,q}) \leq \mathcal{E}_{\text{emp}}(f_{S,q}) + \sqrt{\frac{\log |\mathcal{H}_q| + \log \frac{1}{\delta}}{2m}}$$

The **structural risk minimization** chooses the hypothesis space $\mathcal{H}_{q^*}$ which minimizes the r.h.s. of this inequality

The expected error of function f_{S,q^*} (model) is bounded as

$$\mathcal{E}(f_{S,q^*}) \leq \mathcal{E}_{\text{emp}}(f_{S,q^*}) + \sqrt{\frac{\log |\mathcal{H}_{q^*}| + \log Q + \log \frac{1}{\delta}}{2m}}$$

VC–dimension

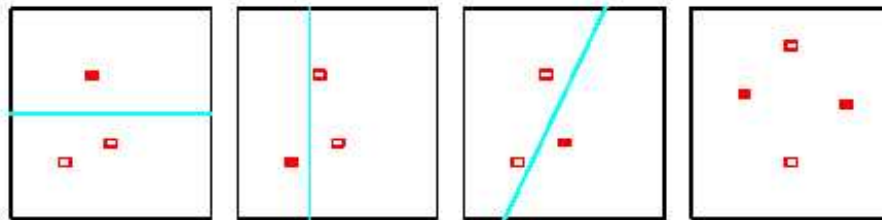
If $\mathcal{H}$ is an infinite set (for example the set of all linear classifiers) things become a bit more complicated but bounds similar to the above one can still be derived which contain a **complexity measure** of $\mathcal{H}$ (aka capacity measure)

For binary classification, the most widely used complexity measure is the **VC–dimension** (from Vapnik and Chervonenkis)

- The VC–dimension of a set of binary classifiers $\mathcal{H}$ is the largest number h of inputs $\mathbf{x}_1, \dots, \mathbf{x}_h$ which can be shattered (classified) in all 2^h possible ways using classifiers in $\mathcal{H}$

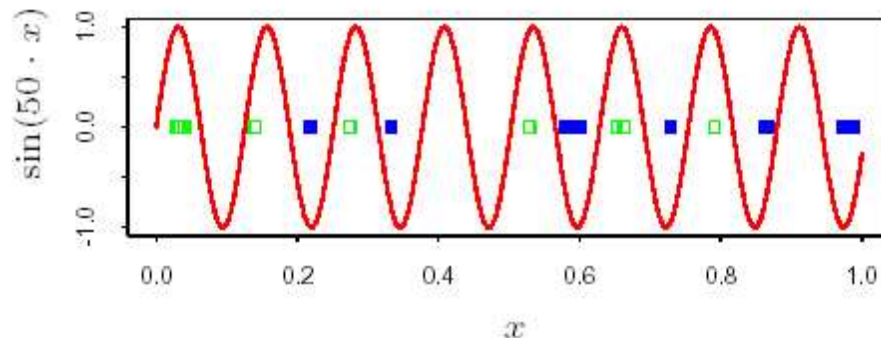
VC–dimension

The VC–dimension of lines in the plane is 3 (in $\mathbb{R}^d$ is $d + 1$)



The VC–dimension of hyper-rectangles in $\mathbb{R}^d$ is $2d$ (exercise)

An example where $h(\mathcal{H})$ is infinite: $\mathcal{H} = \{\text{sign}(\sin ax) : a \in \mathbb{R}\}$



VC–bounds

Theorem: (V. & C.) with confidence at least $1 - \delta$ we have that

$$\mathcal{E}(f_S) \leq \mathcal{E}_{\text{emp}}(f_S) + 2\sqrt{\frac{h(\log \frac{2m}{h} + 1) + \log \frac{2}{\delta}}{m}}$$

Compare to the case that $\mathcal{H}$ is finite, where $h \leq \log N$ ($N := |\mathcal{H}|$)

$$\mathcal{E}(f_S) \leq \mathcal{E}_{\text{emp}}(f_S) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2m}}$$

- It may be that $h \ll \log N \Rightarrow$ VC–bound is better
- When $h = \log N$, VC–bound is slightly worse but essentially the same: $O\left(\frac{h}{m} \log \frac{m}{h}\right)$ vs. $O\left(\frac{h}{m}\right)$

Bias/variance decomposition

$$\mathcal{E}(f_S) - \mathcal{E}(f^*) = \underbrace{\mathcal{E}(f_S) - \mathcal{E}(f_{\mathcal{H}})}_{\text{variance}} + \underbrace{\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f^*)}_{\text{bias}}$$

where $f_{\mathcal{H}} = \operatorname{argmin}_{f \in \mathcal{H}} \mathcal{E}(f)$. Using the VC-bound and the fact that $\mathcal{E}_{\text{emp}}(f_S) \leq \mathcal{E}_{\text{emp}}(f_{\mathcal{H}})$ we obtain that

$$\begin{aligned} \mathcal{E}(f_S) - \mathcal{E}(f_{\mathcal{H}}) &= \mathcal{E}(f_S) - \mathcal{E}_{\text{emp}}(f_S) + \mathcal{E}_{\text{emp}}(f_S) - \mathcal{E}(f_{\mathcal{H}}) \\ &\leq \mathcal{E}(f_S) - \mathcal{E}_{\text{emp}}(f_S) + \mathcal{E}_{\text{emp}}(f_H) - \mathcal{E}(f_{\mathcal{H}}) \\ &\leq 2 \times 2 \sqrt{2 \frac{h(\log \frac{2m}{h} + 1) + \log \frac{2}{\delta}}{m}} \end{aligned}$$

Sample error and approximation error

$$\mathcal{E}(f_S) - \mathcal{E}(f^*) = \underbrace{\mathcal{E}(f_S) - \mathcal{E}(f_{\mathcal{H}})}_{\text{variance}} + \underbrace{\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f^*)}_{\text{bias}}$$

- the variance (or **sample error**) increases with the complexity of $\mathcal{H}$ and decreases with the sample size m
- the bias is independent of the sample and decreases with the complexity of $\mathcal{H}$. It measures the **approximation error** of $\mathcal{H}$ to f^* . For example, for the square loss we have:

$$\mathcal{E}(f) = E_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \mathcal{E}(f^*) + E_{\mathbf{x}}[(f(\mathbf{x}) - f^*(\mathbf{x}))^2]$$

Bias/variance decomposition in statistics

In statistics another way to study the behavior of a learning algorithm is via the average generalization error (over the sampling of the training set S)

We define $\bar{f}(\mathbf{x}) = \mathbf{E}_S[f_S(\mathbf{x})]$. For the square loss we have that

$$\mathbf{E}_S[\mathcal{E}(f)] = \mathcal{E}(f^*) + \underbrace{\mathbf{E}_S[\|\bar{f} - f_S\|^2]}_{\text{variance}} + \underbrace{\|f^* - \bar{f}\|^2}_{\text{bias}}$$

where we have used the notation

$$\|f - g\|^2 := \mathbf{E}_{\mathbf{x}}[(f(\mathbf{x}) - g(\mathbf{x}))^2]$$

When $\bar{f} = f_{\mathcal{H}}$ the bias is the same as the previous notion of bias (approximation error)