VI - Statistical Learning Principles

Math 728 D - Machine Learning & Data Science - Spring 2019

Contents

1. Introduction - Basic Notions
   - “Loss” Quantification
2. Risk Notions
   - Basic Ideas
   - Basic Definitions and Properties
3. General Estimation Strategies - An Orientation
   - Variance-Bias Decomposition
4. A Closer Look at ERM for Classification
   - A Preliminary Result
   - A Better Result - VC-Dimension

Bias and Variance Balance

Many statistical learning algorithms search for estimators or decision functions via the following optimization problem

$$\text{Minimize} \quad \left\{\text{Loss} + \lambda\text{Penalty}\right\} \tag{2.1}$$

- Loss measures the error of fitting the data - bias or approximation error
- Penalty the complexity of the learned function to avoid overfitting
- λ is a regularization parameter that is to balance both effects.

Example: soft-margin SVM: the problem

$$\text{minimize}_{\mathbf{w},b,\xi} : \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{j=1}^{N}\xi_j \quad \text{subject to} \quad y_i\underbrace{(\mathbf{w}\cdot\mathbf{x}^i + b)}_{=f(\mathbf{x}^i)} \geq 1 - \xi_i, \ \ i \leq N, \tag{2.2}$$

can be shown to provide the same solution as

$$\text{minimize}_{\mathbf{w},b} : \quad \underbrace{\sum_{i=1}^{N}\left[1 - y_i(\mathbf{w}\cdot\mathbf{x}^i + b)\right]_+}_{\text{“hinge loss”}} + \underbrace{\lambda\|\mathbf{w}\|_2^2}_{\text{penalty}} \tag{2.3}$$

when $\lambda = \frac{1}{2C}$ (where $[x]_+ := \max\{x, 0\}$).

Explanation

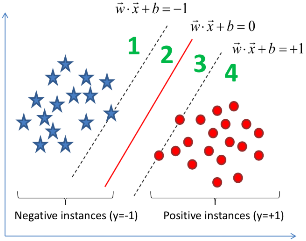
Notice: i-th loss term $[1 - y_i f(\mathbf{x}^i)]_+ = [1 - y_i(\mathbf{w}\cdot\mathbf{x}^i + b)]_+ \neq 0$ if and only if

$$\begin{aligned} &(\mathbf{w}\cdot\mathbf{x}^i + b) < 1 && \text{when} \quad y_i = +1, && \rightsquigarrow \quad 0 < \xi_i := 1 - (\mathbf{w}\cdot\mathbf{x}^i + b) \\ &(\mathbf{w}\cdot\mathbf{x}^i + b) > -1 && \text{when} \quad y_i = -1, && \rightsquigarrow \quad 0 < \xi_i := 1 + (\mathbf{w}\cdot\mathbf{x}^i + b) \end{aligned} \tag{2.4}$$

and

$$(2.3) \quad \Leftrightarrow \quad \text{minimize}_{\mathbf{w},b} : \quad C\sum_{i=1}^{N} \underbrace{\left[1 - y_i f(\mathbf{x}^i)\right]_+}_{\xi_i} + \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \quad y_i f(\mathbf{x}^i) \geq 1 - \xi_i,\ i \leq N.$$

- If the instance is negative, it is penalized only in regions 2,3,4
- If the instance is positive, it is penalized only in regions 1,2,3


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
1
2
3
4
Negative instances (y=-1)
Positive instances (y=+1)

How much faith can one put on the results of learning algorithms?

Basic principle:

- Understand first what is ideally best possible - a ground truth benchmark - which depends entirely on the underlying probabilistic model. It realizes the smallest possible generalization error under the given probability law.

  The generalization error is often referred to as Risk. The minimal risk, so to speak the intrinsic unavoidable error, is, however not known.
- Goal: design algorithms whose risk comes close to the minimal risk.
- Any concrete algorithm will exceed this Risk $\rightsquigarrow$ Excess Risk. This Excess Risk measures the quality of the algorithm. So this is what we would like to estimate.
- How to actually measure this? because Risk and Excess Risk are unknown.
- Idea: formulate a suitable Empirical Risk. A typical learning algorithm minimizes this Empirical Risk $\rightsquigarrow$ Empirical Risk Minimization (ERM).
- Accuracy assessment relies then on comparing the Empirical Risk and the Risk.
- This comparison requires probabilistic estimates - concentration inequalities (see Lecture III) $\rightsquigarrow$ “error estimates” are understood in a probabilistic sense!

Risk Notions - Regression

Always: $\mathfrak{Z}_N = \{(\mathbf{x}^i, y_i), \dots, (\mathbf{x}^N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$ i.i.d. samples from an unknown probability measure P with density p

Regression: $\mathcal{Y} \subset \mathbb{R}^n$ has a continuous range, objective: estimate the regression function $f_p(x) := \mathbb{E}[y|x]$ (see Lecture II, (8.3)).

What is the benchmark - an appropriate Risk notion in this case?

Risk functional: For any $f \in L_2(\mathcal{X}, p_{\mathcal{X}})$, let

$$\mathcal{R}[f] := \int_{\mathcal{Z}} (y - f(x))^2 dP(x, y). \tag{3.1}$$

Factorizing $dP(x, y) = dP(y|x)dP_{\mathcal{X}}(x)$, and the fact that $f^* = f_p := \mathbb{E}[y|x]$ is an orthogonal projection yield

$$\mathcal{R}[f] = \mathcal{R}[f_p] + \|f - f_p\|^2_{L_2(\mathcal{X}, p_{\mathcal{X}})}. \tag{3.2}$$

Thus the risk $\mathcal{R}$ is minimized by the regression function f_p and its minimum is $\mathcal{R}[f_p]$.

The Excess Risk incurred by any concrete estimator $\hat{f}_{\mathfrak{Y}}$ is $\|\hat{f}_{\mathfrak{Z}} - f_p\|^2_{L_2(\mathcal{X}, p_{\mathcal{X}})}$.

Empirical Risk:

$$\widehat{\mathcal{R}}_{\mathfrak{Z}_N}[f] := \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2. \tag{3.3}$$

This is a Random variable over $\mathcal{Z}^N$ (mean square risk).

Risk Notions - Classification

Classification: $\mathcal{Y} = \{+1, -1\}$ (or more generally, has finiteley many elements - classes)

- We have seen in the previous lecture, that constructing a classifier or decision function can be seen as finding a set such that any element in the set gets a positive label $\rightsquigarrow$.
- General Classifier format: every (binary) classifier can be recast as:

$$h_S(x) = \chi_S(x) - \chi_{S^c}(x) = \begin{cases} +1 & \text{if} \quad x \in S, \\ -1 & \text{if} \quad x \notin S, \end{cases} \tag{3.4}$$

  which is well defined for any subset $S \subset \mathcal{X}$ that is $p_{\mathcal{X}}$-measurable.
- The risk of any h_S, defined by (3.4), is given by

$$\mathcal{R}[h_S] = \int_{\mathcal{X}} \mathrm{Prob}\{h_S(x) \neq y\} dp_{\mathcal{X}}, \tag{3.5}$$

  which is the measure of the set of misclassified events.
- Is there an “ideal set S^*” - classifier - that minimizes the risk over all admissible measurable sets? Such a classifier is called Bayes Classifier: which depends only on the underlying (unknow) probability distribution (that serves as a data model) $\rightsquigarrow$

Bayes Classifier

Some usefull ingredients:

- $p(x, y) = p_{\mathcal{X}}(x) \cdot p(y|x)$ a probability density on $\mathcal{Z}$ (see Lecture II, page 18)
- Let
$$\rho(x) := \mathrm{Prob}\{y = 1|x\}$$
denote the probability of a feature point x to have the label $+1$. The expected label is the regression function
$$\eta(x) := \mathbb{E}[y|x] = \rho(x) - (1 - \rho(x)) = 2\rho(x) - 1 \quad \text{(verify this)} \tag{3.6}$$

Define (see (3.4))

$$h^*(x) = h_{S^*}(x) = \chi_{S^*}(x) - \chi_{(S^*)^c}(x) \quad \text{where} \quad S^* := \{x \ : \ \eta(x) \geq 0\}. \tag{3.7}$$

Remark 1

The classifier (3.7) *minimizes the risk* $\mathcal{R}[h_S]$, *i.e.,*

$$S^* = \operatorname*{argmin}_{S} \mathcal{R}[h_S],$$

any any set S' *that also mimizes the risk can differ from* S^* *only by a set of either* $p_{\mathcal{X}}$*-measure zero or where* η *vanishes.*

Risk Notions - Classification

The following properties explain the above definition.

Lemma 2

The Baye's Risk is

$$\mathcal{R}[h^*] = \mathcal{R}[h_{S^*}] = \int_{\mathcal{X}} \min\{\rho, 1-\rho\} dp_{\mathcal{X}}. \tag{3.8}$$

For $C_p := \int_{\mathcal{X}} \rho dp_{\mathcal{X}}$ *one has*

$$\mathcal{R}[h_S] = C_p - \int_S \eta dp_{\mathcal{X}}. \tag{3.9}$$

Moreover, the Excess Risk is given by

$$\mathcal{R}[h_S] - \mathcal{R}[h_{S^*}] = \eta_{S^*} - \eta_S = \int_{S \Delta S^*} |\eta| dp_{\mathcal{X}}, \tag{3.10}$$

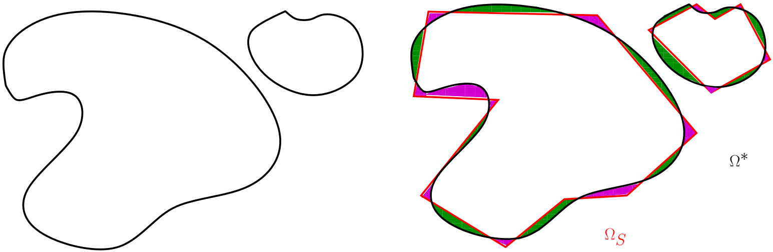
where $\eta_S := \int_S \eta dp_{\mathcal{X}}$ *and* $A \triangle B := (A \setminus B) \cup (B \setminus A)$ *is the symmetric difference of A and B.*

Remark 3

By (3.9), *maximizing* η_S, *which* S^* *does, minimizes the risk.*

Excess Risk - Classification

$$\mathcal{R}(S) - \mathcal{R}(S^*) = \int_{S \Delta S^*} |\eta| dp_{\mathcal{X}} = \eta_{S^*} - \eta_S$$



Empirical Risk: A natural discrete analog to (3.5) is:

$$\widehat{\mathcal{R}}[h_S] := \frac{1}{N} \#\{i : h_S(x_i) \neq y_i\}. \tag{3.11}$$

Proof of Lemma 2: As for (3.8), since $x \in S^* \Leftrightarrow \eta(x) > 0 \Leftrightarrow \rho(x) > 1/2$ we have

$$\begin{aligned} x \in S^* \quad &\Rightarrow \quad \mathrm{Prob}\{y = -1|x\} = 1 - \rho(x) < \tfrac{1}{2} \leq \rho(x), \\ x \notin S^* \quad &\Rightarrow \quad \mathrm{Prob}\{y = 1|x\} = \rho(x) \leq \tfrac{1}{2} \leq 1 - \rho(x), \end{aligned}$$

which yields (3.8).
To show (3.9) $h_S(x) \neq y$ iff either $y = 1$ and $x \in S^c$ or $y = -1$ and $x \in S$. The probability of the first case is $\int\limits_S (1-\rho)dp_{\mathcal{X}}$, and of the second case is $\int\limits_{S^c} \rho dp_{\mathcal{X}}$ i.e.,

$$\int\limits_{\mathcal{X}} \mathrm{Prob}\{h_S(x) \neq y\} dp_{\mathcal{X}} = \int\limits_S (1-\rho)dp_{\mathcal{X}} + \int\limits_{S^c} \rho dp_{\mathcal{X}} \tag{3.12}$$

The right hand side can be rewritten as

$$\int\limits_S (1-\rho)dp_{\mathcal{X}} + \int\limits_X \rho dp_{\mathcal{X}} - \int\limits_S \rho dp_{\mathcal{X}} = C_p + \int\limits_S (1-2\rho)dp_{\mathcal{X}} \overset{(3.6)}{=} C_p - \int\limits_S \eta dp_{\mathcal{X}},$$

which is (3.9).
Concerning (3.10), the first identity follows directly from (3.9). Regarding the second identity, we can rwite

$$\eta_{S^*} - \eta_S = \int\limits_{S^*} \eta dp_{\mathcal{X}} - \int\limits_S \eta dp_{\mathcal{X}} = \int\limits_{S^* \setminus S} \eta dp_{\mathcal{X}} - \int\limits_{S \setminus S^*} \eta dp_{\mathcal{X}} = \int\limits_{S^* \setminus S} |\eta| dp_{\mathcal{X}} + \int\limits_{S \setminus S^*} |\eta| dp_{\mathcal{X}},$$

as claimed. □

Risk Notions for Further Learning Tasks

Pattern Recognition:

Median estimation ⇝ L_1-metric

$$\mathcal{R}[f] := \int_{\mathcal{Z}} \frac{1}{2}|f(x) - y| dP(x,y), \qquad \widehat{\mathcal{R}}_{\mathfrak{Z}_N}[f] := \frac{1}{N}\sum_{i=1}^{N} |f(x_i) - y_i|$$

Density Estimation:

compare with Maximum Likelihood Estimation - log-likelihood function:

$$\mathcal{R}[p] := \int_{\mathcal{X}} (-\log p(x)) dP(x), \qquad \widehat{\mathcal{R}}_{\mathfrak{Z}_N}[p] := -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i)$$

A General Road Map

Given a learning task (classification or regression):

- Understand/define a suitable risk notion $\mathcal{R}$ and the corresponding minimal (Bayes) risk $\mathcal{R}[f^*]$
- Devise a learning algorithm: $\mathfrak{Z}_N \subset \mathcal{Z} \to \hat{f}_{\mathfrak{Z}_N}$, that hopefully minimizes the excess risk

$$\mathcal{E}[\hat{f}] := \mathcal{R}[\hat{f}] - \mathcal{R}[f^*]. \tag{4.1}$$

- Any such algorithm constructs (a regression function estimator or classifier) $\hat{f}$ from a hypothesis class $\mathcal{H}$ (polynomials, splines, kernels, networks, sets,...) that can be used to represent the estimator. The complexity of $\mathcal{H}$ determines how well the optimal f^* can be at best approximated by elements from $\mathcal{H} \rightsquigarrow$

$$f_{\mathcal{H}} \in \operatorname*{argmin}_{g\in\mathcal{H}} \mathcal{R}[g] : \quad \mathcal{R}[f_{\mathcal{H}}] - \mathcal{R}[f^*] =: a_{\mathcal{H}}(f^*, \mathcal{H}) \tag{4.2}$$

  is the best approximation error or bias (a deterministic quantity).
- According to the underlying statistical model, any outcome $\hat{f}$ of an algorithm is based on i.i.d. samples and hence itself a random variable over $\mathcal{Z}^N$. The only chance to estimate the quantity $\mathcal{E}[\hat{f}]$ is to exploit this statistical model.
- As error appraisal one can therefore ask for

$$\mathbb{E}\big[\mathcal{E}[\hat{f}_{\mathfrak{Z}_N}]\big] \quad \text{or} \quad \mathrm{Prob}\{\mathcal{E}[\hat{f}_{\mathfrak{Z}_N}] \geq \epsilon_N\}, \quad N \to \infty. \tag{4.3}$$

Error Decomposition - Road Map

- Thus, for any estimator $\hat{f}$, the excess risk $\mathcal{E}[\hat{f}]$ can be decomposed as

$$\mathcal{R}[\hat{f}] - \mathcal{R}[f^*] = \underbrace{\mathcal{R}[\hat{f}] - \mathcal{R}[f_{\mathcal{H}}]}_{\mathcal{E}_{var}} + \underbrace{\mathcal{R}[f_{\mathcal{H}}] - \mathcal{R}[f^*]}_{a(f^*,\mathcal{H})} \quad (4.4)$$

  where both differences are, by definition nonnegative.
- Once the hypothesis class $\mathcal{H}$ is chosen, the second term $a(f^*,\mathcal{H})$ is deterministic and independent of any learning algorithm. This term is the smaller the richer $\mathcal{H}$ is.
- Both $\hat{f}$ and $f_{\mathcal{H}}$ belong to $\mathcal{H}$. When $\mathcal{H}$ is very rich, one faces the risk of overfitting, i.e., the variance of the random variable $\mathcal{E}_{\rm var}$ is expected to increase, when $\mathcal{H}$ gets richer. As mentioned earlier, a key objective is therefore to find a proper balance between bias and variance.
- To invoke the underlying probabilistic model, involve the empirical risk $\widehat{\mathcal{R}}$ and further decompose $\mathcal{E}_{var}$ as follows

$$\mathcal{E}_{var} = \mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}] + \widehat{\mathcal{R}}[\hat{f}] - \widehat{\mathcal{R}}[f_{\mathcal{H}}] + \widehat{\mathcal{R}}[f_{\mathcal{H}}] - \mathcal{R}[f_{\mathcal{H}}]. \quad (4.5)$$

- If the algorithm is based on minimizing the empirical risk, the second difference is non-positive and we obtain

$$\mathcal{E}_{var} \le \left|\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}] + \widehat{\mathcal{R}}[f_{\mathcal{H}}] - \mathcal{R}[f_{\mathcal{H}}]\right| \le \left|\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}]\right| + \left|\widehat{\mathcal{R}}[f_{\mathcal{H}}] - \mathcal{R}[f_{\mathcal{H}}]\right|. \quad (4.6)$$

  This will be done by invoking concentration inequalities.

Estimating $\mathcal{E}_{var}$

Assumption: The estimator $\hat{f}_{\mathfrak{Z}_N}$ is obtained through Empirical Risk Minimiation (ERM)

$$\hat{f}_{\mathfrak{Z}_N} = \operatorname*{argmin}_{g \in \mathcal{H}} \widehat{\mathcal{R}}[g] \tag{4.7}$$

The bound (4.6) $\mathcal{E}_{var} \leq \left|\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}]\right| + \left|\widehat{\mathcal{R}}[f_{\mathcal{H}}] - \mathcal{R}[f_{\mathcal{H}}]\right|$ involves quantities of the form

$$\left|\mathcal{R}[g] - \widehat{\mathcal{R}}[g]\right| \qquad g \in \mathcal{H}. \tag{4.8}$$

Recall: Regression (5.32), (3.3) $\rightsquigarrow$

$$\mathcal{R}[f] := \int_{\mathcal{Z}} (y - f(x))^2 dP(x,y) = \mathbb{E}[\xi], \qquad \widehat{\mathcal{R}}_{\mathfrak{Z}_N}[f] := \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\xi_i. \tag{4.9}$$

So we need to estimate

$$\left|\frac{1}{N}\sum_{i=1}^{N}\xi_i - \mathbb{E}[\xi]\right|, \quad \xi_i \text{ i.i.d. according to } P \tag{4.10}$$

Estimating $\mathcal{E}_{var}$

Classification: recall from Lemma 2 and (3.11)

$$\mathcal{R}[h_S] = \int_{\mathcal{X}} \mathrm{Prob}\{h_S(x) \neq y dp_{\mathcal{X}} = C_p - \int_S \eta dp_{\mathcal{X}}, \quad \widehat{\mathcal{R}}[h_S] = \frac{1}{N}\#\{i : h_S(x_i) \neq y_i\} \tag{4.11}$$

Consider the empirical quantities

$$p_S := \int_S dp_{\mathcal{X}} \leftrightarrow \hat{p}_S := \frac{1}{N}\sum_{i=1}^{N} \chi_S(x_i) \quad \eta_S := \int_S \eta dp_{\mathcal{X}} \leftrightarrow \hat{\eta}_S := \frac{1}{N}\sum_{i=1}^{N} y_i\chi_S(x_i). \tag{4.12}$$

One can show:

$$\operatorname*{argmin}_{S\in\mathcal{H}} \widehat{\mathcal{R}}[h_S] = \operatorname*{argmin}_{S\in\mathcal{H}} \left(C_p - \hat{\eta}_S\right) = \operatorname*{argmax}_{S\in\mathcal{H}} \hat{\eta}_S \quad \rightsquigarrow \tag{4.13}$$

$$\left|\widehat{\mathcal{R}}[h_S] - \mathcal{R}[h_S]\right| = \left|\eta_S - \hat{\eta}_S\right| \tag{4.14}$$

Since $p_S = \mathbb{E}[\chi_S(X)] = \mathbb{E}_{\mathcal{Z}^N}[\hat{p}_S]$ and $\eta_S\mathbb{E}[Y\chi_S(X)] = \mathbb{E}[\hat{\eta}_S]$

$$\text{(4.14) requires again estimating} \qquad \left|\frac{1}{N}\sum_{i=1}^{N}\xi_i - \mathbb{E}[\xi]\right| \tag{4.15}$$

Estimating $\mathcal{E}_{var}$

Summary:

In all these cases estimating the quantities (4.8) $\Big|\mathcal{R}[g]-\widehat{\mathcal{R}}[g]\Big|$ for $g\in\mathcal{H}$, appearing in the bound (4.6) for $\mathcal{E}_{var}$, amounts to estimating

$$\sup_{\mathfrak{Z}_N\subset\mathcal{Z}}\left|\frac{1}{N}\sum_{i=1}^{N}\xi_i-\mathbb{E}[\xi]\right| \quad \text{for i.i.d } \xi_i$$

This is precisely the subject of the concentration inequalities discussed in Lecture III.

Back to (4.6): $\mathcal{E}_{var}\le\big|\mathcal{R}[\hat f]-\widehat{\mathcal{R}}[\hat f]\big|+\big|\widehat{\mathcal{R}}[f_{\mathcal{H}}]-\mathcal{R}[f_{\mathcal{H}}]\big|$

Goal: show that both terms (being random variables) are small in expectation of with high probability.

- The second term $\big|\widehat{\mathcal{R}}[f_{\mathcal{H}}]-\mathcal{R}[f_{\mathcal{H}}]\big|$ is relatively harmless, because $f_{\mathcal{H}}$ is a fixed object in the hypothesis class $\mathcal{H}$. So as soon as we know something about the variance or the range of the underlying density, one can apply Hoeffding's or Bernstein's inequality (see Lecture III, Theorem 2, (3.5), Corollary 5, (3.14)) to conclude that $\mathbb{R}\big\{\big|\widehat{\mathcal{R}}[f_{\mathcal{H}}]-\mathcal{R}[f_{\mathcal{H}}]\big|>\epsilon\big\}\le 2e^{-cN\epsilon^2}$ (some $c>0$).
- The first term is more critical because we need an estimate that is valid for all $\hat f\in\mathcal{H}$. This is where the complexity of the hypothesis class $\mathcal{H}$ enters - the more complex $\mathcal{H}$ is, the harder it gets to have simultaneous estimates.

ERM in Classification - Summary

Scenario:

- $\mathfrak{Z}_N = \{(x_i, y_i), \ldots, (x_N, y_N)\} \subset \mathcal{X} \times \{+1, -1\}$, i.i.d. according to $P(x, y)$;
- $\eta(x) = \mathbb{E}[y|x] = 2\rho(x) - 1$, $\rho(x) = \mathrm{Prob}\{y = 1|x\}$;
- $p_S := \int\limits_S dp_{\mathcal{X}}, \quad \eta_S := \int\limits_S \eta dp_{\mathcal{X}},$
- $\hat{p}_S := \frac{1}{N}\sum_{i=1}^N \chi_S(x_i), \quad \hat{\eta}_S := \frac{1}{N}\sum_{i=1}^N y_i\chi_S(x_i) \Rightarrow p_S = \mathbb{E}[\hat{p}_S], \ \eta_S = \mathbb{E}[\hat{\eta}_S]$
- $\mathcal{R}[h_S] = C_p - \eta_S, \quad \widehat{\mathcal{R}}[h_S] = C_p - \hat{\eta}_S.$
- Recall from (4.6) that we need to bound for all $S \in \mathcal{H}$ ($\hat{f} \leftrightarrow h_S$)

$$\begin{aligned} \mathcal{E}_{var} &\leq \left|\mathcal{R}[h_S] - \mathcal{R}[h_{S_\mathcal{H}}] - (\widehat{\mathcal{R}}[h_S] - \widehat{\mathcal{R}}[h_{S_\mathcal{H}}])\right| \\ &= \left|\eta_{S_\mathcal{H}} - \eta_S - (\hat{\eta}_{S_\mathcal{H}} - \hat{\eta}_S)\right| && (5.1) \\ &\leq \left|\hat{\eta}_S - \eta_S\right| + \left|\eta_{S_\mathcal{H}} - \hat{\eta}_{S_\mathcal{H}}\right| && (5.2) \end{aligned}$$

where $S_\mathcal{H} := \mathrm{argmin}_{S\in\mathcal{H}}\,\mathcal{R}[h_S]$.

So we want to bound:

$$\mathrm{Prob}\left\{\left|\eta_{S_\mathcal{H}} - \eta_S - (\hat{\eta}_{S_\mathcal{H}} - \hat{\eta}_S)\right| \geq \epsilon\right\} \quad \text{or} \quad \mathrm{Prob}\left\{\left|\hat{\eta}_S - \eta_S\right| \geq \epsilon\right\}, \quad S \in \mathcal{H}.$$

A First Result: $\#\mathcal{H} < \infty$

Proposition 4

Assume that in the above scenario

$$\#\mathcal{H} = H < \infty. \tag{5.3}$$

For any fixed $r \in \mathbb{N}$ *define*

$$e_N(S) := \sqrt{p_{S\Delta S_{\mathcal{H}}}\epsilon_N} + \epsilon_N, \quad \epsilon_N = \epsilon_N(\mathcal{H}) := \frac{10(r\log N + \log \#\mathcal{H})}{3N}. \tag{5.4}$$

Then

$$\mathrm{Prob}_{\mathcal{Z}^N}\{\forall\, S \in \mathcal{H},\ \textit{one has}\ \left|\eta_{S_{\mathcal{H}}} - \eta_S - (\hat\eta_{S_{\mathcal{H}}} - \hat\eta_S)\right| \le e_N(S)\} \ge 1 - 2N^{-r}. \tag{5.5}$$

Thus, in view of (4.4)*, for* $\hat S := \mathrm{argmin}_{S\in\mathcal{H}}\widehat{\mathcal{R}}[h_S]$*, the excess risk bound*

$$\mathcal{R}[h_{\hat S}] - \mathcal{R}[h_{S^*}] \le e_N(\hat S) + a(h_{S^*}, \mathcal{H}) \tag{5.6}$$

holds with probability at least $1 - 2N^{-r}$*.*

Recall: $a(h_{S^*}, \mathcal{H})$ is a best approximation error depending on the regression function $\eta(x) = \mathbb{E}[y|x]$. and $\mathcal{H}$. How to estimate such terms, later.

Some Ingredients

Understand first who the empirical quantities $\hat{p}_S, \hat{\eta}_S$ approximate p_S, η_S, respectively.

Lemma 5

For $p_S, \hat{p}_S, \eta_S, \hat{\eta}_S$ defined in Scenario, one has

$$\mathrm{Prob}\{|p_S - \hat{p}_S| > \delta\} \leq 2e^{-\frac{N\delta^2}{2p_S + 2\delta/3}}, \tag{5.7}$$

and

$$\mathrm{Prob}\{|\eta_S - \hat{\eta}_S| > \delta\} \leq 2e^{-\frac{N\delta^2}{2p_S + 2\delta/3}}. \tag{5.8}$$

Proof: We already know $\mathbb{E}[\hat{p}_S] = p_S$, $\mathbb{E}[\hat{\eta}_S] = \eta_S$. We wish to apply Bernstein's Inequality, Lecture III, Corollary 5. We need a bound M for the random variables $\chi_S(x), y\chi_S(x)$ (see Scenario) which is clearly one. Moreover, for each $S \in \mathcal{H}$

$$\begin{aligned} \mathrm{var}[\chi_S] &= \mathbb{E}\big[(\chi_S - p_S)^2\big] = \int_S \chi_S(x)^2 - 2\chi_S(x)p_S + p_S^2 dp_{\mathcal{X}} \\ &= \int_S (1 - 2p_S^2 + p_S^2) dp_{\mathcal{X}} = p_S(1 - p_S^2) \leq p_S. \end{aligned}$$

Proof of Lemma 5 continued: Similarly

$$\begin{aligned}\mathrm{var}[y\mathcal{X}(x)] &= \mathbb{E}\big[(y\mathcal{X}_S(x)-\eta_S)^2\big] = \int_S y^2\mathcal{X}_S(x)^2 - 2y\mathcal{X}_S(x)\eta_S + \eta_S^2 dp_{\mathcal{X}}\\ &= p_S - 2\eta_S^2 + \eta_S^2 p_S = p_S + \eta_S^2(p_S-2) \leq p_S \leq 1,\end{aligned}$$

Inserting this into Bernstein's inequality Lecture III, (3.14)

$$\mathrm{Prob}\Big(\Big|\frac{1}{N}\Big(\sum_{j=1}^N X_j\Big) - \mu\Big| > \epsilon\Big) \leq 2\exp\Big\{-\frac{N\epsilon^2}{2(\sigma^2+\epsilon M/3)}\Big\}$$

with $M = 1$, $\sigma^2 \leq p_S$ yields (5.7), (5.8). □

We could apply (5.8) to both summands in (5.2) to estimate the probability in (5.5). A somewhat better estimated would be obtained through estimates for (5.1), i.e., for $\big|\eta_{S_{\mathcal{H}}} - \eta_S - (\hat{\eta}_{S_{\mathcal{H}}} - \hat{\eta}_S)\big|$ in (5.5). This can be obtained in exactly the same way as above, as shown next.

Some Ingredients

Recall $S_{\mathcal{H}}$ is the best approximation from the hypothesis class $\mathcal{H}$ to the Bayes set S^*. Consider the random variable

$$Y\chi_S(X) - Y\chi_{S_{\mathcal{H}}}(X) = Y(\chi_{S\setminus S_{\mathcal{H}}} - \chi_{S_{\mathcal{H}}\setminus S}),$$

for which we have again $|Y\chi_S(X) - Y\chi_{S_{\mathcal{H}}}(X)| \leq 1$ and as above one can show

$$\mathrm{var}\big[Y\chi_S(X) - Y\chi_{S_{\mathcal{H}}}(X)\big] \leq p_{S\Delta S_{\mathcal{H}}}. \tag{5.9}$$

Since

$$\mathbb{E}\big[Y(\chi_S(X) - \chi_{S_{\mathcal{H}}}(X))\big] = \mathbb{E}\big[\hat{\eta}_S - \hat{\eta}_{S_{\mathcal{H}}}\big] = \eta_S - \eta_{S_{\mathcal{H}}}, \tag{5.10}$$

the same arguments as above yield:

Lemma 6

For $p_S, \hat{p}_S, \eta_S, \hat{\eta}_S$ defined in Scenario, one has (see (5.1))

$$\mathrm{Prob}_{\mathcal{Z}^N}\big\{|\eta_S - \eta_{S_{\mathcal{H}}} - (\hat{\eta}_S - \hat{\eta}_{S_{\mathcal{H}}})| > \delta\big\} \leq 2e^{-\frac{N\delta^2}{2p_{S\Delta S_{\mathcal{H}}} + 2\delta/3}}. \tag{5.11}$$

Some Ingredients

Lemma 7

For any finite $\mathcal{H}$, $\#\mathcal{H} = H < \infty$ and $\epsilon_N = \epsilon_N(\mathcal{H})$ defined by (5.4)*, we have*

$$\mathrm{Prob}_{\mathcal{Z}^N}\left\{\forall\, S \in \mathcal{H} : |\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})| \le \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\right\} \ge 1 - 2N^{-r}. \tag{5.12}$$

Proof: Substituting $\delta = \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N$ in (5.11), yields for any fixed S

$$\mathrm{Prob}\left\{|\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})| > \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\right\} \le 2\exp\left\{-\frac{N(\sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N)^2}{2p_{S\Delta S_\mathcal{H}} + 2(\sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N)/3}\right\}.$$

$\underline{\text{Case 1.} \quad \epsilon_N \le p_{S\Delta S_\mathcal{H}}:}$ $\rightsquigarrow$ numerator in exponential $\ge N p_{S\Delta S_\mathcal{H}} \epsilon_N$ and denominator $\le \frac{10 p_{S\Delta S_\mathcal{H}}}{3}$. Thus, exponent $\ge \frac{N p_{S\Delta S_\mathcal{H}} \epsilon_N}{10\epsilon_N/3} = \frac{3N p_{S\Delta S_\mathcal{H}}}{10} \ge \frac{3N\epsilon_S}{10}$.

$\underline{\text{Case 2.} \quad \epsilon_N > p_S:}$ $\rightsquigarrow$ numerator in exponential $\ge N\epsilon_N^2$ and denominator $\le \frac{10\epsilon_N}{3}$. Thus, exponent $\ge \frac{N\epsilon_N^2}{10\epsilon_N/3} = \frac{3N\epsilon_N}{10}$.

Proof of Lemma 7 continued: Therefore, in both cases we obtain for a fixed $S \in \mathcal{H}$

$$\mathrm{Prob}\{|\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})| > \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\} \leq 2\exp\left\{-\frac{3N\epsilon_N}{10}\right\}. \tag{5.13}$$

Now recall the definition of

$$\epsilon_N = \epsilon_N(\mathcal{H}) = \frac{10(r\log N + \log \#\mathcal{H})}{3N} \quad \Rightarrow \quad 2\exp\left\{-\frac{3N\epsilon_N}{10}\right\} = 2N^{-r}(\#\mathcal{H})^{-1} \tag{5.14}$$

which says, in view of (5.13), that for each $S \in \mathcal{H}$

$$\mathrm{Prob}\{|\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})| > \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\} \leq 2N^{-r}(\#\mathcal{H})^{-1}.$$

By the Union Bound (Lecture IV, Remark 4) it follows that

$$\begin{aligned} &\mathrm{Prob}\Big\{\text{for every } S \in \mathcal{H} : |\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})| \leq \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\Big\} \\ &\qquad \geq 1 - \sum_{S\in\mathcal{H}} 2N^{-r}(\#\mathcal{H})^{-1} = 1 - 2N^{-r}, \end{aligned} \tag{5.15}$$

which is (5.12). □

The **Proof of Proposition 4** follows now from Lemma 7 and the decomposition (4.4). □

A Major Deficit of Proposition 4

- The assumption $\#\mathcal{H} = H < \infty$ is completely unrealistic. Even for simple linear SVMs

  $$\mathcal{H} = \{S : S \text{ is a half-space } = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} + b \geq 0\} \text{ for some } b \in \mathbb{R},\ \mathbf{w} \in \mathbb{R}^d\}$$

  i.e., $\#\mathcal{H} = \infty$.
- Since $\#\mathcal{H}$ enters through

  $$e_N(S) := \sqrt{p_{S \Delta S_{\mathcal{H}}} \epsilon_N} + \epsilon_N, \quad \epsilon_N = \epsilon_N(\mathcal{H}) := \frac{10(r \log N + \log \#\mathcal{H})}{3N}.$$

  the definition of the threshhold $\epsilon_N(\mathcal{H})$ in (5.4), as well as the Union Bound argument in (5.15), a new ingredient is required.
- Key Observation: any given finite set of data $\mathfrak{Z}_N$ cannot "see" infinitely many different elements in $\mathcal{H}$.
- What should therefore matter is the capacity of the hypothesis class $\mathcal{H}$ to separate subsets of given data.
- This capacity is quantified by the so called Vapnik-Chervonenkis--dimension (VC-dimension) of $\mathcal{H}$, see [5, 4].

Shattering

Given a collection of sets $\mathcal{H}$ in $\mathcal{X}$ and a set of m points $\mathcal{P} = \{x_1, \ldots, x_m\} \subset \mathcal{X}\}$, define:

$$s(\mathcal{H}, \mathcal{P}) := \#\Big\{\mathcal{A} \subset \mathcal{P} : \exists\, S \in \mathcal{H} \text{ such that } \mathcal{A} \subset S,\ \mathcal{P} \setminus \mathcal{A} \subset S^c\Big\}, \tag{5.16}$$

i.e., the number of different subsets of the data set $\mathcal{P}$ that can be separated by some element S in $\mathcal{H}$ from the rest in $\mathcal{P}$.

Note: $s(\mathcal{H}, \mathcal{P}) =$ the number of different outputs $(h_S(x) : x \in \mathcal{P}) \in \{+1, -1\}^{\#\mathcal{P}}$, $S \in \mathcal{H}$, the classifier based on $\mathcal{H}$ can have on $\mathcal{P}$.

Shatter number:

$$s(\mathcal{H}, m) := \max\{s(\mathcal{H}, \mathcal{P}) : \mathcal{P} \subset \mathcal{X},\ \#\mathcal{P} = m\} \tag{5.17}$$

is called the mth shatter number of $\mathcal{H}$ - the separation capacity of $\mathcal{H}$ on sets of cardinality m.

Ideally $\mathcal{H}$ can single out each subset of a set of m points, i.e.,

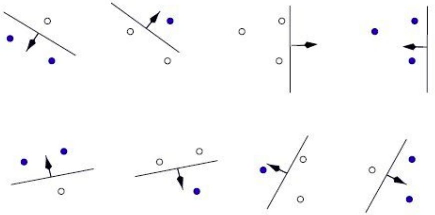
$$s(\mathcal{H}, m) = 2^m = \#\text{ of subsets of some set of cardinality } m. \tag{5.18}$$

In this case one says $\mathcal{H}$ shatters m points.

Shattering

When m increases eventually $\mathcal{H}$ ceases to shatter m points.

Example: $\mathcal{H} =$ set of all half-planes: $\exists\, \mathcal{P} \subset \mathbb{R}^2$, $\#\mathcal{P} = 3 = m$, $s(\mathcal{H}, m) = s(\mathcal{H}, \mathcal{P}) = 2^3 = \#2^{\mathcal{P}}$



- $\exists\, \mathcal{P} \subset \mathbb{R}^2$, $\#\mathcal{P} = 3 = m$, $s(\mathcal{H}, m) = s(\mathcal{H}, \mathcal{P}) = 2^3 = \#2^{\mathcal{P}}$
- Half-planes can no longer shatter $m = 4$ points.
- Thus, shatter numbers desribe the complexity of $\mathcal{H}$. To avoid overfitting, there should by a finite maximal number $m(\mathcal{H})$ such that $\mathcal{H}$ shatters $m(\mathcal{H})$ points but not $m(\mathcal{H}) + 1$ points.
- In this case one has

$$s(\mathcal{H}, k) = s(\mathcal{H}, m(\mathcal{H})) = 2^{m(\mathcal{H})}, \quad k \geq m(\mathcal{H}). \tag{5.19}$$

VC-Dimension

Definition 8

Given a hypohesis class $\mathcal{H} \subset 2^{\mathcal{X}}$ of subsets of the feature space $\mathcal{X}$, then

$$VC(\mathcal{H}) := \operatorname*{argmax}_{m \in \mathbb{N}} \{s(\mathcal{H}, m) = 2^m\} = \log_2 \big(\sup_{k \in \mathbb{N}} s(\mathcal{H}, k) \big) \tag{5.20}$$

is called the *VC*-dimension of $\mathcal{H}$.

Extension to function classes: Consider the epigraph of a function $f : \mathcal{X} \to \mathbb{R}$

$$epi(f) := \{(x, t) \in \mathcal{Z} : t \leq f(x)\}$$

which associates the set $epi(f)$ with f (specialize this to $f = h_S$).

Definition 9

Suppose that $\mathcal{H}$ is a class of functions (e.g. polynomials, splines, wavelets neural nets, etc. used in regression). Then we define

$$VC(\mathcal{H}) = VC\big(\{epi(f) : f \in \mathcal{H}\}\big). \tag{5.21}$$

VC-Dimension

A detailed discussion of VC-dimension concepts can be found in [4]. A useful fact is:

Proposition 10

Let $\mathcal{H}$ be an n-dimensional linear space of functions on $\mathcal{X}$ which are continuous almost everywhere. Then $VC(\mathcal{H}) = n + 1$.

A further useful fact needed later can be found in [1].

Lemma 11

One has

$$VC\big(\{S \triangle S_{\mathcal{H}} : S \in \mathcal{H}\}\big) \leq 2VC(\mathcal{H}). \tag{5.22}$$

Towards A Better Result

We indicate heuristically how the concept of VC-dimension leads to a better result than Proposition 4. Complete proofs can be found in [1].

- Recall: by (5.10) the relevant sets in the variance $|\eta_S - \eta_{S_\mathcal{H}} - (\hat{\eta}_S - \hat{\eta}_{S_\mathcal{H}})|$ are $S\vartriangle S_\mathcal{H}$, $S \in \mathcal{H}$. Let $\mathcal{H}^\Delta := \{S\vartriangle S_\mathcal{H} : S \in \mathcal{H}\}$.
- Intuitively, we do not need to estimate the probability in (5.12) for all S in $\mathcal{H}$ but only for $s(\mathcal{H}^\Delta, N)$ many "representers" in $\mathcal{H}^\Delta$ that can produce the $s(\mathcal{H}^\Delta, N)$ many different outcomes. We argue only formally to indicate how the VC-dimension enters. By (8), we have

$$s(\mathcal{H}^\Delta, N) = 2^{VC(\mathcal{H}^\Delta)} \overset{(5.22)}{\leq} 2^{2VC(\mathcal{H})}. \tag{5.23}$$

- Formally, replace the $\#\mathcal{H}$ in the threshhold $\epsilon_N(\mathcal{H}) := \frac{10(r\log N + \log \#\mathcal{H})}{3N}$ from (5.4) by $2^{2VC(\mathcal{H})}$, to obtain the new threshhold

$$\epsilon_N(\mathcal{H}) := \frac{10\big(r\log N + \log\big(2^{2VC(\mathcal{H})}\big)\big)}{3N} = \frac{10\big(r\log N + (2\log 2)VC(\mathcal{H})\big)}{3N}. \tag{5.24}$$

- Use this threshhold in (5.14) to obtain

$$2\exp\left\{-\frac{3N\epsilon_N}{10}\right\} = 2N^{-r}2^{-2VC(\mathcal{H})}$$

A Better Result

The union bound in (5.15) applies now to at most $s(\mathcal{H}^\Delta, N)$ representers in $\mathcal{H}$ so that this becomes

$$\mathrm{Prob}\Big\{\text{for every } S \in \mathcal{H} : |\eta_S - \eta_{S_\mathcal{H}} - (\hat\eta_S - \hat\eta_{S_\mathcal{H}})| \le \sqrt{\epsilon_N p_{S\Delta S_\mathcal{H}}} + \epsilon_N\Big\}$$
$$\ge 1 - 2^{2VC(\mathcal{H})} 2N^{-r} 2^{-2VC(\mathcal{H})} = 1 - 2N^{-r}. \qquad (5.25)$$

This argument is not rigorous. The critical issue remains to bound quantities like $\sup_{S\in\mathcal{H}} |\mathcal{R}[h_S] - \widehat{\mathcal{R}}[h_S]|$. This requires in the end much more involved arguments such as symmetrization in combination with concentration inequalities, see e.g. [2]. The following result is more advanced and uses VC-dimension in combination with Talagrad's inequality for empirical processes, [1].

Theorem 12

Assume that $VC(\mathcal{H}) < \infty$. Then for a sufficiently large constant $A > 0$ and for any fixed $r \in \mathbb{N}$

$$e_N(S) := \sqrt{p_{S\Delta S_\mathcal{H}} \epsilon_N} + \epsilon_N, \quad \epsilon_N = \epsilon_N(\mathcal{H}) = A \max\{r+1, VC(\mathcal{H})\} \frac{\log N}{N} \qquad (5.26)$$

there exists an absolute constant C_0 such that for any $N \ge 2$

$$\mathrm{Prob}_{\mathcal{Z}^N}\{\forall\, S \in \mathcal{H} : \ \big|\eta_{S_\mathcal{H}} - \eta_S - (\hat\eta_{S_\mathcal{H}} - \hat\eta_S)\big| \le e_N(S)\} \ge 1 - 2N^{-r}. \qquad (5.27)$$

A Better Result

To derive from the variance estimate in Theorem 12 a total excess risk bound, one can introduce the following modulus depending on the mapping e_N from(5.26):

$$\omega(p, e_N) := \sup \left\{ \int_{S\Delta S_{\mathcal{H}}} |\eta| dp_{\mathcal{X}} : S \in \mathcal{H} \text{ and } \int_{S\Delta S_{\mathcal{H}}} |\eta| dp_{\mathcal{X}} \leq 3e_N(S) \right\}. \tag{5.28}$$

Then one can show [1, Corollary 2.4]:

Theorem 13

Suppose that for some $\delta < 1$ (take $\delta = \delta(N) = 1 - C_0 N^{-r}$ from Theorem 12)

$$\mathrm{Prob}_{\mathcal{Z}^N} \left\{ \forall\, S \in \mathcal{H} : |\eta_S - \eta_{S_{\mathcal{H}}} - (\hat{\eta}_S - \hat{\eta}_{S_{\mathcal{H}}})| \leq e_N(S) \right\} \geq 1 - \delta, \tag{5.29}$$

then the ERM-minimizer $\hat{S} \in \mathcal{H}$ satisfies

$$\mathrm{Prob}_{\mathcal{Z}^N} \left\{ \mathcal{R}[h_{\hat{S}}] - \mathcal{R}[h_{S^*}] \leq \omega(p, e_N) + 2a(S^*, \mathcal{H}) \right\} \geq 1 - \delta, \tag{5.30}$$

where $a(S^*, \mathcal{H}) = \inf_{S \in \mathcal{H}} \mathcal{R}[h_S] - \mathcal{R}[h_{S^*}] = \mathcal{R}[h_{S_{\mathcal{H}}}] - \mathcal{R}[h_{S^*}]$ is the best approximation error from $\mathcal{H}$.

Concluding Remarks

- Results of very similar flavor hold also for regression, see e.g. [3].
- The main flavor is always: increasing r in (5.26) improves the success probability while increasing ϵ_N linearly.
- The S-dependent accuracy threshold $e_N(S)$ is always bounded by $2\sqrt{\epsilon_N}$.
- Keeping $VC(\mathcal{H})$ fixed and increasing the number N of samples shows that ϵ_N decreases like $\frac{\log N}{N}$ and the variance estimate (5.27) holds with increasing probability.
- However, keeping $VC(\mathcal{H})$ and hence the complexity of the hypothesis class $\mathcal{H}$ fixed would freeze the bias term $a(h_{S^*}, \mathcal{H})$ and the excess risk would not decrease below the approximation error.
- Thus to decrease the risk, one should also gradually increase the complexity of $\mathcal{H} = \mathcal{H}_N$, ideally, so as to balance the two terms representing variance and bias. This can be done by model selection, see e.g. [1].
- How fast the bias $a(h_{S^*}\mathcal{H}_N)$ as well as the modulus $\omega(p, e_N)$ can tend to zero as N grows, depends on the “regularity” of the decision boundary ∂S^* and on the way how the regression function $\eta(x)$ passes through ∂S^*. So called “margin conditions” describe this in a similar way as differentiability is used in approximation theory to characterize approximation properties of functions. A discussion of such conditions can also be found in [1].

Regression - a Sketch

Always: $\mathfrak{Z}_N = \{(\mathbf{x}^i, y_i), \ldots, (\mathbf{x}^N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$ i.i.d. samples from an unknown probability measure P with density p

Regression: $\mathcal{Y} \subset \mathbb{R}^n$ has a continuous range, objective: estimate the regression function $f^* = f_p(x) := \mathbb{E}[y|x]$ (see Lecture II, (8.3)).

Risks: For any $f \in L_2(\mathcal{X}, p_{\mathcal{X}}) \cap C(\mathcal{X})$, $f : \mathcal{X} \to \mathcal{Y}$, let

$$\mathcal{R}[f] := \int_{\mathcal{Z}} (y - f(x))^2 dP(x, y), \quad \widehat{\mathcal{R}}_{\mathfrak{Z}_N}[f] := \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2. \tag{5.31}$$

Recall: $dP(x, y) = dP(y|x) dP_{\mathcal{X}}(x)$, and the fact that $f_p := \mathbb{E}[y|x]$ is an orthogonal projection yield

$$\mathcal{R}[f] = \mathcal{R}[f_p] + \|f - f_p\|^2_{L_2(\mathcal{X}, p_{\mathcal{X}})}. \tag{5.32}$$

Given a hypothesis class $\mathcal{H}$, let

$$\hat{f} = \hat{f}_{\mathfrak{Z}_N} = \operatorname*{argmin}_{f \in \mathcal{H}} \widehat{\mathcal{R}}[f]. \tag{5.33}$$

Recall:

$$\mathcal{R}[\hat{f}] - \mathcal{R}[f^*] = \underbrace{\mathcal{R}[\hat{f}] - \mathcal{R}[f_{\mathcal{H}}]}_{\mathcal{E}_{var}} + \underbrace{\mathcal{R}[f_{\mathcal{H}}] - \mathcal{R}[f^*]}_{a(f^*, \mathcal{H})} \tag{5.34}$$

Setting

Assumptions:

- $|f_\rho(x)| = |f^*(x)| \leq \bar{M}$, a.e., i.e., y is bounded almost surely.
- $\mathcal{H} \subset C(\mathcal{X})$ is compact and $\exists\, M < \infty$ s.t.

$$|f(x) - y| \leq M, \quad a.e. \quad \forall\, f \in \mathcal{H}. \tag{5.35}$$

Some Consequences: samples are i.i.d.

$$\mathbb{E}\big[\widehat{\mathcal{R}}[f]\big] = \mathcal{R}[f], \quad \mathrm{var}[\widehat{\mathcal{R}}[f]] = \mathrm{var}\big[(Y - f(X))^2\big] =: \sigma^2 \leq M^4, \tag{5.36}$$

$$\rightsquigarrow \quad \mathrm{var}\big[\mathcal{R}[f] - \widehat{\mathcal{R}}[f]\big] = \sigma^2 - \mathcal{R}[f]^2 \tag{5.37}$$

Covering Numbers: $\mathcal{N}(\mathcal{H}, \delta) :=$ the smallest number of balls of radius δ that cover $\mathcal{H}$.

Since $\mathcal{H}$ is compact, $\mathcal{N}(\mathcal{H}, \delta) < \infty$; covering numbers quantify the complexity of $\mathcal{H}$, see [3, 7, 6]

Single Elements in $\mathcal{H}$

Proposition 14

Under the above assumptions: for a given $f \in \mathcal{H}$

$$\mathrm{Prob}_{\mathbb{Z}^N}\left\{|\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \le \delta\right\} \ge 1 - \min \begin{cases} 2e^{-\frac{N\delta^2}{2(\sigma^2 + M^2\delta/3)}} & \textit{(Bernstein's inequality)} \\ 2e^{-\frac{N\delta^2}{2M^2}} & \textit{(Hoeffding's inequality)} \end{cases} \quad (5.38)$$

Notice: when σ^2 is small compared with $M^2\delta$, Bernstein's inequality gives a better estimate.

To estimate the sample error or "variance" term

$$\mathcal{E}_{var} \le \left|\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}]\right| + \left|\widehat{\mathcal{R}}[f_{\mathcal{H}}] - \mathcal{R}[f_{\mathcal{H}}]\right|$$

as in the case of classification, we need to bound

$$\sup_{f \in \mathcal{H}} \left|\mathcal{R}[f] - \widehat{\mathcal{R}}[f]\right|$$

in probability.

Supremum over $\mathcal{H}$

Theorem 15

Under the above assumptions: one has for all $\delta > 0$

$$\mathrm{Prob}_{\mathbb{Z}^N}\left\{\sup_{f\in\mathcal{H}} |\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \le \delta\right\} \ge 1 - \mathcal{N}\left(\mathcal{H}, \frac{\delta}{8M}\right) \min \begin{cases} 2e^{-\frac{N\delta^2}{4(2\sigma^2 + M^2\delta/3)}} \\ 2e^{-\frac{N\delta^2}{2M^2}} \end{cases} \tag{5.39}$$

where $\sigma^2 = \sigma^2(\mathcal{H}) = \sup_{f\in\mathcal{H}} \mathrm{var}\left[(Y - f(X))^2\right] \le M^4$.

Preliminary observation:

Lemma 16

Underthe above assumptions: For any two $f, g \in \mathcal{H}$ *and* $\mathfrak{z}_N \in \mathcal{Z}^N$ *one has*

$$|(\mathcal{R}[f] - \widehat{\mathcal{R}}[f]) - (\mathcal{R}[g] - \widehat{\mathcal{R}}[g])| \le 4M\|f - g\|_{L_\infty(\mathcal{X})}. \tag{5.40}$$

The proof is elementary and uses the decomposition (see [3, Prop. 3])

$$(f(x) - y)^2 - (g(x) - y)^2 = (f(x) - g(x))(f(x) + g(x) - 2y).$$

Proof of Theorem 15: Let $m := \mathcal{N}(\mathcal{H}, \delta/(4M))$ and let f_j, $j = 1, \dots, m$, be the centers of balls $B_j \subset \mathcal{H}$, $j = 1, \dots, m$, covering $\mathcal{H}$. By Lemma 16

$$|(\mathcal{R}[f] - \widehat{\mathcal{R}}[f]) - (\mathcal{R}[f_j] - \widehat{\mathcal{R}}[f_j])| \leq 4M\|f - f_j\|_{L_\infty} \leq 4M\frac{\delta}{4M} = \delta, \quad \forall\, f \in B_j,$$

for any $\mathfrak{z}_N \in \mathcal{Z}^N$. Since therefore $|\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \leq |\mathcal{R}[f_j] - \widehat{\mathcal{R}}[f_j]| + \delta$ we conclude

$$\sup_{f \in B_j} |\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \geq 2\delta \quad \Rightarrow \quad |\mathcal{R}[f_j] - \widehat{\mathcal{R}}[f_j]| \geq \delta.$$

Thus, for each $j \in \{1, \dots, m\}$

$$\mathrm{Prob}_{\mathcal{Z}^N}\Big\{ \sup_{f \in B_j} |\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \geq 2\delta \Big\} \leq \mathrm{Prob}_{\mathcal{Z}^N}\Big\{ |\mathcal{R}[f_j] - \widehat{\mathcal{R}}[f_j]| \geq \delta \Big\} \leq e^{-\frac{N\delta^2}{2(\sigma^2(f_j - y) + M^2\delta/3)}}, \quad (5.41)$$

where we have used Proposition 14 in the last step.

The claim (5.39) follows now from the Union Bound, Lecture IV, Remark 4. □

Sample Error

$$\mathcal{E}_{var} \leq |\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}]| + |\mathcal{R}[f_{\mathcal{H}}] - \widehat{\mathcal{R}}[f_{\mathcal{H}}]| \Rightarrow$$

$$\begin{aligned} \mathrm{Prob}_{\mathcal{Z}^N}\left\{\mathcal{E}_{var} \geq \delta\right\} &\leq \mathrm{Prob}_{\mathcal{Z}^N}\left\{|\mathcal{R}[\hat{f}] - \widehat{\mathcal{R}}[\hat{f}]| + |\mathcal{R}[f_{\mathcal{H}}] - \widehat{\mathcal{R}}[f_{\mathcal{H}}]| \geq \delta\right\} \\ &\leq \mathrm{Prob}_{\mathcal{Z}^N}\left\{\sup_{f \in \mathcal{H}} |\mathcal{R}[f] - \widehat{\mathcal{R}}[f]| \geq \delta/2\right\} \end{aligned}$$

Use Theorem 15 with δ replaced by $\delta/2$ to conclude

Theorem 17

Under the above assumptions one has for $\delta > 0$

$$\mathrm{Prob}_{\mathcal{Z}^N}\left\{\mathcal{E}_{var} \leq \delta\right\} \geq 1 - \mathcal{N}\left(\mathcal{H}, \frac{\delta}{16M}\right) e^{-\frac{N\delta^2}{8(2\sigma^2 + M^2\delta/6)}} \tag{5.42}$$

Thus to achieve target accuracy δ with probability at least $1 - \eta$, it suffices to take

$$N \geq \frac{8\left(2\sigma^2 + \frac{M^2\delta}{6}\right)}{\delta^2}\left\{\ln\left(2\mathcal{N}(\mathcal{H}, \frac{\delta}{16M})\right) + |\ln \eta|\right\}. \tag{5.43}$$

Excess Risk

Main Goal: estimate $(f^*(x) = f_\rho(x) = \mathbb{E}[y|x])$

$$\mathcal{R}[\hat{f}] - \mathcal{R}[f_\rho] = \mathcal{E}_{var} + \mathcal{R}[f_{\mathcal{H}}] - \mathcal{R}[f_\rho] = \mathcal{E}_{var} + a(f_\rho, \mathcal{H}) \quad \rightsquigarrow$$

$$\begin{aligned} \mathrm{Prob}_{\mathcal{Z}^N}\Big\{\mathcal{R}[\hat{f}] - \mathcal{R}[f_\rho] \geq \delta + a(f_\rho, \mathcal{H})\Big\} \quad &= \quad \mathrm{Prob}_{\mathcal{Z}^N}\Big\{\mathcal{E}_{var} + a(f_\rho, \mathcal{H}) \geq \delta + ra(f_\rho, \mathcal{H})\Big\} \\ &= \quad \mathrm{Prob}_{\mathcal{Z}^N}\Big\{\mathcal{E}_{var} \geq \delta\Big\} \qquad (5.44) \end{aligned}$$

- For many choices of $\mathcal{H}$ and f with some smoothness properties one has

$$a(f, \mathcal{H}_D) \leq C(f)D^{-r}, \quad \mathcal{N}(\mathcal{H}_D, \eta) \sim \eta^{-a}, \quad \text{for some } r, a > 0.$$

- (5.43) with $\eta = N^{-b} \rightsquigarrow N \sim \delta^{-2}(d|\log \delta| + b \log N) \sim \delta^{-2}(a+b)\log N$, i.e., $\delta \sim \sqrt{\frac{(a+b)\log N}{N}}$
- Ideally: choose $\mathcal{H}_D$ such that $\delta \approx a(f_\rho, \mathcal{H}) \rightsquigarrow D \sim \left(\frac{N}{(a+b)\log N}\right)^{\frac{1}{2r}}, \rightsquigarrow$
- For A sufficiently large, with probability at least $1 - N^{-r}$ one has $\mathcal{R}[\hat{f}_{\mathfrak{z}_N}] - \mathcal{R}[f_\rho] \lesssim \sqrt{\frac{A \log N}{N}}, N \to \infty.$

References I

[1] P. Binev, A. Cohen, W. Dahmen, R. DeVore, Classification algorithms using adaptive partitioning, Annals of Statistics, 42 (No. 6)(2014), 2141-2163.
Supplement to “Classification algorithms using adaptive partitioning.” DOI:10.1214/14-AOS1234SUPP.

[2] O. Bousquet, S. Boucheron, G. Lugosi, Theory of Classification: A Survey of Recent Advances, ESAIM: Probability and Statistics, EDP Sciences, SMAI, 1999, http:///www.emath.fr/ps/

[3] F. Cucker, S. Smale, On the Mathematical Foundation of Learning, Bulletin of the American Mathematical Society, 39 (No 1)(2001), 1–49.

[4] L. Györfy, M. Kohler, A. Krzyzak, A. and H. Walk *A distribution-free theory of nonparametric regression*, Springer, Berlin, 2002.

[5] V. N. Vapnik. Statistical Learning Theory. Wiley, New York, 1998.

[6] Ding-Xuan Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Transactions on Information Theory (Volume: 49 , Issue: 7 , July 2003)

[7] Ying Guo ; P.L. Bartlett ; J. Shawe-Taylor ; R.C. Williamson, Covering numbers for support vector machines, IEEE Transactions on Information Theory (Volume: 48 , Issue: 1 , Jan 2002)