Machine Learning 3: KL divergence and lower bounds for deviations, PAC learning, No-Free-Lunch theorem

Master 2 Computer Science

Aurélien Garivier

2018-2019

Table of contents

Deviation bounds and kNN

Parenthesis: a nice proof for the technicalities of Bernstein

From [Pollard, MiniEmpirical ex.14, `http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf`]

For any sufficiently smooth real-valued function ϕ defined at least in a neighborhood of 0 let

$$G(x) = \frac{\phi(x) - \phi(0) - x\phi'(0)}{x^2/2} \text{ if } x \neq 0, \text{ and } G(0) = \phi''(0) \ .$$

By Taylor's integral formula

$$\phi(x) - \phi(0) - x\phi'(0) = \int_0^x \phi''(u)(x-u)du = x^2 \int_0^1 \phi''(sx)(1-s)ds \ .$$

Thus, $G(x) = \int \phi''(sx)d\nu(s)$, where $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$.

Hence, if ϕ is convex then $\phi'' \geq 0$ and $G \geq 0$. Moreover, if ϕ'' is increasing then the functions $x \mapsto \phi''(sx)$ for $s \in [0,1]$ are all increasing and G is also increasing as an average of increasing functions. For $\phi(u) = \exp(u)$, this yields that $(\exp(u) - u - 1)/u^2$ is increasing, as required for the proof of Bernstein's inequality.

Similarly, if ϕ'' is convex then G is also convex as an average of convex functions $\Big(x \mapsto \phi''(sx)\Big)_s$. Moreover, by Jensen's inequality applied to convex function $\psi(s) = \phi''(xs)$ with the probability measure $d\nu(s) = 2(1-s)\mathbb{1}\{0 \leq s \leq 1\}ds$

$$G(x) = \int_0^1 \phi''(xs)\, 2(1-s)ds \geq \phi''\left(x\int_0^1 s \times 2(1-s)ds\right) = \phi''\left(\frac{x}{3}\right) \ .$$

For $\phi(u) = (1+u)\log(1+u) - u$, $\phi''(u) = 1/(1+u)$ and this yields:

$$\frac{\phi(u)}{u^2/2} \geq \phi''\left(\frac{u}{3}\right) = \frac{1}{1+u/3} \ .$$

Exercise: for $X_i \overset{iid}{\sim} \mathcal{B}(\mu)$, $\mathbb{P}(\bar{X}_n \geq 2\mu) \leq \exp(-n\times ?)$

Chernoff + Taylor: since $\log(u) \geq (u-1)/u$,

$$\mathrm{kl}(2\mu, \mu) = 2\mu \log(2) + (1-2\mu)\log\frac{1-2\mu}{1-2\mu} \geq 2\mu\log(2) - \mu = \mu(2\log(2)-1) \approx 0.386\,\mu\ .$$

Chernoff with convexity:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{4/3\mu} = \frac{3}{8}\,\mu = 0.375\mu\ .$$

Improved Hoeffding:

$$\mathrm{kl}(2\mu, \mu) \geq \frac{(2\mu-\mu)^2/2}{\max_{\mu\leq u\leq 2\mu} u(1-u)} \geq \frac{\mu^2/2}{2\mu} = \frac{1}{4}\,\mu = 0.25\mu\ .$$

Bennett:

$$2\mu\log\frac{2\mu}{\mu} - (2\mu-\mu) = \mu(2\log(2)-1) \approx 0.386\,\mu\ .$$

Bernstein:

$$\frac{(2\mu-\mu)^2/2}{\mu(1-\mu)+(2\mu-\mu)/3} \geq \frac{\mu^2/2}{\mu+\mu/3}\,\frac{3}{8}\,\mu = 0.375\mu\ .$$

Hoeffding: $2(2\mu-\mu)^2 = 2\mu^2$, very poor (as expected) when μ is small.

Recall: risk bound for k-nearest neighbours

Let $\mathcal{C}_\epsilon$ be an ϵ-covering of $\mathcal{X}$:

$$\forall x \in X, \exists x' \in C_\epsilon : d(x, x') \leq \epsilon \, .$$

Excess risk for k-nearest-neighbours

If η is c-Lipschitz continuous: $\forall x, x' \in \mathcal{X}, \left|\eta(x) - \eta(x')\right| \leq c\, d\big(x, x'\big)$, then for all $k \geq 2$ and all $m \geq 1$:

$$\begin{aligned} L\big(\hat{h}^{kNN}\big) - L\big(h^*\big) &\leq \frac{1}{\sqrt{k\,e}} + \frac{2k|\mathcal{C}_\epsilon|}{m} + 4c\epsilon \\ &\leq \frac{1}{\sqrt{k\,e}} + (2+4c)\left(\frac{\alpha k}{m}\right)^{\frac{1}{d+1}} \quad \begin{cases} \text{for } \epsilon = \left(\frac{\alpha k}{m}\right)^{\frac{1}{d+1}} \,, \\ \text{if } |\mathcal{C}_\epsilon| \leq \alpha \epsilon^{-d} \end{cases} \\ &\leq (3+4c)\left(\frac{\alpha}{m}\right)^{\frac{1}{d+3}} \quad \text{for } k = \left(\frac{m}{\alpha}\right)^{\frac{2}{d+3}} \, . \end{aligned}$$

Bias-variance decomposition of the risk.

Room for improvement

- Lower bound? in $m^{-\frac{1}{d}}$.
- Margin conditions
  $\implies$ fast rates
- More regularity?
  $\implies$ weighted nearest neighbors
- Is regularity required everywhere?
  $\implies$ What matters are the balls of mass $\approx k/m$ near the decision boundary.

Research Article 1

Classification in general finite dimensional spaces with the k-nearest neighbor rule

by Sébastien Gadat, Thierry Klein, and Clément Marteau

Annals of Statistics Volume 44, Number 3 (2016), 982-1009.

arXiv: arXiv:submit/1106085

CLASSIFICATION WITH THE NEAREST NEIGHBOR RULE IN GENERAL FINITE DIMENSIONAL SPACES

By Sébastien Gadat and Thierry Klein and Clément Marteau

Toulouse School of Economics, Université Toulouse I Capitole
Institut Mathématiques de Toulouse, Université Paul Sabatier

Given an n-sample of random vectors $(X_i, Y_i)_{1 \leq i \leq n}$ whose joint law is unknown, the long-standing problem of supervised classification aims to *optimally* predict the label Y of a given a new observation X. In this context, the nearest neighbor rule is a popular flexible and intuitive method in non-parametric situations. Even if this algorithm is commonly used in the machine learning and statistics communities, less is known about its prediction ability in general finite dimensional spaces, especially when the support of the density of the observations is $\mathbb{R}^d$. This paper is devoted to the study of the statistical properties of the nearest neighbor rule in various situations. In particular, attention is paid to the marginal law of X, as well as the smoothness and margin properties of the *regression function* $\eta(X) = \mathbb{E}[Y|X]$. We identify two necessary and sufficient conditions to obtain uniform consistency rates of classification and to derive sharp estimates in the case of the nearest neighbor rule. Some numerical experiments are proposed at the end of the paper to help illustrate the discussion.

1. Introduction. The supervised classification model has been at the core of numerous contributions to statistical literature in recent years. It continues to provide interesting problems, both from the theoretical and practical point of views. The classical task in supervised classification is to predict a feature $Y \in \mathcal{M}$ when a variable of interest $X \in \mathbb{R}^d$ is observed, the set $\mathcal{M}$ being finite. In this paper, we focus on the binary classification problem where $\mathcal{M} = \{0, 1\}$.

In order to provide a prediction of the label Y of X, it is assumed that a training set $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is at our disposal, where (X_i, Y_i) are i.i.d. and with a common law $\mathbb{P}_{X,Y}$. This training set $\mathcal{S}_n$ makes it possible to retrieve some information on the joint law of (X, Y) and to provide, depending on some technical conditions, a pertinent prediction. In particular,

AMS 2000 subject classifications: Primary 62G05; secondary 62G20
Keywords and phrases: Supervised classification, nearest neighbor algorithm, plug in rules, minimax classification rates

Rates of convergence for nearest neighbor classification

by Kamalika Chaudhuri and Sanjoy Dasgupta

Advances in Neural Information Processing Systems 27 (NIPS 2014)

`https://papers.nips.cc/paper/5439-rates-of-convergence-for-nearest-neighbor-classification`

Rates of convergence for nearest neighbor classification

Kamalika Chaudhuri
Computer Science and Engineering
University of California, San Diego
kamalika@cs.ucsd.edu

Sanjoy Dasgupta
Computer Science and Engineering
University of California, San Diego
dasgupta@cs.ucsd.edu

Abstract

We analyze the behavior of nearest neighbor classification in metric spaces and provide finite-sample, distribution-dependent rates of convergence under minimal assumptions. These are more general than existing bounds, and enable us, as a by-product, to establish the universal consistency of nearest neighbor in a broader range of data spaces than was previously known. We illustrate our upper and lower bounds by introducing a new smoothness class customized for nearest neighbor classification. We find, for instance, that under the Tsybakov margin condition the convergence rate of nearest neighbor matches recently established lower bounds for nonparametric classification.

1 Introduction

In this paper, we deal with binary prediction in metric spaces. A classification problem is defined by a metric space $(\mathcal{X}, \rho)$ from which instances are drawn, a space of possible labels $\mathcal{Y} = \{0, 1\}$, and a distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. The goal is to find a function $h : \mathcal{X} \to \mathcal{Y}$ that minimizes the probability of error on pairs (X, Y) drawn from $\mathbb{P}$; this error rate is the *risk* $R(h) = \mathbb{P}(h(X) \neq Y)$. The best such function is easy to specify: if we let μ denote the marginal distribution of X and η the conditional probability $\eta(x) = \mathbb{P}(Y = 1|X = x)$, then the predictor $1(\eta(x) \geq 1/2)$ achieves the minimum possible risk, $R^* = \mathbb{E}_X[\min(\eta(X), 1 - \eta(X))]$. The trouble is that $\mathbb{P}$ is unknown and thus a prediction rule must instead be based only on a finite sample of points $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn independently at random from $\mathbb{P}$.

Nearest neighbor (NN) classifiers are among the simplest prediction rules. The *1-NN classifier* assigns each point $x \in \mathcal{X}$ the label Y_i of the closest point in $X_1, \ldots, X_n$ (breaking ties arbitrarily, say). For a positive integer k, the *k-NN classifier* assigns x the majority label of the k closest points in $X_1, \ldots, X_n$. In the latter case, it is common to let k grow with n, in which case the sequence $(k_n : n \geq 1)$ defines a *k_n-NN classifier*.

The asymptotic consistency of nearest neighbor classification has been studied in detail, starting with the work of Fix and Hodges [7]. The risk of the NN classifier, henceforth denoted R_n, is a random variable that depends on the data set $(X_1, Y_1), \ldots, (X_n, Y_n)$; the usual order of business is to first determine the limiting behavior of the expected value $\mathbb{E}R_n$ and to then study stronger modes of convergence of R_n. Cover and Hart [2] studied the asymptotics of $\mathbb{E}R_n$ in general metric spaces, under the assumption that every x in the support of μ is either a continuity point of η or has $\mu(\{x\}) > 0$. For the 1-NN classifier, they found that $\mathbb{E}R_n \to \mathbb{E}_X[2\eta(X)(1 - \eta(X))] \leq 2R^*(1 - R^*)$; for k_n-NN with $k_n \uparrow \infty$ and $k_n/n \downarrow 0$, they found $\mathbb{E}R_n \to R^*$. For points in Euclidean space, a series of results starting with Stone [15] established consistency without any distributional assumptions. For k_n-NN in particular, $R_n \to R^*$ almost surely [5].

These consistency results place nearest neighbor methods in a favored category of nonparametric estimators. But for a fuller understanding it is important to also have rates of convergence. For

Kullback-Leibler divergence

Kullback-Leibler divergence

Definition

Let P and Q be two probability distributions on a measurable set Ω. The Kullback-Leibler divergence from Q to P is defined as follows:

- if P is not absolutely continuous with respect to Q, then $\mathrm{KL}(P, Q) = +\infty$;
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

$$\mathrm{KL}(P, Q) = \int_\Omega \log \frac{dP}{dQ}\, dP = \int_\Omega \frac{dP}{dQ} \log \frac{dP}{dQ}\, dQ\ .$$

Property: $0 \leq \mathrm{KL}(P, Q) \leq +\infty$, $\mathrm{KL}(P, Q) = 0$ iff $P = Q$.

If $P \ll Q$ and $f = \frac{dP}{dQ}$, $\int_\Omega f \log(f)\, dQ = \int_\Omega \left[f \log(f)\right]_+ dQ - \int_\Omega \left[f \log(f)\right]_-\, dQ$, the later is finite since $\left[f \log(f)\right]_- \leq 1/e$.

Examples:

$\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = \mathrm{kl}(p, q)$, $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

Properties

Tensorization of entropy:

If $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, then

$$\mathrm{KL}(P, Q) = \mathrm{KL}(P_1, Q_1) + \mathrm{KL}(P_2, Q_2)\,.$$

Contraction of entropy data-processing inequality:

Let $(\Omega, \mathcal{A})$ be a measurable space, and let P and Q be two probability measures on $(\Omega, \mathcal{A})$. Let $X : \Omega \to (\mathcal{X}, \mathcal{B})$ be a random variable, and let P^X (resp. Q^X) be the push-forward measures, ie the laws of X wrt P (resp. Q). Then

$$\mathrm{KL}\left(P^X, Q^X\right) \leq \mathrm{KL}(P, Q)\,.$$

Pinsker's inequality:

Let $P, Q \in \mathfrak{M}_1(\Omega, \mathcal{A})$. Then

$$\|P - Q\|_{TV} \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \leq \sqrt{\frac{\mathrm{KL}(P, Q)}{2}}\,.$$

Proof: contraction

Contraction: if $\mathrm{KL}(P, Q) = +\infty$, the result is obvious. Otherwise, $P \ll Q$ and there exists $\frac{dP}{dQ} : \Omega \to \mathbb{R}$ such that for all measurable $f : \Omega \to \mathbb{R}$, $\int_\Omega f \, dP = \int_\Omega f \, \frac{dP}{dQ} \, dQ$.

- We first prove that $P^X \ll Q^X$ and, if $\gamma(x) := \mathbb{E}_Q\left[\frac{dP}{dQ} \middle| X = x\right]$ is the Q-a.s. unique function such that $\mathbb{E}_Q\left[\frac{dP}{dQ} \middle| X\right] = \gamma(X)$, then $\gamma = \frac{dP^X}{dQ^X}$. Indeed, for all $B \in \mathcal{B}$,

$$
\begin{aligned}
P^X(B) = P(X \in B) &= \int_{X \in B} \frac{dP}{dQ} dQ = \mathbb{E}_Q\left[\frac{dP}{dQ} \mathbb{1}\{X \in B\}\right] \\
&= \mathbb{E}_Q\left[\mathbb{E}_Q\left[\frac{dP}{dQ} \mathbb{1}\{X \in B\} \middle| X\right]\right] = \mathbb{E}_Q\left[\mathbb{1}\{X \in B\} \mathbb{E}_Q\left[\frac{dP}{dQ} \middle| X\right]\right] \\
&= \mathbb{E}_Q\left[\mathbb{1}\{X \in B\} \gamma(X)\right] = \int_{X \in B} \gamma(X) dQ = \int_B \gamma dQ^X
\end{aligned}
$$

  and hence $P^X \ll Q^X$ and $\frac{dP^X}{dQ^X} = \gamma$.
- Now,

$$
\begin{aligned}
\mathrm{KL}\left(P^X, Q^X\right) &= \int_{\mathcal{X}} \gamma \log \gamma \; dQ^X = \int_\Omega \gamma(X) \log \gamma(X) \, dQ \\
&= \mathbb{E}_Q\left[\phi\left(E_Q\left[\frac{dP}{dQ} \middle| X\right]\right)\right] \quad \text{where } \phi := x \mapsto x \log(x) \text{ is convex} \\
&\leq \mathbb{E}_Q\left[\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right) \middle| X\right]\right] \qquad \text{by (conditional) Jensen's inequality} \\
&= \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right] = \mathrm{KL}(P, Q) \, .
\end{aligned}
$$

Proof: Pinsker

Let $A \in \mathcal{A}$, $p = P(A)$ and $q = Q(A)$. By contraction,

$$\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^{\mathbb{1}_A}, Q^{\mathbb{1}_A}) = \mathrm{KL}\left(\mathcal{B}\big(P(A)\big), \mathcal{B}\big(Q(A)\big)\right) = \mathrm{kl}\left(P(A), Q(A)\right) \geq 2\big(P(A) - Q(A)\big)^2 .$$

Application: Lower bound
"Chernoff's bound is asymptotically almost tight"

Let $\mu \in (0,1)$. $X_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{B}(\mu)$, and let $x \in (\mu, 1]$. Then

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\bar{Y}_n > x) \geq - \mathrm{kl}(x, \mu) \ .$$

Proof: Let $\epsilon > 0$ and on the same probability space let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{B}(x+\epsilon)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{B}(\mu)$. Then

$$\begin{aligned}
n\,\mathrm{kl}(x+\epsilon, \mu) &= \mathrm{KL}\left(P^{\mathbf{X}}, P^{\mathbf{Y}}\right) \qquad \text{by tensorization} \\
&\geq \mathrm{KL}\left(P^{\mathbb{1}\{\bar{X}_n \geq x\}}, P^{\mathbb{1}\{\bar{Y}_n \geq x\}}\right) \qquad \text{by contraction} \\
&= \mathrm{kl}\left(\mathbb{P}(\bar{X}_n \geq x), \mathbb{P}(\bar{Y}_n \geq x)\right) \\
&\geq \mathbb{P}(\bar{X}_n \geq x) \log \frac{1}{\mathbb{P}(\bar{Y}_n \geq x)} - \log(2)
\end{aligned}$$

since $\mathrm{kl}(p,q) = -h(p) + p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q}$. Hence, by Hoeffding's inequality,

$$\liminf_m \frac{1}{n} \log \mathbb{P}(\bar{Y}_n > x) \geq \liminf_n \frac{-n\,\mathrm{kl}(x+\epsilon,\mu) + \log(2)}{n\big(1 - \exp(-2n\epsilon^2)\big)} = -\,\mathrm{kl}(x+\epsilon, \mu)$$

for all $\epsilon > 0$, and we conclude by the continuity of $\mathrm{kl}(\cdot, \mu)$.
Note that one can also derive non-asymptotic lower bounds.

PAC learning

Learning framework

- Underlying distribution D on $\mathcal{X} \times \mathcal{Y}$.
- Sample $S \overset{iid}{\sim} D$ (otherwise: transductive learning).
- $h : \mathcal{X} \to \mathcal{Y}$, $h \in \mathcal{H}$ hypothesis class.
- loss function $l(y, y')$ (regression, classification)
- generalization error (loss) $L_D(h)$
- training error $L_S(h)$
- Realizable assumption: there exists h^* such that $L_S(h^*) = 0$.
- Antonym: *agnostic* learning.

Empirical risk minimization with inductive bias

Definition

Any learning algorithm $\hat{h}_m$ of the form

$$ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\arg\min}\, L_S(h)$$

is called a *empirical risk minimizer*.

Risk of overfitting

PAC learnability: "probably approximately correct"

Definition

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_m$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D_X on $\mathcal{X}$ and for every labelling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, D_X, f$ then when $S = \big((X_1, f(X_1)), \dots, (X_m, f(X_m))\big)$ with $(X_i)_{1 \le i \le m} \overset{iid}{\sim} D_X$,

$$\mathbb{P}\Big(L_{(D_X,f)}(\hat{h}_m) \ge \epsilon\Big) \le 1 - \delta$$

for all $m \ge m_{\mathcal{H}}(\epsilon, \delta)$.
The smallest possible function $m_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

Remark: Valiant's PAC requires also sample complexity and running time polynomial in $1/\epsilon$ and $1/\delta$.

Examples

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbb{1}\{x \leq a\}$ is PAC-learnable with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{log \frac{2}{\delta}}{\epsilon} \right\rceil .$$
Proof: let a^* be such that $L_D(h_{a*}) = 0$ and let $a_0 = \inf\{a : D_X([a, a^*]) \leq \epsilon\}$ and $a_1 = \sup\{a : D_X([a^*, a]) \leq \epsilon\}$.
An ERM is $\hat{h}_S(x) = \mathbb{1}_{x \leq T}$ where $T \in [B_0, B_1]$, with $B_0 = \max\{x : (x, 1)^i nS\}$ and $B_1 = \min\{x : (x, 0)^i nS\}$. Then
$P(L(\hat{h}_S) \geq \epsilon) \leq= \mathbb{P}(B_0 < a_0) + \mathbb{P}(B_1 > a_1)$. Since $D_X(a_0, a^*) \geq \epsilon$ and
$\mathbb{P}(B_0 < a_0) \leq (1 - D_X([a_0, a^*])^m \leq \exp(-m\epsilon)$.
- Exercise: Learning axis-aligned rectangles: given real numbers $a_1 \leq b_1$ and $a_2 \leq b_2$, let
$$h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2\,; \\ 0 & \text{otherwise}\,. \end{cases}$$
Let $\mathcal{H}^2_{\mathrm{rec}} = \{h_{(a_1, b_1, a_2, b_2)} : a_1 \leq b_1 \text{ and } a_2 \leq b_2\}$. Show that $\mathcal{H}^2_{\mathrm{rec}}$ is PAC-learnable, with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{4 log \frac{4}{\delta}}{\epsilon} \right\rceil .$$

Finite hypothese classes are PAC-learnable

The sample complexity of finite hypothese classes in the realizable case is smaller than $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$:

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\epsilon, \delta \in (0,1)$ and let m be an integer that satisfies

$$m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon} .$$

Then, for any labeling function f and for any distribution D_X on $\mathcal{X}$, under the realizability assumption, with probability at least $1 - \delta$ over the choice of iid sample S of size m, any ERM hypothesis $\hat{h}_m$ is such that

$$L_{(D_X, f)}(\hat{h}_m) \leq \epsilon .$$

Proof

The realizability assumption implies that an ERM $\hat{h}_S$ has empirical risk $L_S(\hat{h}_S) = 0$. Hence,

$$
\begin{aligned}
\mathbb{P}\left(L(\hat{h}_S) \geq \epsilon\right) &= D_X^{\otimes m}\left(\left\{S \in \mathcal{X}^m : \exists h \in \mathcal{H}, L_S(h) = 0 \text{ and } L_D(h) \geq \epsilon\right\}\right) \\
&= D_X^{\otimes m}\left(\bigcup_{h : L_D(h) \geq \epsilon} S_h\right) \quad \text{where } S_h = \left\{S \in \mathcal{X}^m : L_s(h) = 0\right\} \\
&\leq \sum_{h : L_D(h) \geq \epsilon} D_X^{\otimes m}(S_h) \\
&= \sum_{h : L_D(h) \geq \epsilon} \prod_{i=1}^{m} \underbrace{D_X\left(\left\{x \in \mathcal{X} : h(x) = f(x)\right\}\right)}_{= 1 - L_D(h) \leq 1 - \epsilon} \\
&\leq \sum_{h : L_{(D_X, f)}(h) \geq \epsilon} \prod_{i=1}^{m} (1 - \epsilon) \leq \left|\mathcal{H}\right| (1 - \epsilon)^m \leq \left|\mathcal{H}\right| \exp(-m\epsilon) \,.
\end{aligned}
$$

This quantity is smaller than δ for $m \geq \dfrac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$.