

DSAA3073: Theories in Data Science

Week 3 Validation Sheet · From Empirical Risk to Finite-Sample Complexity

Coverage: Week 3

Duration: 30 minutes · Total: 130 marks

Answer all questions. For multiple-choice questions, select exactly one answer. Show the essential reasoning requested; unsupported answers may receive no credit.

Part A: Multiple choice

A1 [5 marks]. On a sample S , a classifier h_1 makes 3 mistakes out of 12 and h_2 makes 2 mistakes out of 12. Which conclusion follows from these facts alone?

- (a) $R(h_2) = \frac{1}{6}$.
- (b) h_2 is an empirical risk minimizer within $\{h_1, h_2\}$.
- (c) h_2 has smaller population risk than h_1 .
- (d) The distribution is realizable by $\{h_1, h_2\}$.

Solution. (b). Under zero-one loss, the empirical risks are $\hat{R}_S(h_1) = \frac{3}{12} = \frac{1}{4}$ and $\hat{R}_S(h_2) = \frac{2}{12} = \frac{1}{6}$. Therefore h_2 minimizes empirical risk over the displayed two-element class. These sample counts do not determine either population risk, and a nonzero empirical risk does not establish realizability. Hence neither (a), (c), nor (d) follows.

Marking. 3 marks for (b); 2 for distinguishing the empirical comparison from unsupported population-risk or realizability conclusions.

A2 [5 marks]. Which observation is sufficient to establish that a distribution is realizable by a binary class $\mathcal{H}$?

- (a) Some $h \in \mathcal{H}$ has zero empirical risk on one observed sample.
- (b) ERM returns a unique hypothesis.
- (c) Some $h^* \in \mathcal{H}$ has population risk $R(h^*) = 0$.
- (d) The sample size exceeds the number of hypotheses.

Solution. (c). Realizability is the distributional condition $\exists h^* \in \mathcal{H} : R(h^*) = 0$. A classifier may fit one finite sample perfectly while still making errors elsewhere in the population, so (a) is insufficient. Uniqueness of ERM and a comparison between sample size and class cardinality are not parts of the definition.

Marking. 3 marks for (c); 2 for identifying population risk, rather than one realized training error, as the defining quantity.

A3 [5 marks]. On one three-point set C , a binary class realizes exactly seven distinct label vectors. Which statement is necessarily correct?

- (a) C is shattered.
- (b) $\Pi_{\mathcal{H}}(3) = 7$.
- (c) $\Pi_{\mathcal{H}}(3) \geq 7$, but C is not shattered.
- (d) $\text{VCdim}(\mathcal{H}) = 3$.

Solution. (c). A three-point set is shattered only when all $2^3 = 8$ binary vectors occur, so seven vectors are insufficient. Because the growth function is a maximum over all three-point sets, this one restriction proves the lower bound $\Pi_{\mathcal{H}}(3) \geq 7$, not equality. It also does not prove that any three-point set is shattered.

Marking. 3 marks for (c); 2 for using both the eight-vector shattering requirement and the maximum in the growth-function definition.

A4 [5 marks]. Suppose $\text{VCdim}(\mathcal{H}) \leq 2$. What does Sauer–Shelah guarantee at $m = 5$?

- (a) $\Pi_{\mathcal{H}}(5) = 10$.
- (b) $\Pi_{\mathcal{H}}(5) \leq 1 + 5 + 10 = 16$.
- (c) $|\mathcal{H}| \leq 16$.
- (d) Every five-point set is shattered.

Solution. (b). Sauer–Shelah gives

$$\Pi_{\mathcal{H}}(5) \leq \sum_{i=0}^2 \binom{5}{i} = 1 + 5 + 10 = 16.$$

This is an upper bound on the number of distinct restrictions, not an exact value and not a bound on the total cardinality of the hypothesis class. Since $16 < 2^5$, it also rules out shattering a five-point set.

Marking. 3 marks for (b); 2 for evaluating the binomial sum and interpreting it as a growth-function upper bound.

Part B: Definition in use

B1 [20 marks]. A finite class $\mathcal{H} = \{h_1, \dots, h_K\}$ is evaluated on the same i.i.d. sample S . For every index j fixed before observing S , bounded-loss concentration gives

$$P(|\hat{R}_S(h_j) - R(h_j)| > \gamma) \leq 2 \exp(-2m\gamma^2).$$

After seeing all empirical risks, ERM returns $h_{J(S)}$. An analyst simply replaces the fixed j by $J(S)$ in the displayed inequality. Audit that step, and derive a correct confidence- $1 - \delta$ statement that protects whichever hypothesis ERM selects. Give a sufficient sample-size condition.

Solution. The displayed inequality has the quantifier order “for each index j fixed before the sample, bound the probability of its bad event.” The ERM index $J(S)$ is a function of the same empirical risks computed from S . Substituting it for fixed j changes the premise: the selected row may be the row whose empirical risk was unusually favorable. Thus the fixed-index statement cannot be applied to $h_{J(S)}$ by direct substitution.

For each fixed j , define

$$E_j(\gamma) = \{|\hat{R}_S(h_j) - R(h_j)| > \gamma\}.$$

A sample on which the selected classifier violates the tolerance belongs to at least one of these fixed-index events. Therefore

$$\{|\hat{R}_S(h_{J(S)}) - R(h_{J(S)})| > \gamma\} \subseteq \cup_{j=1}^K E_j(\gamma).$$

The union bound, which does not require the E_j to be independent, gives

$$P(\cup_{j=1}^K E_j(\gamma)) \leq \sum_{j=1}^K P(E_j(\gamma)) \leq 2K \exp(-2m\gamma^2).$$

Consequently, with probability at least $1 - 2K \exp(-2m\gamma^2)$, one simultaneous event holds:

$$\forall h \in \mathcal{H}, \quad |\hat{R}_S(h) - R(h)| \leq \gamma.$$

This event is strong enough for $h_{J(S)}$ because it controls every candidate before the data-dependent index is inspected.

To obtain confidence at least $1 - \delta$, it is sufficient that

$$2K \exp(-2m\gamma^2) \leq \delta.$$

Taking natural logarithms and rearranging yields

$$m \geq \frac{\ln(2\frac{K}{\delta})}{2\gamma^2}.$$

An integer certificate is $m = \left\lceil \frac{\ln(2\frac{K}{\delta})}{2\gamma^2} \right\rceil$. This repaired statement protects not only the ERM output but every possible sample-dependent choice from the finite class.

Marking. 5 marks for identifying that $J(S)$ is sample-dependent and stating the quantifier failure; 4 for defining the fixed-index bad events and relating selected failure to their union; 5 for the union calculation $2K \exp(-2m\gamma^2)$; 3 for the simultaneous confidence statement; 3 for the sufficient integer sample size. Preserve method and follow-through credit after an algebra error in a valid simultaneous-event repair.

Part C: Basic skills

C1 [7 marks]. A finite binary class has $|\mathcal{H}| = 100$. Assume realizability and that ERM returns a consistent hypothesis. Compute a sufficient sample size for $\varepsilon = 0.20$ and $\delta = 0.10$ using

$$m \geq \left(\frac{1}{\varepsilon}\right) \left(\ln|\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)\right).$$

Round upward.

Solution. The supplied formula applies under the stated zero-one-loss realizability and consistency conditions. Substitution gives

$$m \geq \left(\frac{1}{0.20}\right) (\ln 100 + \ln 10) = 5 \ln(1000) \approx 5(6.9078) \approx 34.54.$$

Since the sample size is an integer, rounding upward gives $m = 35$ as a sufficient sample size. The calculation supplies a sufficient certificate; it does not claim that 35 samples are necessary.

Marking. 2 marks for selecting the supplied realizable-consistent formula under its stated conditions; 3 for correct substitution and logarithmic calculation; 2 for rounding upward to 35 and interpreting sufficiency.

C2 [8 marks]. On $C = \{x_1, x_2, x_3\}$, five hypotheses induce the vectors

$$000, \quad 001, \quad 001, \quad 011, \quad 111.$$

Compute $|\mathcal{H}|_C$, give the conclusion about $\Pi_{\mathcal{H}}(3)$, and decide whether C is shattered.

Solution. A restriction counts distinct label vectors, so the duplicate row 001 is counted once. Hence

$$\mathcal{H}|_C = \{000, 001, 011, 111\}, \quad |\mathcal{H}|_C = 4.$$

Since the growth function maximizes over all three-point sets, this fixed set proves $\Pi_{\mathcal{H}}(3) \geq 4$. It does not prove equality, because another three-point set might realize more vectors. Finally, shattering requires all $2^3 = 8$ vectors, and only four occur here, so C is not shattered.

Marking. 3 marks for collapsing the duplicate and obtaining four restrictions; 3 for the lower-bound conclusion $\Pi_{\mathcal{H}}(3) \geq 4$ without claiming equality; 2 for correctly rejecting shattering using the eight-vector requirement.

Part D: Construction and obstruction

D1 [17 marks]. Let $\mathcal{T} = \{h_t : t \in \mathbb{R}\}$ be the threshold class $h_t(x) = \mathbf{1}_{\{x \geq t\}}$ on $\mathbb{R}$. Determine $\text{VCdim}(\mathcal{T})$. Your proof must include a concrete shattered set for the lower bound and an obstruction that works for every set of the next larger size.

Solution. The exact value is $\text{VCdim}(\mathcal{T}) = 1$.

Lower bound. Fix any one point, for example $C = \{0\}$. To realize label 1, choose $t = 0$, so $h_0(0) = 1$. To realize label 0, choose $t = 1$, so $h_1(0) = 0$. Both binary labelings of this one-point set occur, hence $\mathcal{T}$ shatters C and $\text{VCdim}(\mathcal{T}) \geq 1$.

Universal upper bound. Take an arbitrary pair of distinct real numbers and order it as $x_1 < x_2$. The labeling $(1, 0)$ is impossible. Indeed, if a threshold labels x_1 positive, then $x_1 \geq t$. Since $x_2 > x_1$, we also have $x_2 \geq t$, so x_2 must be positive as well. Thus every two-point set has at least this missing labeling and no two-point set is shattered. Hence $\text{VCdim}(\mathcal{T}) < 2$.

Combining the existential one-point witness with the obstruction for every two-point set gives $\text{VCdim}(\mathcal{T}) = 1$.

Marking. 3 marks for the correct VC value; 5 for realizing both labels on one explicit point; 3 for beginning the upper bound with an arbitrary ordered pair; 4 for proving that $(1, 0)$ is impossible; 2 for combining the quantified lower and upper bounds.

D2 [18 marks]. Let $\mathcal{I} = \{h_{a,b} : a \leq b\}$ be the interval class $h_{a,b}(x) = \mathbf{1}_{\{a \leq x \leq b\}}$ on $\mathbb{R}$. Establish its exact VC dimension. For the lower bound, use the set $\{-2, 3\}$ and give an explicit interval for each of its four target labelings. For the upper bound, your obstruction must apply to an arbitrary three-point set.

Solution. The exact value is $\text{VCdim}(\mathcal{I}) = 2$.

Lower bound. On the ordered set $\{-2, 3\}$, the four target vectors can be realized as follows:

$$(0, 0) : [5, 6], \quad (1, 0) : [-3, -1], \quad (0, 1) : [2, 4], \quad (1, 1) : [-3, 4].$$

Direct membership checks show that each interval contains exactly the points assigned label 1. Therefore this two-point set is shattered and $\text{VCdim}(\mathcal{I}) \geq 2$.

Universal upper bound. Let an arbitrary three-point set be ordered as $z_1 < z_2 < z_3$. Consider the target labeling $(1, 0, 1)$. Any interval that contains both z_1 and z_3 contains every real number between its endpoints. Since $z_1 < z_2 < z_3$, it must also contain z_2 , contradicting the requested middle label 0. Thus $(1, 0, 1)$ is missing on every three-point set, so no such set is shattered and $\text{VCdim}(\mathcal{I}) < 3$.

The two-point construction and arbitrary-triple obstruction together give $\text{VCdim}(\mathcal{I}) = 2$.

Marking. 3 marks for the correct VC value; 7 for four valid explicit intervals and membership verification on $\{-2, 3\}$; 3 for quantifying over an arbitrary ordered triple; 3 for the interval-convexity obstruction to $(1, 0, 1)$; 2 for combining the two proof directions. Equivalent explicit intervals receive full credit.

Part E: Transfer to a new ordered class

E1 [25 marks]. Define a binary class $\mathcal{P}$ on $\mathbb{R}$ as follows: the positive region of a classifier is either a left ray $(-\infty, t]$ or a right ray $[t, \infty)$, for some $t \in \mathbb{R}$. Thus each classifier labels either an initial portion or a final portion of the real line as 1.

Determine the exact growth function $\Pi_{\mathcal{P}}(m)$ for every integer $m \geq 1$ and the exact value of $\text{VCdim}(\mathcal{P})$. Give a complete proof of both claims; a count from one specially chosen set alone does not establish the growth function or the VC upper bound.

Solution. The decisive representation is to order an arbitrary m -point set. Let $C = \{x_1, \dots, x_m\}$ with

$$x_1 < x_2 < \dots < x_m.$$

A left ray can label as positive exactly an initial segment of this order. Its restriction vectors are

$$0^m, \quad 10^{m-1}, \quad 110^{m-2}, \quad \dots, \quad 1^m.$$

There are $m + 1$ such prefix vectors, including the all-zero and all-one vectors. Similarly, a right ray produces exactly the $m + 1$ suffix vectors

$$0^m, \quad 0^{m-1}1, \quad 0^{m-2}11, \quad \dots, \quad 1^m.$$

For $m \geq 1$, the prefix and suffix lists overlap exactly twice: at 0^m and 1^m . To justify that there are no other duplicates, observe that every nonconstant prefix vector begins with 1 and ends with 0, whereas every nonconstant suffix vector begins with 0 and ends with 1. Hence

$$|\mathcal{P}|_C = (m + 1) + (m + 1) - 2 = 2m.$$

The argument began with an arbitrary set of m distinct real points. Every such set therefore has exactly $2m$ restrictions, so maximizing over C gives

$$\Pi_{\mathcal{P}}(m) = 2m \quad \text{for every } m \geq 1.$$

We now determine the VC dimension with both required quantifiers. For the lower bound, take two points $x_1 < x_2$. The prefix and suffix restrictions include

$$00, \quad 10, \quad 01, \quad 11.$$

Thus all four binary labelings occur on this one two-point set, and $\text{VCdim}(\mathcal{P}) \geq 2$.

For the upper bound, take an arbitrary triple $z_1 < z_2 < z_3$. A classifier in $\mathcal{P}$ can only produce a prefix or suffix of positive labels. In particular, it cannot produce the alternating labeling $(1, 0, 1)$: a left ray containing z_3 contains the two smaller points, while a right ray containing z_1 contains the two larger points. Thus every triple has a missing labeling and no triple is shattered. Equivalently, the already proved exact count gives only $\Pi_{\mathcal{P}}(3) = 6 < 8$. Hence $\text{VCdim}(\mathcal{P}) < 3$.

Combining both directions yields

$$\text{VCdim}(\mathcal{P}) = 2,$$

while the exact growth rule remains $\Pi_{\mathcal{P}}(m) = 2m$ for all $m \geq 1$. Notice that the class is infinite even though its restrictions grow only linearly.

Marking. 6 marks for choosing an ordered representation of an arbitrary m -point set and identifying the prefix and suffix families; 5 for counting their two overlaps and proving $\Pi_{\mathcal{P}}(m) = 2m$ for every $m \geq 1$; 5 for the existential two-point shattering construction; 6 for a universal arbitrary-triple obstruction; 3 for the exact VC conclusion and correct interpretation of infinite class versus finite-sample growth. Preserve method and follow-through credit after a minor counting slip. A count on only one chosen set, or a VC value without both quantifiers, earns at most 10 marks.

Part F: Compare two valid finite-class routes

F1 [15 marks]. Let $|\mathcal{H}| = 64$ and $\delta = 0.05$. Compare the following two requests.

- Request S asks for one event on which every $h \in \mathcal{H}$ satisfies

$$|\hat{R}_S(h) - R(h)| \leq \gamma \quad \text{with } \gamma = 0.10.$$

- Request R assumes binary zero-one loss, realizability, and a consistent ERM, and asks that its returned hypothesis have risk at most $\varepsilon = 0.10$.

Compute a sufficient sample size for each request using the appropriate Week 3 finite-class route. Then explain why the smaller number cannot replace the larger one when Request S is required.

Solution. Request S is a two-sided simultaneous-deviation request for every member of a finite class. Its sufficient sample size is

$$m_S \geq \frac{\ln\left(2 \frac{|\mathcal{H}|}{\delta}\right)}{2\gamma^2} = \frac{\ln\left(2 \cdot \frac{64}{0.05}\right)}{2(0.10)^2} = \frac{\ln 2560}{0.02} \approx 392.39.$$

Rounding upward gives $m_S = 393$.

Request R has the stronger assumptions of zero-one loss, realizability, and consistency, and it asks only that the returned consistent hypothesis avoid population risk above ε . The direct bad-consistent route gives

$$m_R \geq \left(\frac{1}{\varepsilon}\right) \left(\ln|\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)\right) = \left(\frac{1}{0.10}\right) (\ln 64 + \ln 20) \approx 10(4.1589 + 2.9957) \approx 71.55.$$

Thus $m_R = 72$ is sufficient.

The two numbers certify different events. The m_S argument controls two-sided empirical-to-population deviations simultaneously for every classifier and has tolerance cost $\frac{1}{\gamma^2}$. The m_R argument controls only the possibility that a population-bad classifier nevertheless makes zero training errors; it uses realizability and

consistency and has risk cost $\frac{1}{\epsilon}$. With $m = 72$, that specialized argument may certify the risk of a consistent ERM, but it does not imply that every empirical risk is within 0.10 of its population risk. Therefore the smaller sample certificate cannot replace $m_S = 393$ when Request S is the goal.

Marking. 5 marks for selecting the simultaneous-deviation route, substituting correctly, and rounding to $m_S = 393$; 4 for selecting the consistent-realizable route, substituting correctly, and rounding to $m_R = 72$; 4 for comparing the different events, assumptions, and $\frac{1}{\gamma^2}$ versus $\frac{1}{\epsilon}$ roles; 2 for explicitly rejecting use of the smaller certificate for Request S. Preserve route-selection and follow-through credit after arithmetic errors.