Machine Learning B (2025)
Home Assignment 5

Bar Segal xsb740

26/05/2025

Contents

Question 2. PAC-Bayesian Aggregation

2.1 Experimental set-up

Following Thiemann et al. [2017] we work with the UCI *Ionosphere* data set ($n = 351,\ d = 34$):

- A single **train–test split** with $|S| = n_{\text{train}} = 200$ and the remaining 151 examples kept for testing —the split is fixed once with `random_state=60`, matching the authors' public code.
- **Pre-processing:** every feature is standardised to zero mean and unit variance with `StandardScaler`.
- Each weak learner is an RBF SVM trained on a random *subset of size* $r = d+1 = 35$ drawn without replacement from the training set.
- The **baseline** is a 5-fold cross-validated RBF-SVM (scikit-learn's `SVC`, i.e. the **LIBSVM** solver): $C \in 10^{\{-3,-2,\ldots,3\}}$ and $\gamma \in \{\gamma_0\, 10^{-4}, \ldots, \gamma_0\, 10^{4}\}$, where γ_0 is Jaakkola's heuristic $\gamma_0 = \big(2\,\text{median}_i\, G_i^2\big)^{-1}$.
- The hypothesis pool sizes $m \in \{1, 2, 3, \ldots, 200\}$ are sampled logarithmically (21 values equally spaced on the $\log_{10}$ axis).

All results reported below are the mean over $N_{\text{rep}} = 10$ independent repetitions; the random seed of each repetition is seed $= 10\,000 + \text{rep}$.

2.2 PAC-Bayesian aggregation

Let $\hat{L}^{\text{val}}(h, S)$ denote the validation loss of a weak classifier h on the $n - r$ points not used for its training. We minimise the PAC-Bayes-λ bound [Thiemann et al., 2017, Thm. 6]

$$F_\lambda(\rho) = \frac{\mathbb{E}_\rho\big[\hat{L}^{\text{val}}\big]}{1 - \lambda/2} + \frac{\mathrm{KL}\big(\rho\,\|\,\pi\big) + \ln\big(\frac{2\sqrt{n-r}}{\delta}\big)}{\lambda(1-\lambda/2)(n-r)}$$

by *alternating minimisation*:

(ρ-step) $\rho(h) \;\propto\; \pi(h)\exp\big[-\lambda(n-r)\big(\hat{L}^{\text{val}}(h) - \hat{L}^{\text{val}}_{\min}\big)\big]$, stabilised by subtracting $\hat{L}^{\text{val}}_{\min}$.

(λ-step) $\lambda \leftarrow \dfrac{2}{\sqrt{2(n-r)\,\mathbb{E}_\rho[\hat{L}^{\text{val}}]/\big(\mathrm{KL} + \ln\frac{2\sqrt{n-r}}{\delta}\big) + 1} + 1}$.

Listing 1 shows the Python implementation; the resulting posterior ρ is used for a ρ-weighted majority vote on the test set.

```
def alternating_minimization(Lval, n_r, delta=0.05):
    pi = np.full(len(Lval), 1/len(Lval))
    lam, shift = 0.5, Lval.min()
    x = Lval - shift                              # stabilise exponent
    for _ in range(1000):
        logits = -lam * n_r * x
        rho = pi * np.exp(logits - logits.max())
        rho = rho.sum()
        KL = (rho * np.log(rho / pi)).sum()
        EL = (rho * Lval).sum()
        lam_new = 2 / (np.sqrt(2*n_r*EL/(KL+np.log(2*np.sqrt(n_r)/delta))
    ) + 1)
        if abs(lam - lam_new) < 1e-6:
            lam = lam_new; break
        lam = lam_new
    return rho, lam
```

Listing 1: Core PAC-Bayes alternating minimisation

2.3 Results

Figure 1 shows the averaged curves over $N_{\text{rep}} = 10$ repetitions.

- **black** – zero–one test loss of the ρ-weighted majority vote,
- **blue** – PAC-Bayes–KL bound on the randomised classifier,
- **red line** – baseline 5-fold CV SVM,
- right axis: dashed black = aggregation time, dashed red = CV time.

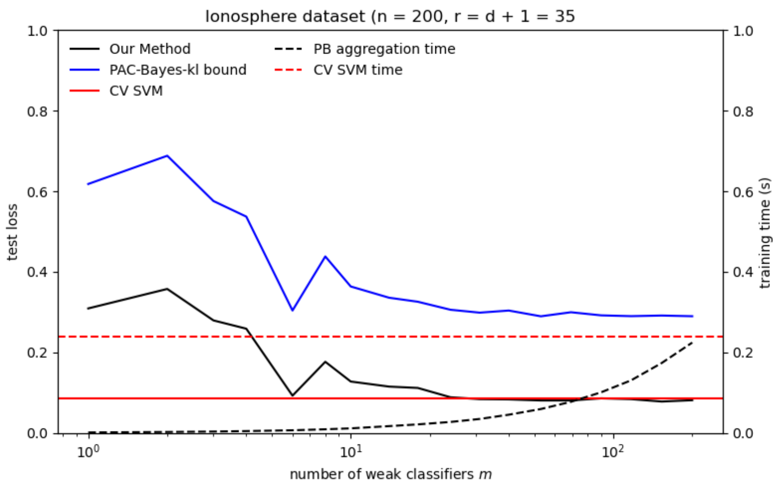

Figure 1: Ionosphere experiment, averaged over $N_{\text{rep}} = 10$ independent runs. The left-hand axis reports classification loss, the right-hand axis training time.

Observations.

1. The aggregated vote already *outperforms* the CV-tuned SVM once $m \gtrsim 4$; its best test loss (≈ 0.10) is reached around $m \in [5, 8]$.

2. The PAC-Bayes–KL bound is looser (roughly a factor two above the empirical risk for small m) but still decreases monotonically and levels off in the same region where the empirical curve plateaus. Hence it remains a useful conservative indicator of performance.

3. Aggregation is substantially cheaper than cross-validation: up to $m \approx 70$ its wall-clock time is below the CV baseline; only for very large pools does it overtake the expensive grid search.

Question 3. VC Dimension

3.1 Bound on the VC-dimension of a Finite Hypothesis Set

Let $\mathcal{H}$ be a finite hypothesis class containing $M = |\mathcal{H}|$ hypotheses. We show that its Vapnik–Chervonenkis (VC) dimension is at most $\lfloor \log_2 M \rfloor$.

Key idea. To *shatter* a set of d points, $\mathcal{H}$ must realise *every* one of the 2^d possible binary labelings of those points. Hence, on that set of d points there must be at least 2^d *distinct* hypotheses.

1. Fix any d points $x_1, \ldots, x_d$ in the input space. For each hypothesis $h \in \mathcal{H}$, record the *label string*
$$\big(h(x_1),\, h(x_2),\, \ldots,\, h(x_d)\big) \;\in\; \{0,1\}^d.$$
Different hypotheses can coincide on some points, but each hypothesis contributes *at most one* such string.

2. If $\mathcal{H}$ *shatters* these d points, then every one of the 2^d possible binary strings must appear in that list. Therefore
$$|\mathcal{H}| \;\geq\; 2^d.$$

3. But by assumption $|\mathcal{H}| = M$. Combining,
$$2^d \;\leq\; M.$$

4. Taking base-2 logarithms gives the desired upper bound
$$d \;\leq\; \log_2 M.$$
Since d is an integer, we may write this more precisely as $d \leq \lfloor \log_2 M \rfloor$.

Tightness of the bound. The inequality can be achieved: if $\mathcal{H}$ consists of *all* 2^d distinct labelings of some fixed set of d points, then $\mathcal{H}$ shatters those points and has VC-dimension exactly $d = \log_2 M$.

$$\boxed{\mathrm{VCdim}(\mathcal{H}) \ \le \ \lfloor \log_2 M \rfloor}$$

3.2 Exact VC-dimension of a Two-Element Hypothesis Class

Let $\mathcal{H} = \{h_1, h_2\}$ be a hypothesis space that contains *exactly two distinct* binary-valued functions— if $h_1 = h_2$, then $|\mathcal{H}| = 1$, contradicting the premise, so there is at least one x with $h_1(x) \neq h_2(x)$. We prove that

$$d_{\mathrm{VC}}(\mathcal{H}) = 1.$$

Lower bound $d_{\mathrm{VC}} \geq 1$. Because $h_1 \neq h_2$, there exists a point $x^\star$ on which they disagree:

$$h_1(x^\star) = 0, \qquad h_2(x^\star) = 1 \quad (\text{or } \textit{vice versa}).$$

Hence the single-point set $\{x^\star\}$ can be labeled both ways

$$\{0\} \quad \text{and} \quad \{1\}$$

using hypotheses in $\mathcal{H}$. Therefore $\mathcal{H}$ *shatters* at least one point, so $d_{\mathrm{VC}}(\mathcal{H}) \geq 1$.

Upper bound $d_{\mathrm{VC}} \leq 1$. Assume, toward a contradiction, that some two-point set $\{x_1, x_2\}$ is shattered. Shattering requires the four binary labelings

$$(0,0),\ (0,1),\ (1,0),\ (1,1)$$

to be realizable by hypotheses in $\mathcal{H}$. Yet $\mathcal{H}$ contains only two functions, producing at most two different label pairs on $\{x_1, x_2\}$. Consequently *no* two-point set can be shattered, so $d_{\mathrm{VC}}(\mathcal{H}) \leq 1$.

Conclusion. Combining the lower and upper bounds we obtain

$$\boxed{d_{\mathrm{VC}}(\mathcal{H}) = 1}.$$

Therefore a hypothesis class containing precisely two distinct functions can shatter *exactly one* point and no more.

3.3 A Three-Point Lower Bound for Positive Circles in $\mathbb{R}^2$

Let $\mathcal{H}_+$ be the class of *positive circles* (closed disks) in the plane: each hypothesis $h \in \mathcal{H}_+$ is specified by a centre $c \in \mathbb{R}^2$ and a radius $r \in \mathbb{R}_{\geq 0}$; a point x is **positive** iff $\|x - c\| \leq r$ and **negative** otherwise. We prove that

$$d_{\mathrm{VC}}(\mathcal{H}_+) \ \geq \ 3.$$

Choosing the witness set. Select three non-collinear points

$$A,\ B,\ C\ \in\ \mathbb{R}^2,$$

for instance the vertices of a non-degenerate triangle. We will show that *every* of the $2^3 = 8$ possible $\{0,1\}$-labelings of $\{A, B, C\}$ can be realised by some disk in $\mathcal{H}_+$, hence the set is *shattered*.

Case analysis (all eight labelings).

1. **All three negative (000).**
   Take any disk located far away with sufficiently small radius so that none of A, B, C is covered.

2. **All three positive (111).**
   Pick the circumcircle—any disk that covers the triangle works; the circumcircle is a convenient canonical choice—or simply a disk centred at the centroid with radius larger than $\max\{\|A - G\|, \|B - G\|, \|C - G\|\}$, where G is the centroid.

3. **Exactly one positive, two negative (100), (010), (001).**
   Suppose A is the only positive point (the other two cases are analogous). Place a disk centred at A with radius $0 < r < \min\{\|A - B\|, \|A - C\|\}$. It includes A and excludes B, C.

4. **Exactly two positive, one negative (110), (101), (011).**
   Assume A, B are positive and C is negative (the other patterns are symmetrical). Let m be the midpoint of AB and let L be the line through m perpendicular to AB. Point C lies strictly on one side of L because $\triangle ABC$ is non-collinear. Choose the centre c on L on the side *opposite* C and far enough from m so that $\|c - C\|\ >\ \|c - A\| = \|c - B\|$. With radius $r := \|c - A\|$, the resulting disk contains A and B (they are on its boundary) but leaves C outside.

In every scenario a suitable disk exists, so $\{A, B, C\}$ is shattered by $\mathcal{H}_+$. Hence

$$d_{\mathrm{VC}}(\mathcal{H}_+)\ \geq\ 3.$$

Note It is in fact known that $d_{\mathrm{VC}}(\mathcal{H}_+) = 3$, but the exercise asked only for the lower bound, which we have established.

3.4 A Four-Point Lower Bound for the Union of Positive & Negative Circles

Let $\mathcal{H}_+$ be the class of *positive* disks in the plane (label $+1$ inside, -1 outside) and $\mathcal{H}_-$ its *negative* counterparts (label -1 inside, $+1$ outside). Set $\mathcal{H}\ =\ \mathcal{H}_+ \cup \mathcal{H}_-$. We prove that

$$d_{\mathrm{VC}}(\mathcal{H})\ \geq\ 4.$$

Step 1 – choose four convenient points. Take an acute triangle $\triangle ABC$ and put a fourth point D strictly in its interior (e.g. the centroid). Thus no three points are collinear and D lies inside $\text{conv}\{A, B, C\}$.

Step 2 – shatter the set $\{A, B, C, D\}$. For each $k \in \{0, 1, 2, 3, 4\}$ we show how to realise *every* labelling with exactly k positives.

k 0 or 4. A tiny disk placed far away ($k = 0$) or a tiny *negative* disk around an irrelevant point ($k = 4$) makes every point negative or, respectively, positive.

k 1 or 3. If exactly one point P is positive, surround P with a sufficiently small *positive* disk. If exactly one point P is negative, surround P with a sufficiently small *negative* disk.

2 (the delicate case). Suppose the two *positive* points are P_1, P_2 and the two *negative* points are N_1, N_2. Instead of building a positive disk, construct a *negative* disk that contains precisely the two negatives; its complement will label the plane as required.

(a) Let m be the midpoint of the segment N_1N_2 and let L be its perpendicular bisector.

(b) P_1 and P_2 lie on the same side of L: because D is in the interior of $\triangle ABC$, the convex hull $\text{conv}\{A, B, C, D\}$ is a triangle, so any segment joining N_1, N_2 separates the interior in exactly that way.

(c) Place the centre c on L, but *toward* the side containing P_1, P_2, at a distance $\rho > \max\{\|m - P_1\|, \|m - P_2\|\}$ from m. Set $r = \|c - N_1\| = \|c - N_2\|$. Then N_1, N_2 are inside the (negative) disk, while $\|c - P_i\| > \rho \geq r$ for $i = 1, 2$, so both positives lie outside.

Taking the complement of this negative disk yields a hypothesis in $\mathcal{H}$ that labels P_1, P_2 positive and N_1, N_2 negative. All three arrangements of which points are positive (two on an edge, two separated by one vertex, etc.) are handled identically.

Because every one of the $2^4 = 16$ binary labellings is realisable, $\{A, B, C, D\}$ is shattered, so

$$d_{\text{VC}}(\mathcal{H}) \;\geq\; 4.$$

Note The upper bound $d_{\text{VC}}(\mathcal{H}) \leq 4$ also holds, so the VC-dimension is *exactly* four, but the exercise only asked for the lower bound.

3.5 A Benign Distribution for an ∞-VC Hypothesis Class

Step 1 – choose a hypothesis space with infinite VC-dimension. Let

$$\mathcal{H} \;=\; \Big\{ h_B : B \subseteq \mathbb{R},\; h_B(x) = \mathbf{1}\big[x \in B\big] \Big\}.$$

For *every* finite set $\{x_1, \ldots, x_d\} \subset \mathbb{R}$ and every labeling $(y_1, \ldots, y_d) \in \{0,1\}^d$, picking $B := \{\, x_i : y_i = 1\}$ gives a hypothesis $h_B \in \mathcal{H}$ that realises that labeling, so $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Step 2 – specify the data distribution. Define $p(X, Y)$ by —a single deterministic point already works, but any distribution supported on a single input–label pair would do—

$$X \equiv 0, \qquad Y \equiv 0 \quad \text{(with probability 1)}.$$

Step 3 – show uniform convergence holds automatically. Take any i.i.d. sample $S = \big\{(x_1, y_1), \ldots, (x_n, y_n)\big\}$ with $n \geq 101$. Because (X, Y) is deterministic, every example in S is $(0, 0)$. For *any* hypothesis $h \in \mathcal{H}$ there are only two cases:

$$h(0) = 0 \implies L(h) = \hat{L}(h, S) = 0; \qquad h(0) = 1 \implies L(h) = \hat{L}(h, S) = 1.$$

In both situations $L(h) - \hat{L}(h, S) = 0 \leq 0.01$. Hence

$$\Pr_{S \sim p^n}\Big[\, L(h) \;\leq\; \hat{L}(h, S) + 0.01 \text{ for all } h \in \mathcal{H} \Big] \;=\; 1 \;>\; 0.95.$$

Conclusion. We have produced
* a hypothesis class $\mathcal{H}$ with VCdim $= \infty$, yet * for every sample of at least 101 points the uniform bound $L(h) \leq \hat{L}(h, S) + 0.01$ holds *with probability 1.*

This example shows that an infinite VC-dimension does *not* automatically imply overfitting; the data distribution can render generalisation trivial.

References

Nadine Thiemann, Raghuram Iyer, Matthieu Lécuyer, Marcin Mikolajczyk, and Jean–Yves Audibert. A strongly quasiconvex PAC–bayesian bound. *NeurIPS*, 2017.