Machine Learning B (2025)
Home Assignment 6

Bar Segal xsb740

02/06/2025

Contents

Analysing AdaBoost

Question 1: Closed–form expression for the AdaBoost weights

We show, by induction over the boosting round b, that the weight update

$$w_i^{(b+1)} = \frac{w_i^{(b)} \exp\bigl(-\alpha_b\, y_i\, h_b(x_i)\bigr)}{\sum_{j=1}^{n} w_j^{(b)} \exp\bigl(-\alpha_b\, y_j\, h_b(x_j)\bigr)}$$

can be written in the closed form

$$\boxed{w_i^{(b+1)} = \frac{\exp\bigl(-y_i\, f_b(x_i)\bigr)}{\displaystyle\sum_{j=1}^{n} \exp\bigl(-y_j\, f_b(x_j)\bigr)}} \qquad \text{with} \quad f_b(x) = \sum_{p\le b} \alpha_p\, h_p(x).$$

Base case $b = 0$. Before any weak learner is trained we take the uniform distribution $w_i^{(1)} = 1/n$. Because $f_0 \equiv 0$, the formula gives $w_i^{(1)} = \exp(0)/\sum_j \exp(0) = 1/n$, so the claim holds.

Inductive step. Assume the closed form is true for round b. Insert that expression into the AdaBoost update:

$$\begin{aligned} w_i^{(b+1)} &= \frac{\exp\bigl(-y_i\, f_{b-1}(x_i)\bigr) \exp\bigl(-\alpha_b\, y_i\, h_b(x_i)\bigr)}{\displaystyle\sum_{j=1}^{n} \exp\bigl(-y_j\, f_{b-1}(x_j)\bigr) \exp\bigl(-\alpha_b\, y_j\, h_b(x_j)\bigr)} \\ &= \frac{\exp\bigl(-y_i\, \bigl[f_{b-1} + \alpha_b h_b\bigr](x_i)\bigr)}{\displaystyle\sum_{j=1}^{n} \exp\bigl(-y_j\, \bigl[f_{b-1} + \alpha_b h_b\bigr](x_j)\bigr)} = \frac{\exp\bigl(-y_i\, f_b(x_i)\bigr)}{\displaystyle\sum_{j=1}^{n} \exp\bigl(-y_j\, f_b(x_j)\bigr)}. \end{aligned}$$

Thus the statement holds for $b + 1$, completing the induction. □

Question 2: Number of boosting rounds for zero training error

Assume each weak learner satisfies $\Pr\bigl[\varepsilon_b \le \frac{1}{2} - \gamma\bigr] = 1 - \delta'$ for some $\gamma \in (0, \frac{1}{2}]$ and $\delta' \in (0, 1)$, where ε_b is the weighted error in round b. Define the edge $\gamma_b = \frac{1}{2} - \varepsilon_b$; whenever the guarantee holds, $\gamma_b \ge \gamma$.

AdaBoost training-error bound. For any sequence of errors $\varepsilon_1, \ldots, \varepsilon_B$,

$$\text{train-error} \le \prod_{b=1}^{B} Z_b, \qquad Z_b = 2\sqrt{\varepsilon_b(1 - \varepsilon_b)} = \sqrt{1 - 4\gamma_b^2}.$$

If $\gamma_b \geq \gamma$ in every round,

$$\text{train-error} \;\leq\; \bigl(1-4\gamma^2\bigr)^{B/2} \;\leq\; e^{-2\gamma^2 B}. \tag{$*$}$$

Choosing B^*. Requiring the bound ($*$) to be below $1/n$ ensures zero mistakes on the n training points:

$$e^{-2\gamma^2 B^*} \;<\; \frac{1}{n} \implies \boxed{B^* \;=\; \left\lceil \tfrac{\ln n}{2\gamma^2} \right\rceil}.$$

Overall confidence. All B^* rounds must satisfy the weak-learner guarantee. By a union bound,

$$\Pr\bigl[\text{all rounds good}\bigr] \;\geq\; 1 - B^*\,\delta'.$$

Hence AdaBoost's final hypothesis h obeys

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\bigl(h(x_i) \neq y_i\bigr) = 0 \quad \text{with probability at least } 1-\delta^*,$$

where

$$\boxed{\delta^* \;=\; 1-(1-\delta')^{B^*} \;\leq\; B^*\delta'}\,.$$

Therefore both B^* and δ^* depend only on γ, n (and on δ' supplied by the weak learner), exactly as required. □

Landcover Classification

AdaBoost baseline with central pixel features

Feature design Each 13×13 image patch contains 6 spectral bands measured at 12 time stamps ($13 \times 13 \times 6 \times 12 = 12\,168$ values per pixel). Training AdaBoost on the full cube would be slow because every weak tree would need to evaluate thousands of candidate splits and the grid-search would explode in runtime. To accelerate the experiment we retain only the 72 values of the central pixel (6 bands × 12 dates). This cuts the dimensionality by a factor of 170 and reduces the whole grid search to well under a minute on a laptop.

The drawback is the loss of spatial context: texture cues and neighbourhood correlations are invisible to the classifier, so geometrically defined classes such as *artificial_surface* become harder to recognise.

Model selection. A discrete AdaBoost ensemble with decision-tree weak learners was tuned on

$$\text{max\_depth} \in \{1,2,3\}, \qquad n\text{\_estimators} \in \{50,100,200\},$$

using 2-fold cross–validation as required. The best configuration was

$$\boxed{\text{max\_depth} = 3,\ n\text{\_estimators} = 200}, \qquad \text{CV accuracy} = 0.7505.$$

Training each candidate took 1.7–13 s, so the complete grid (18 fits) finished in about one minute.

Test performance. After refitting that model on the full training set we obtained

$$\textbf{Test accuracy} = 0.72.$$

Per-class metrics are listed in Table 1, and the confusion matrix is shown in Fig. 1.

Class	Precision	Recall	F_1	Support
cultivated_land	0.43	0.45	0.44	44
forest	0.88	0.83	0.86	1155
grassland	0.45	0.60	0.52	224
shrubland	0.28	0.34	0.30	153
water	0.88	0.54	0.67	13
artificial_surface	0.48	0.26	0.34	42
bareland	0.71	0.58	0.64	108
Overall accuracy		**0.72**		

Table 1: Per-class precision, recall and F_1 on the test set.

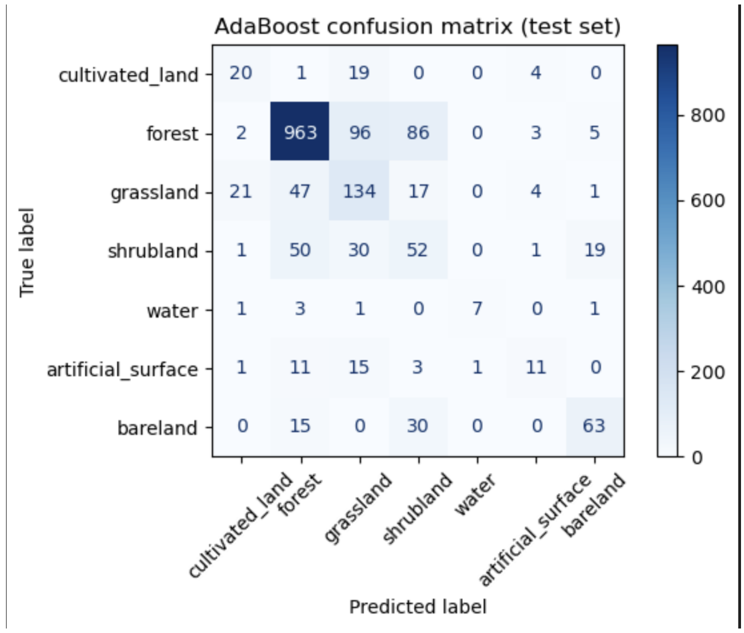

Figure 1: Confusion matrix of the AdaBoost model on the test set.

Discussion. The majority class *forest* is recognised very reliably ($F_1 = 0.86$) thanks to abundant training examples. Classes with few pixels, especially *water* and *artificial_surface*, suffer from data scarcity and the lack of spatial features—many urban pixels are confused with *bareland*. Nonetheless, the 72-feature baseline already delivers a respectable 72 % overall accuracy with a training time of only seconds, providing a solid starting point for richer feature sets or deeper boosted trees.

Majority Vote

Throughout, let the risk (generalisation error) of a classifier $h : X \to \{0,1\}$ be

$$L(h) \;=\; \Pr_{x\sim P}\big[h(x) \neq Y(x)\big], \qquad L(\mathrm{MV}) \;=\; \Pr_{x\sim P}\big[\mathrm{MV}(x) \neq Y(x)\big],$$

and let the uniformly weighted majority vote over a finite hypothesis set $H = \{h_1, \dots, h_M\}$ be

$$\mathrm{MV}(x) \;=\; \mathbf{1}\Big(\sum_{j=1}^{M} h_j(x) \;>\; M/2\Big).$$

Part 1: $L(\mathrm{MV}) = 0$ while every $L(h) \geq \frac{1}{3}$

Take an instance space $X = \{x_1, x_2, x_3\}$ with the uniform distribution and (for convenience) a constant label $Y(x) = 0$. Define three hypotheses as indicated:

instance	Y	h_1	h_2	h_3
x_1	0	**1**	0	0
x_2	0	0	**1**	0
x_3	0	0	0	**1**

- Each h_j is wrong on exactly one of the three equiprobable points, hence $L(h_j) = \frac{1}{3}$ for $j = 1, 2, 3$.
- At every point two of the three hypotheses are correct, therefore the majority vote is always correct and $L(\mathrm{MV}) = 0$.

Part 2: $L(\mathrm{MV}) > L(h)$ for every $h \in H$

Let $X = \{a, b, c, d\}$ be uniformly distributed and again $Y(x) = 0$ for all x. Define

instance	Y	h_1	h_2	h_3
a	0	**1**	**1**	0
b	0	**1**	0	**1**
c	0	0	**1**	**1**
d	0	0	0	0

- Each hypothesis errs on two of the four points, so $L(h_j) = \frac{1}{2}$.
- On a, b, c at least two hypotheses predict 1 (wrong), hence $\mathrm{MV}(x) = 1$ there; on d everyone is correct. Therefore $L(\mathrm{MV}) = \frac{3}{4} > \frac{1}{2}$.

Part 3: Independent errors majority vote converges to 0

Assume $|H| = M$ and that all hypotheses share the same error rate

$$L(h) = p = \frac{1}{2} - \varepsilon < \frac{1}{2}, \qquad \varepsilon > 0,$$

with independent errors. For a random point x let $S = \sum_{j=1}^{M} \mathbf{1}\{h_j(x) \neq Y(x)\} \sim \mathrm{Binomial}(M, p)$. The majority vote is wrong when $S > M/2$, hence

$$L(\mathrm{MV}) \ = \ \Pr\big[S > M/2\big] \ = \ \sum_{k=\lfloor M/2 \rfloor + 1}^{M} \binom{M}{k} p^k (1-p)^{M-k}.$$

Using Hoeffding's inequality with mean $\mu = Mp$ and deviation $M\varepsilon$,

$$L(\mathrm{MV}) \ = \ \Pr[S - \mu \geq M\varepsilon] \ \leq \ \exp\big(-2M\varepsilon^2\big) \xrightarrow[M \to \infty]{} 0.$$

therefore, under independent and better than chance voters, the majority vote becomes arbitrarily accurate even though each individual classifier may be close to random guessing ($L(h) \lesssim \frac{1}{2}$).

Bottom line. Independent errors with every $L(h) < \frac{1}{2}$ guarantee improvement and ultimately $L(\mathrm{MV}) \to 0$. In contrast, if some voters are worse than chance or if errors are highly correlated, the majority can perform poorly and its error can approach twice the maximum individual error, as demonstrated in Part 2(a).

Occam's kl-razor vs. PAC-Bayes-kl

Let $S = \{(x_i, y_i)\}_{i=1}^{n}$ be an i.i.d. sample, let $\ell \in [0, 1]$ be a bounded loss, and denote by $\hat{L}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(h(x_i), y_i\big)$ the empirical loss of $h\!: X \to [0, 1]$. For a (data–independent) prior π over a countable hypothesis set $\mathcal{H}$, write $\mathrm{kl}(q \| p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}$ for the binary relative entropy.

Part 1: Proof of Theorem 3.40 (Occam's kl-razor for soft selection)

Statement of the theorem Under the conditions of Theorem 3.38(or 3.41 in the new lecture notes), for any $\delta \in (0, 1]$ it holds that

$$\Pr\bigg(\exists\, \rho \ : \ \mathrm{kl}\big(\, \mathbb{E}_\rho[\hat{L}(h, S)] \,\|\, \mathbb{E}_\rho[L(h)]\big) \ \geq \ \frac{\mathbb{E}_\rho\Big[\ln \dfrac{1}{\pi(h)}\Big] + \ln \frac{1}{\delta}}{n}\bigg) \ \leq \ \delta.$$

Proof. Fix any posterior distribution ρ over $\mathcal{H}$ (possibly data–dependent). Because the function $(q, p) \mapsto \mathrm{kl}(q\|p)$ is jointly convex, Jensen's inequality gives

$$\mathrm{kl}\big(\mathbb{E}_\rho[\hat{L}] \,\|\, \mathbb{E}_\rho[L]\big) \;\le\; \mathbb{E}_\rho\Big[\mathrm{kl}\big(\hat{L}(h,S) \,\|\, L(h)\big)\Big]. \tag{1}$$

Now apply Theorem 3.38 simultaneously to every $h \in \mathcal{H}$ and note that, with probability at least $1-\delta$,

$$\mathrm{kl}\big(\hat{L}(h,S) \,\|\, L(h)\big) \;\le\; \frac{\ln\frac{1}{\pi(h)\delta}}{n} \quad \text{for all } h \in \mathcal{H}.$$

Taking ρ expectation on both sides and recalling that $\mathbb{E}_\rho[\ln\frac{1}{\pi(h)\delta}] = \mathbb{E}_\rho[\ln\frac{1}{\pi(h)}] + \ln\frac{1}{\delta}$, we obtain

$$\mathbb{E}_\rho\Big[\mathrm{kl}\big(\hat{L}(h,S) \,\|\, L(h)\big)\Big] \;\le\; \frac{\mathbb{E}_\rho\big[\ln\frac{1}{\pi(h)}\big] + \ln\frac{1}{\delta}}{n}.$$

Combining this bound with (1) proves the required inequality for the fixed ρ. Because the probability statement of Theorem 3.38 holds uniformly over h, the same argument remains valid after taking a supremum over all posteriors ρ, yielding Theorem 3.40(or 3.43 in the new lecture note).

Part 2: Comparison with the PAC-Bayes-kl inequality (Theorem 3.26)

Common ground Both inequalities upper bound the binary KL divergence between the expected empirical loss and the expected true loss of a stochastic classifier ρ, with high probability over the sample S.

Complexity term

- **Occam's kl-razor (Theorm 3.40):** $\dfrac{\mathbb{E}_\rho\big[\ln\frac{1}{\pi(h)}\big]}{n}$.
- **PAC-Bayes-kl (Thm 3.26):** $\dfrac{\mathrm{KL}(\rho\|\pi)}{n} = \dfrac{\mathbb{E}_\rho\big[\ln\frac{\rho(h)}{\pi(h)}\big]}{n}$.

When the posterior ρ has high entropy so that $\rho(h) \ll 1$ for most h, $\mathrm{KL}(\rho\|\pi)$ may greatly exceed $\mathbb{E}_\rho[\ln\frac{1}{\pi(h)}]$, making Theorm 3.40 potentially tighter. Conversely, if ρ concentrates on a few hypotheses, the extra $\ln\rho(h)$ term can shrink $\mathrm{KL}(\rho\|\pi)$ below the Occam cost, so PAC-Bayes may win.

Prior coverage Theorm 3.40 requires only a countable $\mathcal{H}$ (because of the union bound in Theorem 3.41), whereas PAC-Bayes-kl works unchanged for any measurable $\mathcal{H}$ (including continuous parameter spaces).

λ-dependence Occam's bound carries $\ln\frac{1}{\delta}$, while PAC-Bayes-kl uses $\ln\frac{2\sqrt{n}}{\delta}$ (the factor $2\sqrt{n}$ being negligible for large n).

Advantages of PAC-Bayes-kl

- Applicable to continuous hypothesis classes.

- Produces a data dependent complexity measure $\mathrm{KL}(\rho\|\pi)$ that can adapt to how sharply the posterior concentrates.
- Enjoys many refined versions (localised priors, margin based losses,...).

Advantages of Occam's kl-razor (soft selection)

- Simpler complexity term, sometimes significantly smaller than $\mathrm{KL}(\rho\|\pi)$.
- No need to evaluate or approximate $\rho(h)$ inside the bound.
- Removes the extra $\sqrt{n}$ factor in the δ term.

Disadvantages of each

- **Occam:** restricted to countable $\mathcal{H}$; bound can be loose when ρ is highly concentrated.
- **PAC-Bayes:** requires computing $\mathrm{KL}(\rho\|\pi)$ (often intractable); the extra $\ln(2\sqrt{n})$ gives a slightly larger constant.

Take away Occam's soft selection inequality (Theorm 3.40) is a natural extension of the classical Occam bound to randomised classifiers and can outperform PAC-Bayes when the posterior remains diffuse. PAC-Bayes-kl, on the other hand, is more general and adaptive, making it the tool of choice for modern, high capacity or continuous hypothesis spaces.