DSC 190/291 · Assignment 4

UCSD · Spring 2026

Released: Friday, April 24 · Due: Friday, May 1, 11:59 PM

AI policy. AI assistance is allowed and encouraged in this course. You may use AI to learn the material, explore proof structure, test examples, debug code or formalizations, and improve exposition. However, you are responsible for checking correctness and for standing behind every proof step, derivation, formalization, experiment, and explanation you submit. Use AI as a collaborator, not as an oracle: do not submit anything you cannot explain and verify. The AI usage report is a required component of the assignment.

Submission. Submit a single PDF on Gradescope containing your write-up, figures, and discussion. Also place any supporting artifacts for the assignment in your course repository under the appropriate assignment directory. This may include code, Lean files, notebooks, scripts, data, or other materials needed to inspect or reproduce your work. Your submission should make it clear how the repository artifacts relate to the write-up.

Part A: Sparse Linear Predictors and Model Selection (40 points)

This problem studies structural risk minimization in a setting where the complexity parameter is sparsity. The goal is to connect non-uniform learning bounds with the computational cost of searching over sparse predictors.

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Use the convention $\text{sign}(z) = +1$ for $z \geq 0$ and $\text{sign}(z) = -1$ for $z < 0$. For $1 \leq k \leq d$, define the class of k-sparse homogeneous linear classifiers

$$\mathcal{H}_k = \{x \to \text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d, \|w\|_0 \leq k\}.$$

Let $\mathcal{H} = \bigcup_{k=1}^{d} \mathcal{H}_k$. Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$, and let $S = ((x_1, y_1), ..., (x_n, y_n)) \sim \mathcal{D}^n$. When a weight vector w is used as the argument of L_S or $L_{\mathcal{D}}$, it denotes the classifier $x \to \text{sign}(\langle w, x \rangle)$.

1. **VC dimension through supports.**

   Prove an upper bound of the form

   $$\text{VCdim}(\mathcal{H}_k) = O\left(k \log\left(e\frac{d}{k}\right)\right).$$

   Hint: write $\mathcal{H}_k$ as a union over the possible supports of w, then combine the growth-function bounds for those subclasses.

2. **Penalty-based SRM.**

   Choose a prior $p_k > 0$ over sparsity levels $k = 1, ..., d$ with $\sum_{k=1}^{d} p_k \leq 1$; for example, you may take $p_k \propto \frac{1}{k^2}$. Define an SRM rule of the form

   $$\hat{h} \in \arg\min_{1 \leq k \leq d, h \in \mathcal{H}_k} \{L_S(h) + \text{pen}(k, n, \delta)\}.$$

Derive a high-probability oracle inequality of the form

$$L_{\mathcal{D}}\big(\hat{h}\big) \leq \inf_{1\leq k\leq d, h\in\mathcal{H}_k} \left\{ L_{\mathcal{D}}(h) + C\sqrt{\frac{k\log\big(e\frac{d}{k}\big) + \log\Big(\frac{1}{p_k}\Big) + \log\big(\frac{1}{\delta}\big)}{n}} \right\}$$

for a universal constant C.

You may use the class-level SRM theorem from lecture, but you must instantiate every term in the bound and explain what statistical cost is paid for choosing the support and what cost is paid for choosing the sparsity level.

3. **Comparison with dense halfspaces.**

   Suppose there exists $w^* \in \mathbb{R}^d$ with $\|w^*\|_0 \leq s$ and $L_{\mathcal{D}}(w^*) \leq \eta$. Use the SRM bound above to state a sample size sufficient to guarantee

   $$L_{\mathcal{D}}\big(\hat{h}\big) \leq \eta + \varepsilon$$

   with probability at least $1-\delta$.

   Then compare this with the sample size obtained by learning over all homogeneous halfspaces in $\mathbb{R}^d$. In what regime is the sparse bound smaller?

4. **Validation as model selection.**

   Consider the following validation rule. Assume n is even, and split the sample into two equal parts S_1 and S_2. For each $k = 1, ..., d$, let

   $$w_k \in \arg\min_{\|w\|_0\leq k} L_{S_1}(w).$$

   Then choose the output $w_{\hat{k}}$ with

   $$\hat{k} \in \arg\min_{1\leq k\leq d} L_{S_2}(w_k).$$

   Prove a validation-style oracle bound of the form

   $$L_{\mathcal{D}}(w_{\hat{k}}) \leq \inf_{1\leq k\leq d, w:\|w\|_0\leq k} \left\{ L_{\mathcal{D}}(w) + C\sqrt{\frac{k\log\big(e\frac{d}{k}\big) + \log\big(\frac{d}{\delta}\big)}{n}} \right\}$$

   for a universal constant C.

   Hint: after conditioning on S_1, the validation step is a finite-class selection problem over the candidates $w_1, ..., w_d$.

   Then compare this validation rule to the penalty-based SRM rule above: what samples are used to fit predictors, what samples are used to choose the complexity level, and what statistical price is paid for that separation?

5. **Computation.**

   Now restrict attention to the realizable case. Suppose the sample is consistent with some k-sparse homogeneous halfspace. Describe the brute-force algorithm that enumerates supports $I \subseteq [d]$ with $|I| = k$ and solves a halfspace feasibility problem on each support. Estimate its runtime as a function of d, k, and n. In which regimes of k is this polynomial

in d? Explain how this illustrates the Week 4 distinction between sample complexity and computational complexity.

Part B: PAC-Bayes for Thresholds (45 points)

PAC-Bayes bounds use $\mathrm{KL}(Q \,\|\, P)$ as the complexity term. This problem compares that term with VC dimension for thresholds on a finite ordered domain.

Throughout, let

$$\mathcal{X}_N = \{1, 2, ..., N\}, \quad \mathcal{Y} = \{0, 1\},$$

and define the threshold class

$$\mathcal{H}_N = \{h_t : t \in \{1, ..., N+1\}\},$$

where

$$h_t(x) = \mathbf{1}[x \geq t].$$

Thus h_1 labels every point by 1, and h_{N+1} labels every point by 0.

Use the PAC-Bayes bound from lecture: For any distribution $\mathcal{D}$ over $\mathcal{X}_N \times \mathcal{Y}$ and any prior P, with probability at least $1 - \delta$ over $S = ((x_1, y_1), ..., (x_n, y_n)) \sim \mathcal{D}^n$, simultaneously for all posteriors Q,

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{\mathrm{KL}(Q \,\|\, P) + \log\left(2\frac{n}{\delta}\right)}{2(n-1)}}.$$

Here

$$L_{\mathcal{D}}(Q) = \mathbb{E}_{h \sim Q}[L_{\mathcal{D}}(h)], \quad L_S(Q) = \mathbb{E}_{h \sim Q}[L_S(h)].$$

We work in the realizable setting.

1. **VC dimension and point-posterior PAC-Bayes.**

   Prove that $\mathrm{VCdim}(\mathcal{H}_N) = 1$. Let P be the uniform prior over $\mathcal{H}_N$, let $A(S)$ be any deterministic ERM rule that returns a consistent threshold $h_{\hat{t}}$, and define $Q_S = \delta_{h_{\hat{t}}}$. Compute $\mathrm{KL}(Q_S \,\|\, P)$, plug it into the PAC-Bayes theorem, and write the resulting bound on $L_{\mathcal{D}}(Q_S)$. Finally, explain in two or three sentences why this PAC-Bayes certificate scales with $\log(N+1)$ even though a VC-style realizable guarantee for thresholds should not scale with $\log N$.

2. **Version-space posterior.**

   Define $a(S) = \max\{x_i : y_i = 0\}$, with $a(S) = 0$ if there are no negative examples, and define $b(S) = \min\{x_i : y_i = 1\}$, with $b(S) = N + 1$ if there are no positive examples. Prove that the set of thresholds consistent with S is

   $$V(S) = \{t : a(S) < t \leq b(S)\}$$

   and hence $|V(S)| = b(S) - a(S)$.

Let Q_V be the uniform posterior over $\{h_t : t \in V(S)\}$. For the uniform prior P over $\mathcal{H}_N$, compute $\mathrm{KL}(Q_V \,\|\, P)$ exactly. Prove that $L_S(Q_V) = 0$, and explain why Q_V can give a better PAC-Bayes bound than a point posterior when $|V(S)|$ is large.

3. **Spreading helps KL, but can hurt true risk.**

   Let P be the uniform prior over $\mathcal{H}_N$. Let $\mathcal{D}$ be the realizable distribution whose marginal on $\mathcal{X}_N$ is uniform and whose labels are generated by h_τ. For any nonempty set $W \subseteq \{1, ..., N+1\}$, let Q_W be uniform over $\{h_t : t \in W\}$. Prove that

   $$L_{\mathcal{D}}(Q_W) = \frac{1}{N|W|} \sum_{t \in W} |t - \tau|.$$

   Construct a concrete example with $N \geq 20$, a true threshold τ, and a realizable sample S such that $|V(S)|$ is large,

   $$\mathrm{KL}\big(Q_{V(S)} \,\|\, P\big) < \mathrm{KL}\big(\delta_{h_\tau} \,\|\, P\big),$$

   but

   $$L_{\mathcal{D}}\big(Q_{V(S)}\big) > L_{\mathcal{D}}\big(\delta_{h_\tau}\big).$$

   Explain why this does not contradict PAC-Bayes.

4. **Every fixed prior can be attacked.**

   Let P be any prior over $\mathcal{H}_N$, fixed before seeing the sample. Assume $N \geq 3$.

   Prove that there exists an internal threshold $\tau \in \{2, ..., N\}$ such that

   $$P(h_\tau) \leq \frac{1}{N-1}.$$

   For this τ, define a realizable distribution $\mathcal{D}_\tau$ by

   $$\mathbb{P}_{(x,y)\sim\mathcal{D}_\tau}[x = \tau - 1, y = 0] = \frac{1}{2}, \quad \mathbb{P}_{(x,y)\sim\mathcal{D}_\tau}[x = \tau, y = 1] = \frac{1}{2}.$$

   Prove that with probability at least $1 - 2^{1-n}$ over $S \sim \mathcal{D}_\tau^n$, the sample contains both support points.

   On this event, prove that $V(S) = \{\tau\}$. Conclude that any posterior Q with $L_S(Q) = 0$ must be $Q = \delta_{h_\tau}$. Prove that, on this event,

   $$\mathrm{KL}(Q \,\|\, P) \geq \log(N-1)$$

   for every zero-empirical-error posterior Q.

   Explain what this lower bound does and does not show. In particular, why is it a limitation of zero-empirical-error PAC-Bayes certificates with a fixed prior, rather than a lower bound on the sample complexity of learning thresholds?

Part C: AI Usage Report (15 points)

Write a short report describing how you used AI in this assignment. Do not just list tools; explain what role AI played in your work and how you checked the result. Address:

1. Describe the parts of the assignment for which you used AI. For example: exploring examples, proposing conjectures, checking algebra, debugging code or formalizations, or improving exposition.
2. Describe concrete AI suggestions you accepted and explain why.
3. Describe concrete AI suggestions you rejected or substantially modified, and explain what was wrong, incomplete, or unhelpful about them.
4. Describe how you verified the correctness of what you submitted. Be specific about the relevant kind of work in this assignment: proof, derivation, code, experiment, or exposition.

AI workflow. Also describe concrete updates to your AI workflow that resulted from this assignment. This may include changes to `CLAUDE.md`, `AGENTS.md`, prompts, checklists, scripts, or skills. **Explain the 5 most recent changes you made to your AI workflow and why**.

If you did not use AI for some part of the assignment, say so explicitly.