DSC 190/291 · Assignment 6 Solutions

UCSD · Spring 2026

Part A: Boosting Sparse Linear Predictors and the ℓ_1 Margin

(45 points)

This problem continues the sparse-linear-model theme from Homeworks 4 and 5. Let $\varphi : \mathcal{X} \to \{-1,+1\}^d$ be a fixed feature map. For each $j \in [d]$ and $\sigma \in \{-1,+1\}$, define $b_{j,\sigma}(x) = \sigma\varphi_j(x)$, and let $\mathcal{B} = \{b_{j,\sigma} : j \in [d], \sigma \in \{-1,+1\}\}$. For a $\{-1,+1\}$-valued predictor b, its edge under a distribution Q is $\frac{1}{2} - L_Q(b) = \mathbb{E}_Q \frac{[yb(x)]}{2}$.

Assume $\mathcal{D}$ is realizable with margin by an s-sparse predictor: there is $w^\star \in \mathbb{R}^d$ with $\|w^\star\|_0 \le s$ and $y\langle w^\star, \varphi(x)\rangle \ge 1$ for every (x,y) in the support of $\mathcal{D}$. Equivalently, $\frac{w^\star}{\|w^\star\|_1}$ has ℓ_1-normalized margin at least $\frac{1}{\|w^\star\|_1}$. When a finite sample is used, write $S = ((x_1,y_1),...,(x_n,y_n))$.

You may use without proof that sparse linear classifiers have VC dimension $O\big(s\log(e\frac{d}{s})\big)$, so exact sparse ERM has realizable sample complexity $O\Big(\frac{s\log(e\frac{d}{s})+\log(\frac{1}{\delta})}{\varepsilon}\Big)$ up to logarithmic factors, but is computationally difficult when s is part of the input.

1. **(12 points) From sparse margin to a weak coordinate.**

   Assume additionally that $\|w^\star\|_\infty \le B$. For any distribution Q supported on examples satisfying the margin condition, prove that some $b \in \mathcal{B}$ has $L_Q(b) \le \frac{1}{2} - \frac{1}{2\|w^\star\|_1}$, and hence $L_Q(b) \le \frac{1}{2} - \frac{1}{2sB}$.

   Then describe an $O(nd)$ weighted ERM weak learner for $\mathcal{B}$ on weighted examples $(x_i, y_i, D_i)_{i=1}^n$. Hint: relate weighted error to the signed correlation $\sum_i D_i y_i b(x_i)$.

   Solution.

   Let Q be supported on examples satisfying

   $$y\langle w^\star, \varphi(x)\rangle \ge 1.$$

   Taking expectation under Q gives

   $$1 \le \mathbb{E}_Q[y\langle w^\star, \varphi(x)\rangle] = \sum_{j=1}^d w_j^\star \mathbb{E}_Q\big[y\varphi_j(x)\big].$$

   Write

   $$a_j = \mathbb{E}_Q\big[y\varphi_j(x)\big].$$

   Then

   $$1 \le \sum_{j=1}^d w_j^\star a_j \le \sum_{j=1}^d |w_j^\star| \max_l |a_l| = \|w^\star\|_1 \max_l |a_l|.$$

Hence some coordinate j satisfies

$$|a_j| \geq \frac{1}{\|w^\star\|_1}.$$

Choose $\sigma = +1$ if $a_j \geq 0$ and $\sigma = -1$ otherwise, and set

$$b(x) = \sigma\varphi_j(x).$$

Then

$$\mathbb{E}_Q[yb(x)] = |a_j| \geq \frac{1}{\|w^\star\|_1}.$$

Since $yb(x) \in \{-1, +1\}$,

$$\mathbb{E}_Q[yb(x)] = 1 - 2L_Q(b).$$

Therefore

$$L_Q(b) \leq \frac{1}{2} - \frac{1}{2\|w^\star\|_1}.$$

If $\|w^\star\|_\infty \leq B$ and $\|w^\star\|_0 \leq s$, then

$$\|w^\star\|_1 \leq sB,$$

so

$$L_Q(b) \leq \frac{1}{2} - \frac{1}{2sB}.$$

For weighted ERM over $\mathcal{B}$, compute

$$c_j = \sum_{i=1}^{n} D_i y_i \varphi_j(x_i)$$

for every coordinate j. For the predictor $b_{j,\sigma}$,

$$\sum_{i=1}^{n} D_i y_i b_{j,\sigma}(x_i) = \sigma c_j,$$

and, since $\sum_i D_i = 1$, its weighted error is

$$\frac{1}{2}(1 - \sigma c_j).$$

Thus the best sign for coordinate j is the sign of c_j, and the best coordinate is one maximizing $|c_j|$. Computing all c_j takes one scan over the n examples and d coordinates, hence $O(nd)$ time.

2. **(14 points) Boosting guarantee and comparison with sparse ERM.**

   Run AdaBoost over $\mathcal{B}$. At every round, the reweighted distribution is supported on the same margin-realizable sample, so the previous part applies with $\gamma = \frac{1}{2sB}$. Prove the exponential training-error bound, choose T so that the training error is zero, and then use the sparse-

linear VC bound, applied with the number of activated coordinates in place of s, to obtain a realizable generalization guarantee.

State the resulting sample complexity up to logarithmic factors, and compare it with exact sparse ERM. Your comparison should identify the statistical price paid by boosting, the computational advantage, and the role of B.

Solution.

On AdaBoost round t, the current distribution D^t is supported on the same sample S. Since every point in S satisfies the margin condition, the previous part gives a weak rule with edge at least

$$\gamma = \frac{1}{2sB}.$$

Equivalently, if $\varepsilon_t = L_{D^t}(h_t)$, then

$$\varepsilon_t \le \frac{1}{2} - \gamma.$$

The standard AdaBoost potential argument gives

$$L_S^{0/1}(H_T) \le L_S^{\exp}(f_T) \le \prod_{t=1}^{T} \sqrt{1-4\gamma^2} \le \exp(-2\gamma^2 T).$$

The empirical 0-1 error is a multiple of $\frac{1}{n}$, so it is zero whenever

$$\exp(-2\gamma^2 T) < \frac{1}{n}.$$

It is enough to take

$$T > \log \frac{n}{2\gamma^2} = 2s^2 B^2 \log n.$$

The final classifier has the form

$$H_T(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t \sigma_t \varphi_{j_t}(x) \right).$$

After combining repeated coordinates, this is a sparse linear classifier over the original d features with support size at most

$$r = \min(T, d).$$

The realizable VC bound for r-sparse linear classifiers gives, up to logarithmic factors,

$$L_{\mathcal{D}}(H_T) \le O\left(\frac{r \log\left(e\frac{d}{r}\right) + \log\left(\frac{1}{\delta}\right)}{n} \right)$$

for a consistent output. Substituting

$$T = O(s^2 B^2 \log n)$$

and hiding the extra logarithmic dependence on n gives the sample complexity

$$n = \tilde{O}\left(\frac{s^2 B^2 \log d + \log\left(\frac{1}{\delta}\right)}{\varepsilon}\right).$$

Exact sparse ERM has realizable sample complexity

$$\tilde{O}\left(\frac{s \log\left(e\frac{d}{s}\right) + \log\left(\frac{1}{\delta}\right)}{\varepsilon}\right),$$

but it requires searching over sparse supports when s is part of the input. Boosting replaces that search by T coordinate scans, each costing $O(nd)$. The statistical price is that the boosted classifier may activate about $O(s^2 B^2 \log n)$ coordinates rather than s coordinates. The coefficient bound enters because the weak edge is controlled by $\frac{1}{\|w^\star\|_1}$; the assumptions $\|w^\star\|_0 \le s$ and $\|w^\star\|_\infty \le B$ imply the polynomial edge $\frac{1}{2sB}$.

3. **(9 points) Why the coefficient bound matters.**

   For odd $s = 2m + 1$, set $d = s$. Construct a distribution supported on s labeled examples with $y = +1$ and $\varphi(x) \in \{-1, +1\}^s$ such that some $w^\star \in \mathbb{R}^s$ has margin 1, but every coordinate predictor has edge at most $2^{-\Omega(s)}$. Also show that the realizing vector in your construction satisfies $\|w^\star\|_\infty = 2^{\Omega(s)}$.

   Hint: index rows by $0, (1,+), (1,-), ..., (m,+), (m,-)$ with weights $q_0 = 1$ and $q_{r,+} = q_{r,-} = 2^{r-1}$. It suffices to build an invertible $s \times s$ sign matrix whose every column has q-weighted sum 1; try a base column with $(+,-)$ on every pair, columns that flip one pair, and carry columns with $(-,-)$ on earlier pairs, $(+,+)$ on one pair, and $(+,-)$ later. Prove realizability, compute the best coordinate edge, and explain why this rules out any polynomial-in-s AdaBoost guarantee based only on sparsity.

Solution.

Let the rows be indexed by

$$0, (1,+), (1,-), ..., (m,+), (m,-).$$

Define weights

$$q_0 = 1, \quad q_{r,+} = q_{r,-} = 2^{r-1}, \quad Q = \sum_i q_i = 2^{m+1} - 1.$$

We construct an $s \times s$ sign matrix A with columns

$$c_0, c_1, ..., c_m, d_1, ..., d_m.$$

The column c_0 has value $+1$ on row 0 and values $(+1, -1)$ on every pair $(r,+), (r,-)$. For $r \ge 1$, the column c_r is the same as c_0 except that pair r is flipped to $(-1, +1)$. The column d_r has value -1 on row 0; on pairs $\ell < r$ it has $(-1,-1)$; on pair r it has $(+1,+1)$; and on pairs $\ell > r$ it has $(+1,-1)$.

Every column has q-weighted sum 1. For c_0 and the c_r columns, every pair cancels and row 0 contributes 1. For d_r, the later pairs cancel and the weighted sum is

$$-1 + 2 \cdot 2^{r-1} - 2\sum_{\ell<r} 2^{\ell-1} = 1.$$

The columns are linearly independent. Indeed, $c_r - c_0$ is supported only on pair r and spans the pair-difference direction for that pair. After quotienting by the span of c_0 and all these pair-difference directions, the images of $d_1, ..., d_m$ in the pair-sum coordinates are triangular with nonzero diagonal. Hence A is invertible.

Let the s support points be the rows of A, let all labels be $y = +1$, and put probability $\frac{q_i}{Q}$ on row i. Since A is invertible, there is a unique vector $w^\star$ satisfying

$$Aw^\star = (1, ..., 1).$$

Therefore every support point has margin exactly 1.

For every coordinate j,

$$\mathbb{E}\big[y\varphi_j(x)\big] = \frac{1}{Q}\sum_i q_i A_{i,j} = \frac{1}{Q},$$

because every column has q-weighted sum 1. Thus the best coordinate predictor has edge

$$\frac{1}{2Q} = 2^{-\Theta(s)},$$

and no coordinate predictor has a larger edge.

Finally, taking expectation in the margin identity gives

$$1 = \mathbb{E}[y\langle w^\star, \varphi(x)\rangle] = \sum_j w_j^\star \mathbb{E}\big[y\varphi_j(x)\big] = \frac{1}{Q}\sum_j w_j^\star.$$

Hence

$$\sum_j w_j^\star = Q,$$

so

$$\|w^\star\|_\infty \geq \frac{Q}{s} = 2^{\Omega(s)}.$$

The standard AdaBoost training-error guarantee based on this coordinate edge certifies zero training error after

$$T = \Theta(\gamma^{-2}\log n) = 2^{\Omega(s)}\log n$$

rounds. Thus sparsity alone cannot give a polynomial-in-s coordinate-AdaBoost guarantee; the coefficient or ℓ_1-margin scale matters.

4. **(10 points) Experiment: sparse boosting versus a convex surrogate.**

   Implement AdaBoost over the coordinate class $\mathcal{B}$. Compare an easy sparse-margin distribution of your choice with the construction from the previous part. Report training error, exponential loss, observed edge, support size, and normalized margin over rounds. Compare

AdaBoost with one convex surrogate method, for example logistic regression or hinge-loss minimization with an ℓ_1 constraint or penalty. Explain what the experiment illustrates about support sparsity, coefficient size, ℓ_1 margin, and computational tractability.

Solution.

One possible experiment is the following.

For the easy distribution, choose a small support $J \subset [d]$, take a vector $w^\star$ with moderate coefficients on J, draw $\varphi(x) \in \{-1, +1\}^d$, and set

$$y = \operatorname{sign}(\langle w^\star, \varphi(x) \rangle)$$

while rejecting or resampling points with margin below 1. For the hard distribution, use the construction in the previous part, sampling rows according to the weights $\frac{q_i}{Q}$.

Run AdaBoost over the coordinate class by computing the weighted correlations

$$c_j^t = \sum_i D_i^t y_i \varphi_j(x_i)$$

and selecting the coordinate and sign with largest absolute correlation. Track over rounds:

- training error of H_T;
- exponential loss $L_S^{\exp}(f_T)$;
- observed edge $\frac{1}{2} - L_{D^t}(h_t)$;
- number of distinct activated coordinates;
- normalized margin, for example

$$\min_i \frac{y_i f_T(x_i)}{\sum_{t=1}^T |\alpha_t|}.$$

On the easy distribution, the observed edge should remain visibly bounded away from 0 for enough rounds to drive training error down quickly. On the construction above, the initial best coordinate edge is about $\frac{1}{2Q}$ when the empirical sample reflects the row weights, so the AdaBoost progress predicted by the worst-case bound is extremely slow.

A convex surrogate comparison can use logistic regression with an ℓ_1 penalty, or hinge-loss minimization with an ℓ_1 constraint. Such methods optimize a tractable convex objective over an ℓ_1-controlled predictor, but they do not directly optimize coordinate-wise 0-1 agreement and need not return a proper s-sparse classifier. The experiment illustrates that support sparsity, coefficient size, ℓ_1 margin, and computational tractability are distinct: a classifier may be sparse but have exponentially large coefficients, and then coordinate boosting has only an exponentially small guaranteed edge.

Part B: Agnostic Halfspace Hardness via Boosting (40 points)

Part A used boosting constructively; here boosting is a reduction tool.

Let $\mathcal{X}_d = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Let $\mathcal{H}_d$ be the class of affine halfspaces $h_{w,b}(x) = \text{sign}(\langle w, x\rangle + b)$, with $\text{sign}(z) = +1$ for $z \geq 0$. Let $\mathcal{I}_{d,k}$ be the class of intersections of k halfspaces: the output is $+1$ iff all k halfspaces output $+1$. For a size function $k = k(d)$, polynomial-size means $k(d) \leq d^c$ for some fixed constant c, and $k(d) = \omega(1)$ means $k(d) \to \infty$.

Use these black-box hardness facts: under standard uSVP hardness, intersections of d^r affine halfspaces are not efficiently PAC learnable in the realizable case, even improperly, for every fixed $r > 0$; under RSAT, the same is true for intersections of $k(d) = \omega(1)$ affine halfspaces.

1. **(10 points) A weak halfspace inside an intersection.**

   Prove: if $\mathcal{D}$ is realizable by some $g \in \mathcal{I}_{d,k}$, then some affine halfspace has error at most $\frac{1}{2} - \frac{1}{2k^2}$.

   Hint: split on $p = \mathbb{P}[y = +1]$. For small p, use the constant -1 halfspace; for large p, average over the k halfspaces defining the realizing intersection.

Solution.

Let

$$g = \text{AND}(h_1, ..., h_k)$$

realize $\mathcal{D}$, and let

$$p = \mathbb{P}[y = +1].$$

If

$$p \leq \frac{1}{2} - \frac{1}{2k^2},$$

then the constant -1 classifier, which is an affine halfspace, has error p and satisfies the desired bound.

Now suppose

$$p > \frac{1}{2} - \frac{1}{2k^2}.$$

If $k = 1$, then h_1 itself realizes the distribution and has error 0, so assume $k \geq 2$. Every positive example is correctly classified by all h_j, because an intersection outputs $+1$ only when all constituent halfspaces output $+1$. Every negative example is correctly classified by at least one h_j, because the intersection outputs -1 only when at least one constituent halfspace outputs -1. Thus, over the k constituent halfspaces, a positive example contributes no errors and a negative example contributes at most $k - 1$ errors. Therefore

$$\sum_{j=1}^{k} L_{\mathcal{D}}(h_j) \leq (k-1)(1-p),$$

so some h_j has error at most

$$(k-1)\frac{1-p}{k}.$$

Since

$$1 - p < \frac{1}{2} + \frac{1}{2k^2},$$

we have

$$(k-1)\frac{1-p}{k} < (k-1)\frac{k^2+1}{2k^3} \le \frac{k^2-1}{2k^2} = \frac{1}{2} - \frac{1}{2k^2},$$

where the middle inequality is equivalent to $(k-1)^2 \ge 0$. Hence some affine halfspace has error at most $\frac{1}{2} - \frac{1}{2k^2}$.

2. **(10 points) From an agnostic learner to a weak learner.**

   Suppose, hypothetically, that affine halfspaces are efficiently properly agnostically PAC learnable. Use the previous lemma to construct a weak learner for $\mathcal{J}_{d,k}$ in the realizable case. Give γ such that the returned halfspace has error at most $\frac{1}{2} - \gamma$, and explain why the weak learner is polynomial-time when $k(d) \le d^c$ for a fixed constant c.

   Solution.

   Assume there is an efficient proper agnostic learner A for affine halfspaces. On any distribution realizable by $\mathcal{J}_{d,k}$, the previous lemma gives

   $$\inf_{h \in \mathcal{H}_d} L_{\mathcal{D}}(h) \le \frac{1}{2} - \frac{1}{2k^2}.$$

   Run A with excess accuracy

   $$\varepsilon_A = \frac{1}{4k^2}$$

   and any fixed confidence, say $\frac{2}{3}$. With probability at least $\frac{2}{3}$, the returned affine halfspace has error at most

   $$\frac{1}{2} - \frac{1}{2k^2} + \frac{1}{4k^2} = \frac{1}{2} - \frac{1}{4k^2}.$$

   Thus this is a weak learner for $\mathcal{J}_{d,k}$ with edge

   $$\gamma = \frac{1}{4k^2}.$$

   If $k(d) \le d^c$ for a fixed constant c, then $\frac{1}{\varepsilon_A} = 4k(d)^2 \le 4d^{2c}$. Since the assumed agnostic learner is efficient, its sample size and runtime are polynomial in d, $\frac{1}{\varepsilon_A}$, and the confidence parameter. Therefore the resulting weak learner runs in polynomial time for polynomial-size intersections.

3. **(12 points) Boosting the weak learner.**

   Use AdaBoost and the boosted-halfspace VC bound $\tilde{O}(Td)$ to obtain a realizable learner for $\mathcal{J}_{d,k}$. At every round, the weighted distribution is supported on examples still realized by the same intersection. State the sample and runtime dependence up to logarithmic factors.

Solution.

Use the weak learner from the previous part with

$$\gamma = \frac{1}{4k^2}.$$

AdaBoost gives

$$L_S^{0/1}(H_T) \le \exp(-2\gamma^2 T).$$

To make the empirical error zero on a sample of size n, it is enough to take

$$T > \log \frac{n}{2\gamma^2} = O(k^4 \log n).$$

On each round, the distribution D^t is supported on the training sample. Since the original sample is realized by the same intersection, each reweighted distribution is also realized by that intersection. Thus the weak-learning guarantee applies at every round. Running each weak call with failure probability, for example, $\frac{\delta}{2T}$ and union bounding over rounds makes all calls succeed with probability at least $1 - \frac{\delta}{2}$.

The final classifier is a weighted vote of T affine halfspaces. The VC dimension of T-term boosted halfspace votes is $\tilde{O}(Td)$. Therefore, for a consistent boosted classifier, the realizable VC bound gives population error at most ε once

$$n = \tilde{O}\left(\frac{Td + \log\left(\frac{1}{\delta}\right)}{\varepsilon}\right).$$

Substituting

$$T = O(k^4 \log n)$$

gives

$$n = \tilde{O}\left(\frac{dk^4 + \log\left(\frac{1}{\delta}\right)}{\varepsilon}\right).$$

The runtime is T calls to the assumed agnostic halfspace learner, plus the usual AdaBoost bookkeeping. Up to logarithmic factors, this is polynomial in d, k, $\frac{1}{\varepsilon}$, and $\log\left(\frac{1}{\delta}\right)$ whenever the agnostic halfspace learner is polynomial-time. In particular, if $k(d) \le d^c$ for a fixed constant c, the boosted learner is polynomial-time in d, $\frac{1}{\varepsilon}$, and $\log\left(\frac{1}{\delta}\right)$.

4. **(8 points) Consequence for agnostic halfspaces.**

   Prove the implication: if affine halfspaces are efficiently properly agnostically PAC learnable, then intersections of $k(d) \le d^c$ affine halfspaces are efficiently learnable in the realizable case, for every fixed c.

   Explain the contradiction with the black-box hardness facts by taking $k(d) = d^r$ under uSVP, or any polynomially bounded $k(d) = \omega(1)$ under RSAT. Conclude the implication for efficient proper agnostic PAC learning of halfspaces.

Solution.

Combining the previous two parts, an efficient proper agnostic learner for affine halfspaces would give an efficient weak learner for intersections of k halfspaces, and AdaBoost would convert that weak learner into an efficient realizable learner. Therefore, for every fixed c, intersections of $k(d) \leq d^c$ affine halfspaces would be efficiently PAC learnable in the realizable case.

Under standard uSVP hardness, this contradicts the stated hardness of learning intersections of d^r affine halfspaces for every fixed $r > 0$. Under RSAT, it contradicts the stated hardness for every polynomially bounded $k(d) = \omega(1)$. Hence, under these assumptions, affine halfspaces are not efficiently properly agnostically PAC learnable.