ACTIVEHEDGE: HEDGE MEETS ACTIVE LEARNING

{Bhuvesh Kumar, Jacob Abernethy} @ Georgia Tech, Venkatesh Saligrama @ Boston University

SUMMARY

We consider the classical problem of multiclass prediction with expert advice, but with an active learning twist. The learner aims to minimize regret while querying the labels of only a small number of examples; the learner is also allowed a very short *burn-in* phase where it can fast-forward and query certain highly-informative examples. We design ActiveHedge, that utilizes Hedge as a subroutine, and show that under a very particular combinatorial constraint (which we refer as ζ Compactness) on the matrix of expert predictions, we can obtain a very strong regret guarantee while querying very few labels.

PREDICTION WITH EXPERT ADVICE

Sequential classification of M points into K classes using advice from N experts

For $t = 1, \cdots, M$:

1. Receive advice $X_t \in [K]^N$
2. Predict label $\hat{y}_t \in \Delta_K$
3. Query true label y_t
4. Suffer loss $\ell(\hat{y}_t, y_t) := \frac{1}{2}\|\hat{y}_t - y_t\|_1$

Regret

$$\text{REG}_{\texttt{alg}} := \sum_{t=1}^{M} \ell(\hat{y}_t, y_t) - \min_{j\in[N]} \sum_{t=1}^{M} \ell(X_{t,j}, y_t).$$

Hedge

In round t, Hedge weigh's each expert j's advice by $w_{t,j}$,

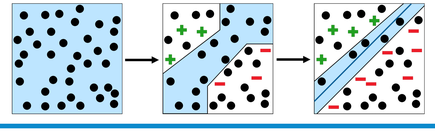
$$w_{t,j} \propto \exp\left(-\eta \sum_{t'=1}^{t} \ell(X_{t',j}, y_{t'})\right)$$

Theorem 2 (Freund and Schapire [1995]). *Give L^* such that $\min_{j\in[N]} \sum_{t=1}^{M} \ell(X_{t,j}, y_t) \le L^*$, then, choosing $\eta = \log\left(1 + \sqrt{\frac{2\ln N}{L^*}}\right)$*

$$\text{REG}_{\textit{Hedge}} \le \sqrt{2L^* \ln N} + \ln N$$

ACTIVE LEARNING

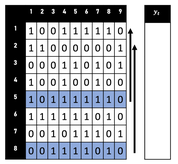
Unlike supervised learning, the learner starts with unlabeled pool of actively chooses and requests labels for the most informative points



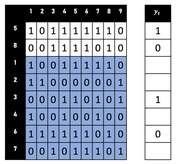
ACTIVE ONLINE LEARNING

Active twist to the learning with expert. It's the Hedge setting, but with three key modifications:

1. Expert predictions, $\mathbf{X}$ is known
2. The learner aims to make only a small number of label queries, limiting the number of times y_t is observed.
3. We allow a very brief *burn-in* phase, where the learner can *fast-forward* to act on particular examples , and query their labels, out of turn.



Burn-in Phase



Second Phase

ζ- COMPACTNESS

Measures the active learnability of expert prediction matrix $\mathbf{X}$

For any subset $V \subseteq [N]$ of experts, the *points of contention* of V is

$$\text{POC}_{\mathbf{X}}(V) := \{i \in [M] \mid \exists j, j' \in V : X_{i,j} \neq X_{i,j'}\}$$

Definition 1 (ζ- Compactness). *An expert prediction matrix $\mathbf{X}$ is ζ-compact if for all $V \subset [N]$ with $|V| \ge 2$,*

$$\frac{|\text{POC}_{\mathbf{X}}(V)|}{\max_{j,j'\in V} |\text{POC}_{\mathbf{X}}(\{j,j'\})|} \le \zeta$$

	1	2	3	4	5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	0
4	1	1	0	0	0
5	1	0	0	0	0

Large set of contentions imply large pairwise contentions also, more actively learnable, **small** ζ

	1	2	3	4	5
1	0	0	0	0	1
2	0	0	0	1	0
3	0	0	1	0	0
4	0	1	0	0	0
5	1	0	0	0	0

Large set of contentions made by similar experts, less actively learnable, **larger** ζ

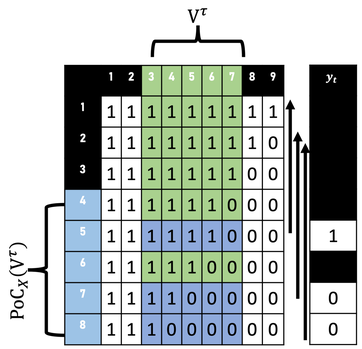
Usually ζ is a small constant

Theorem 3 (Informal). *There is a poly time algorithm to calculate the compactness ζ of a matrix up to an approximation factor of 3.*

ACTIVEHEDGE

Burn-in Phase

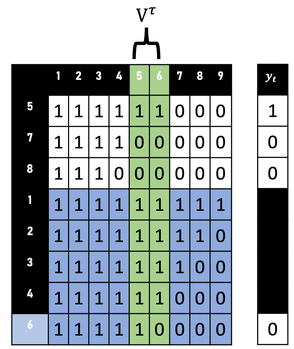
- Maintain a set of candidate experts V^τ
- Select k points from $\text{POC}_{\mathbf{X}}(V)$
- Predict labels using Hedge and request labels
- Shrink V^τ
- Repeat T times



Burn in Phase: Actively selecting informative points

Second Phase

- Sequentially go through remaining points
- Predict using Hedge and request label if point in $\text{POC}_{\mathbf{X}}(V)$
- Otherwise predict using any expert in V^τ



Burn in Phase: Actively selecting informative points

REGRET AND LABEL COMPLEXITY

Theorem 1. *Given a ζ-compact matrix $\mathbf{X}$ such that $\min_{j\in[N]} \sum_{t=1}^{M} \ell(X_{t,j}, y_t) \le L^* = \epsilon M$, with probability at least $1 - \rho$, for* ActiveHedge

1. *the number labels queried is no more than*

$$O\left(\zeta \log\left(\frac{N}{\rho} \log \frac{1}{10\epsilon\zeta}\right) \log \frac{1}{10\epsilon\zeta} + \epsilon\zeta M\right)$$

2. *the length of Phase I is no more than $\tilde{O}(\zeta)$ rounds;*
3. *and finally we have that*

$$\text{REG}_{\textit{ActiveHedge}} \le \sqrt{2\epsilon M \ln N} + \ln N$$

Same regret as Hedge using only $\tilde{O}(\zeta L^*)$ labels (as compared to M labels for Hedge)

REFERENCES

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

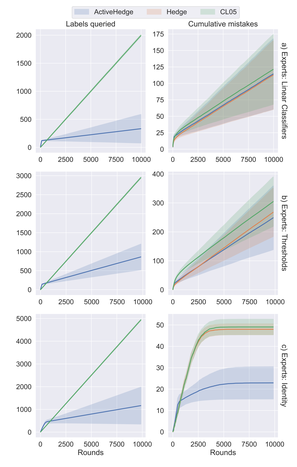
Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

PRELIMINARY EXPERIMENTS

We consider three different classes of experts for our experiments.

- In a) we consider linear classifiers passing through the origin as experts
- In b) we consider multidimensional experts as thresholds
- in c) we consider identity like matrix

We compare ActiveHedge with Hedge (Freund and Schapire [1995]) and the label efficient learner of [Cesa-Bianchi et al., 2005] (referred as CL05)



Labels queried and the cumulative mistakes of ActiveHedge, Hedge, and Cesa-Bianchi et al. [2005](CL05) in 3 different settings