Machine Learning Theory. Lecture 12.

Wouter M. Koolen

- Mixability
  - Prediction with Expert Advice Protocol
  - How to actually combine predictions?
- AdaBoost

  Following book chapter 10

Prediction with Expert Advice

Prediction with Expert Advice

Prediction with *expert advice* for loss function $\mathcal{L} : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$:

Protocol: for $t = 1, 2, \ldots$

- Experts announce actions $a_t^1, \ldots, a_t^K \in \mathcal{A}$.
- Learner chooses an action $a_t \in \mathcal{A}$.
- Adversary reveals outcome $x_t \in \mathcal{X}$.
- Learner incurs loss $\mathcal{L}(a_t, x_t)$.

Goal is small regret w.r.t. best expert:

$$R_T \;=\; \sum_{t=1}^{T} \mathcal{L}(a_t, x_t) - \min_k \sum_{t=1}^{T} \mathcal{L}(a_t^k, x_t)$$

Hedge (ignores advice) gives $R_T \leq O(\sqrt{T})$. Can get $\ln K$ for mix loss.
Q: When can we do better?

Mixable loss: Reduction to mix loss

Crux: exp-concavity is convenient but *too strong*.

Definition: We say $\mathcal{L}(a, x)$ is η-*mixable* if

$$\forall P \in \triangle_{\mathcal{A}} \exists a_P \in \mathcal{A} \forall x \in \mathcal{X} \quad \mathcal{L}(a_P, x) \ \le \ \frac{-1}{\eta} \ln \left(\mathop{\mathbb{E}}_{a \sim P} e^{-\eta \mathcal{L}(a,x)} \right)$$

Mapping from P to witness a_P called *substitution function*.

Sufficient to check mixability for all *binary support* P.

Note: mixability is reparametrisation invariant (exp-concavity not so).

Exp-concavity: mixable with the mean as the substitution function.

Mixable losses regret bound

Mixable losses behave just enough like the mix loss to carry the AA regret bound through.

Theorem 1. *For any η-mixable loss $\mathcal{L}$ there is an algorithm guaranteeing*

$$R_T \;\le\; \frac{\ln K}{\eta}$$

Proof sketch. Run the AA with losses $\ell_t^k = \eta\mathcal{L}(a_t^k, x_t)$. Given weights $\boldsymbol{w}_t$, construct distribution P_t with $P_t(a) = \sum_{k:a_t^k=a} w_t^k$. Then play $a_t = a_{P_t}$. Then apply the bound for AA, obtaining a $\ln K/\eta$ mix loss regret bound. The actual loss incurred is smaller (by mixability). □

Interestingly, Vovk (JCSS 1998) shows the converse.

Square loss is mixable

Let's consider

$$\mathcal{L}(a, x) = (a - x)^2$$

where $\mathcal{A} = \mathcal{X} = [-1, +1]$.

Square loss is mixable with $\eta = \frac{1}{2}$ (exp-concave only with $\eta = 1/8$).

The substitution function is

$$\boldsymbol{w}, a^1, \ldots, a^K \;\mapsto\; \frac{m_{\frac{1}{2}}(-1) - m_{\frac{1}{2}}(+1)}{4}$$

where $m_\eta(x) \;=\; \frac{-1}{\eta} \ln \sum_{k=1}^{K} w^k e^{-\eta(a^k - x)^2}$

See (Vovk 1990, Haussler, Kivinen, Warmuth, 1998)

For $\mathcal{A} = \mathcal{X} = [-Y, +Y]$ the constant becomes $\eta = \frac{1}{2Y^2}$.

Mixable loss list

Popular mixable losses:

- mix loss, log loss, entropic loss
- square loss, Brier loss
- Hellinger loss $\mathcal{A} = \mathcal{X} = [0, 1]$:

$$\mathcal{L}(a, x) \ := \ \frac{1}{2}\Big((\sqrt{1-x} - \sqrt{1-a})^2 + (\sqrt{x} - \sqrt{a})\Big)$$

Characterisation of mixability: (Van Erven, Reid, Williamson 2012).