Statistical Learning Theory

Machine learning: using algorithmic means to become more successful at a given task in a fixed random environment on the basis of past experience.

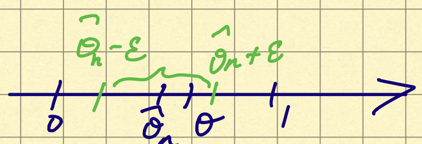
Example/illustration: coin tossing

- biased coin, θ (prob. of HEADS) unknown
- goal (to be able to make predictions on outcomes of tosses): "learn" θ

$$X_1, X_2, \dots, X_n \overset{iid}{\sim} \mathrm{Bern}(\theta)$$

$$\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad X^n := (X_1, \dots, X_n)$$

Fix $\varepsilon \in (0,1)$ [accuracy parameter]



$$G_{n,\varepsilon}(\theta) := \{x^n \in \{0,1\}^n : |\hat{\theta}_n - \theta| \le \varepsilon\}$$

$$B_{n,\varepsilon}(\theta) := \{x^n \in \{0,1\}^n : |\hat{\theta}_n - \theta| > \varepsilon\}$$

Fix $\delta \in (0,1)$ [confidence parameter]

$$\mathbb{P}_\theta\{B_{n,\varepsilon}(\theta)\} \le \delta$$

— can we guarantee this for large enough n w/o prior knowledge of θ?

Yes! Chernoff-Hoeffding bound:

$$\mathbb{P}_\theta\{B_{n,\varepsilon}(\theta)\} = \mathbb{P}_\theta\{|\hat{\theta}_n - \theta| > \varepsilon\} \le 2e^{-2n\varepsilon^2}$$

Implications:

- prob. of "bad set" of samples decays exponentially with n (number of tosses)
- the bound is valid of <u>all</u> θ

Given δ (confidence parameter), we need at least

$$\boxed{n \geq \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right)}$$ ← sample complexity of coin tossing

tosses to capture θ in an interval of width 2ε centered on $\hat{\theta}_n$.

Sample Complexity : $n(\varepsilon, \delta)$

polynomial in $\frac{1}{\varepsilon}$

polylogarithmic in $\frac{1}{\delta}$ (poly. in $\log \frac{1}{\delta}$)

– computational learning theory viewpoint: these are "easy" problems (L. Valiant)

Statistical Learning vs. Classical Statistics

success in a given task vs. parameter estimation

I) Ideal case (no learning needed): known underlying distribution

1) Binary classification (pattern recognition)

(X, Y) X is a "feature" taking values in some set $\mathcal{X}$

$Y \in \{0, 1\}$ is a binary label

$(X, Y) \sim P$

Observe X, predict Y

Classifier (predictor) $f : \mathcal{X} \to \{0, 1\}$

Loss (risk) of f on P: $L_P(f) := P\{f(X) \neq Y\}$

$L_P^* := \min_{f: \mathcal{X} \to \{0,1\}} L_P(f)$ — minimum loss

Claim: the optimal classifier is

$$f_P^*(x) = \begin{cases} 1, & \text{if } \eta(x) \geq 1/2 \\ 0, & \text{if } \eta(x) < 1/2 \end{cases}$$

where $\eta(x) := P[Y=1 | X=x] = \mathbb{E}_P[Y|X=x]$.

Proof: fix an arbitrary classifier f

$$L_P(f) = P\{f(X) \neq Y\}$$

$$= \mathbb{E}_P\{\mathbb{1}_{\{f(X)\neq Y\}}\}$$

$\mathbb{1}_{\{\cdot\}}$ - indicator fcn

$$= \int_{\mathcal{X}\times\{0,1\}} \mathbb{1}_{\{f(x)\neq y\}} P(dx, dy)$$

$$= \int_{\mathcal{X}} P_X(dx) \Big\{ P[Y=1|X=x]\mathbb{1}_{\{f(x)\neq 1\}} + P[Y=0|X=x]\mathbb{1}_{\{f(x)\neq 0\}} \Big\}$$

$$= \int_{\mathcal{X}} P_X(dx) \underbrace{\Big\{ \eta(x)\mathbb{1}_{\{f(x)\neq 1\}} + (1-\eta(x))\mathbb{1}_{\{f(x)\neq 0\}} \Big\}}_{:= \ell(f,x)}$$

$$\ell(f,x) = \begin{cases} 1-\eta(x) & \text{if } f(x)=1 \\ \eta(x) & \text{if } f(x)=0 \end{cases}$$

Optimality: $\min_f \ell(f,x) = \min\{1-\eta(x), \eta(x)\}$

take $f(x)=1$ if $1-\eta(x) \leq \eta(x) \iff \eta(x) \geq \frac{1}{2}$.
$f(x)=0$ if $\eta(x) \leq 1-\eta(x)$

$$L_P(f) \geq L_P(f_P^*) = \mathbb{E}[\min\{1-\eta(X), \eta(X)\}]. \quad \blacksquare$$

2) Minimum Mean Square Error (MMSE) estimation

$(X, Y) \sim P$ X takes values in $\mathbb{R}^p$
Y takes values in $\mathbb{R}$

Predictor (estimator) $f : \mathbb{R}^p \to \mathbb{R}$

Loss: $L_P(f) := \mathbb{E}\,(f(X) - Y)^2$

$L_P^* = \min_f L_P(f)$

Claim: the optimal predictor is the conditional mean,

$$f_P^*(x) := \mathbb{E}[Y \mid X = x].$$

Proof (sketch)

$$\begin{aligned} L_P(f) &= \mathbb{E}_P\,(f(X) - Y)^2 \\ &= \mathbb{E}_P\,(f(X) - f_P^*(X) + f_P^*(X) - Y)^2 \\ &= \mathbb{E}_P\,(Y - f_P^*(X))^2 + 2\,\mathbb{E}_P\big[(Y - f_P^*(X))(f_P^*(X) - f(X))\big] \\ &\quad + \mathbb{E}_P\,(f(X) - f_P^*(X))^2 \end{aligned}$$

cross-term $= 0$ (iterated expectation)

$$\begin{aligned} \Rightarrow L_P(f) &= \mathbb{E}\,(Y - \mathbb{E}(Y \mid X))^2 + \mathbb{E}_P\,(f - f_P^*)^2 \\ &\geq \mathbb{E}\,(Y - \mathbb{E}(Y \mid X))^2 \\ &= L_P^* \end{aligned}$$

∎

Takeaway: if $P = \mathcal{L}(X, Y)$ is known, no learning is needed, it's just optimization.

Learning arises when P is unknown, and you get iid samples from P.