BASICS of SUPERVISED LEARNING

Main goal (informally): Given a set of observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, 2, \dots, n$, predict the label $y \in \mathcal{Y}$ of a previously unseen input $x \in \mathcal{X}$.

<u>Nomenclature</u> : $\mathcal{X}$: input set, domain set

$x \in \mathcal{X}$: input, feature, covariate

$y \in \mathcal{Y}$: output, label, response

$S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \dots, n\}$: training set, data set

previously unseen $(x, y) \in \mathcal{X} \times \mathcal{Y}$: test data

<u>Some examples</u> :

x	y
email	SPAM or not
image	CAT or DOG
image of a digit	0, 1, 2, ..., 9

Formal introduction to supervised learning :

First, let's describe the statistical nature of the problem.

$S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \dots, n\}$ is given to the learner,

$(x, y) \in \mathcal{X} \times \mathcal{Y}$ is the test data.

We assume that

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), (x, y)$ are random variables,

drawn <u>independently</u> from a distribution $P = P_X \otimes P_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$, which is <u>unknown</u> to the learner.

The goal is (intuitively) to come up with a mapping $\hat{f}_S : \mathcal{X} \to \mathcal{Y}$ s.t. $\underline{\hat{f}_S(x) \approx y}$ on the test data $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

How to measure?

<u>Loss function</u> : $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Given a predictor $f : \mathcal{X} \to \mathcal{Y}$, $\ell(f(x), y)$ is the loss incurred when the true label is y, and the prediction is $f(x)$.

For any prediction rule $f : \mathcal{X} \to \mathcal{Y}$, the performance criterion is

$$R(f) = \mathbb{E}_{(x,y)\sim P}\left[\ell(f(x), y)\right] \longrightarrow \text{population risk}$$

The ultimate goal in supervised learning:

$$R^* = \inf_{\substack{f : \mathcal{X} \to \mathcal{Y} \\ f \text{ is measurable}}} R(f) \quad \text{is the Bayes risk}$$

Find a predictor $\hat{f} = \hat{f}_S$ based on S s.t. $R(\hat{f}) \approx R^*$ in expectation or with high probability..

Main Problem Classes in Supervised Learning

① Classification $\quad \mathcal{X}$ an arbitrary set

$$\left.\begin{array}{l} \mathcal{Y} = \{0,1\} \text{ or } \mathcal{Y} = \{-1,1\} \\ \ell(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\} \end{array}\right\} \text{ binary classification}$$

$\mathcal{Y} = \{0, 1, \dots, m-1\}$ multi-class classification.

② Regression $\quad \mathcal{Y} = \mathbb{R}, \quad \ell(\hat{y}, y) = |\hat{y} - y|^2.$

Remarks

① why learning? The underlying distribution of the nature, P, is unknown to the learner. Only partial knowledge via $S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ is available.

② Why supervised? A supervisor pre-labels the training input data, the learner is trained under this supervision.

③ Why i.i.d.?

- (x_i, y_i) and (x, y) are identically distributed
  $\Rightarrow$ all training samples provide useful info about the system and (x, y).
- independence $\Rightarrow$ each sample yields maximal info.

④ Realizable vs. agnostic: If $\exists f^* : \mathcal{X} \to \mathcal{Y}$ s.t. $y = f^*(x), \forall x \in \mathcal{X}$, the problem is called "realizable", and the task is to learn f^*.

Our setting is more general as it does not assume a deterministic mapping $x \mapsto f^*(x) = y$. The relation between $x \in \mathcal{X}$ and its label $y \in \mathcal{Y}$ can be random, governed by the conditional distribution $P_{Y|X}$.

Basics of Bayesian Decision-Making

Assume that P is known. $\to$ no learning

What is R^*? How is $R(f)$ minimized?

Recall : X, Y, Z random variables.

$$\mathbb{E}\big[\mathbb{E}[X \mid Y, Z] \mid Y\big] = \mathbb{E}[X \mid Y] \qquad \text{(tower property)}$$

By using the above property, for any $f : \mathcal{X} \to \mathcal{Y}$,

$$R(f) = \mathbb{E}\big[\ell(f(x), y)\big] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x'), y')\, dP(x', y')$$

$$= \mathbb{E}\Big[\mathbb{E}\big[\ell(f(x), y) \mid x\big]\Big]$$

$$= \int_{\mathcal{X}} \mathbb{E}\big[\ell(f(x'), y) \mid x = x'\big]\, dP_X(x')$$

where P_X is the marginal distribution of X. Using this, we characterize R^* and the optimal prediction rule in the general case :

PROPOSITION For any $x' \in \mathcal{X}$, let

$$f^*(x') \in \underset{y' \in \mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}\big[\ell(y', y) \mid x = x'\big] = \int \ell(y', y'')\, dP_{Y|X}(y'' \mid x')$$

Then, $R(f^*) = R^*$, i.e., f^* is a Bayes optimal predictor.

Note : If we knew the conditional dist'n $P_{Y|X}$, we would be able to find f^*. Without the a priori knowledge of P or $P_{Y|X}$, it is not possible.

Let us make the above proposition more explicit on our two broad problem classes.

Corollary (Bayes optimal predictor for binary classification)

Consider $\mathcal{Y} = \{0, 1\}$, and $\ell(y, y') = \mathbb{1}\{y \neq y'\}$. Given the knowledge of P,

$$f^*(x) = \begin{cases} 1, & \text{if } P_{Y|X}(1 \mid x) > 1/2, \\ 0, & \text{if } P_{Y|X}(1 \mid x) < 1/2, \end{cases}$$

Pf:

$$R(f) = \mathbb{P}(f(x) \neq y) = \mathbb{P}(f(x)\neq 1, y=1) + \mathbb{P}(f(x)\neq 0, y=0)$$

$$= \int_{\mathcal{X}} \left(\mathbb{1}\{f(x')\neq 1\} \mathbb{P}(y=1 \mid x=x') + \mathbb{1}\{f(x')\neq 0\} \mathbb{P}(y=0 \mid x=x') \right) d\mathbb{P}_x(x')$$

$$= \int_{\mathcal{X}} \Big(P_{Y|X}(1|x') + \underbrace{\mathbb{1}\{f(x')\neq 0\} \left[P_{Y|X}(0|x') - P_{Y|X}(1|x') \right]}_{(+)} \Big) dP_X(x')$$

How to minimize $R(f)$?

For $x' \in \mathcal{X}$, • if

$$P_{Y|X}(0|x') - P_{Y|X}(1|x') > 0, \quad \text{then set} \quad f(x') = 0$$

so that

$$\mathbb{1}\{f(x')\neq 0\} = 0, \quad \text{and} \quad (+) = 0.$$

Otherwise, $(+)$ would be strictly positive.

Recall that $P_{Y|X}(0|x') + P_{Y|X}(1|x') = 1, \ \forall x' \in \mathcal{X}$.

Thus, set $f(x') = 0$ if

$$1 - 2P_{Y|X}(1|x') > 0 \Rightarrow P_{Y|X}(1|x') < 1/2$$

• if $P_{Y|X}(0|x') - P_{Y|X}(1|x') = 1 - 2P_{Y|X}(1|x') \leq 0$,

then one must set $f(x') = 1$ so that

$$\mathbb{1}\{f(x')\neq 0\} = 1 \quad \text{and} \quad (+) < 0. \qquad \blacksquare$$

<u>Corollary 2</u> (Bayes optimal predictor for the regression problem)

Given P, $\mathcal{Y} = \mathbb{R}$ and $R(f) = \mathbb{E}[(f(x)-y)^2]$, $f: \mathcal{X} \to \mathcal{Y}$, the Bayes optimal predictor is

$$f^*(x') = \mathbb{E}[y \mid x = x'], \quad \forall x' \in \mathcal{X}.$$

Pf: $R(f) = \mathbb{E}[(y - f(x))^2]$

$$= \mathbb{E}\left[\left(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - f(x) \right)^2 \right]$$

$$= \mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right)^2\right] + \mathbb{E}\left[\left(f(x)-\mathbb{E}[y|x]\right)^2\right]$$

$$+ \; 2 \cdot \underbrace{\mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right)\left(f(x)-\mathbb{E}[y|x]\right)\right]}_{(*)}$$

Note that

$$(*) = \mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right)\left(f(x)-\mathbb{E}[y|x]\right)\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right) \cdot \underbrace{\left(f(x)-\mathbb{E}[y|x]\right)}_{\sigma(x)\text{-measurable}} \,\middle|\, x\right]\right]$$

$$= \mathbb{E}\left[\left(f(x)-\mathbb{E}[y|x]\right) \cdot \mathbb{E}\left[y - \mathbb{E}[y|x] \,\middle|\, x\right]\right]$$

$$= \mathbb{E}\left[\left(f(x)-\mathbb{E}[y|x]\right) \cdot \underbrace{\left(\mathbb{E}[y|x] - \mathbb{E}\left[\mathbb{E}[y|x] \,|\, x\right]\right)}_{=0}\right]$$

$$= 0$$

Thus,

$$R(f) = \underbrace{\mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right)^2\right]}_{\text{does not depend on } f} + \underbrace{\mathbb{E}\left[\left(f(x)-\mathbb{E}[y|x]\right)^2\right]}_{\text{can be minimized iff } f(x) = \mathbb{E}[y|x] \text{ a.s.}}$$

Thus,

$$R^* = \mathbb{E}\left[\left(y-\mathbb{E}[y|x]\right)^2\right]$$

$$= \mathbb{E}\left[\mathrm{Var}(y|x)\right].$$

In summary, the optimal prediction rule <u>heavily</u> depends on $P_{Y|X}$:

- for binary classification, $f^*(x') = \begin{cases} 1, & \text{if } P_{Y|X}(1|x') \geq 1/2, \\ 0, & \text{otherwise} \end{cases}$

- for regression, $f^*(x') = \mathbb{E}[y|x']$.