Lecture 22

AoI in supervised learning.

Reading:

Shisher et al AoI workshop 2021

Empirical risk minimization (ERM).

- Decision maker predicts $Y \in \mathcal{Y}$ by taking an action $a = \psi(X) \in \mathcal{A}$ based on feature $X \in \mathcal{X}$.
  where $\psi : \mathcal{X} \mapsto \mathcal{A}$ is a decision function.
- loss function: $L : \mathcal{Y} \times \mathcal{A} \mapsto \mathbb{R}$
  $L(y, a)$ is the loss if action a is chosen when $Y = y$.
- risk: $\frac{1}{n} \sum_{i=1}^{n} L(y_i, \psi(x_i))$.
  where $(x_i, y_i)_{i=1}^{n}$ is a set of training data samples.
- Def: P_{XY} as the empirical distribution of the traing data set.
  Then, risk is

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, \psi(x_i)).$$

$$= E_{X,Y \sim P_{XY}} [L(Y, \psi(X))].$$

- Empirical risk minimization (ERM).

$$\min_{\psi \in \mathcal{F}} E_{X,Y \sim P_{XY}} [L(Y, \psi(X))].$$

$\mathcal{F}$: set of allowed decision functions.

Examples:

- log-loss: action $a = Q_Y$ is distribution of Y.

$$L_{\log}(y, Q_Y) = -\log[Q_Y(y)]$$

$$E_{Y \sim P_Y}[-\log Q_Y(Y)].$$

cross-entropy

- Quadratic-loss:

action $a = \hat{y}$.

$$L_2(y, \hat{y}) = (y - \hat{y})^2.$$

$$E[(Y - \hat{Y})^2].$$

$\mathcal{F}$:

- the set of linear function:

$$\psi(x) = \vec{a} \cdot \vec{x} + \vec{c}.$$

- the set of function denoted by neural networks.

We consider a fundamental lower bound of ERM. given by.

$$\min_{\psi \in \Psi} E_{XY \sim P_{XY}}[L(Y, \psi(X))],$$

where Ψ is the set of all functions from $\mathcal{X}$ to $\mathcal{A}$.

$$\mathcal{F} \subseteq \Psi.$$

Goal: We will show that the optimum objective value is a version of conditional entropy $H_L(Y|X)$.

Shannon Information Theory:

- Entropy:

$$H(Y) = -\sum_{y \in \mathcal{Y}} P_Y(y) \log P_Y(y).$$

- Conditional Entropy given $X = x$:

$$H(Y|X=x) = -\sum_{y \in \mathcal{Y}} P_{Y|X=x}(y) \log P_{Y|X=x}(y).$$

- Conditional Entropy given X:

$$H(Y|X) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}(x,y) \log P_{Y|X=x}(y)$$

$$= \sum_{x \in \mathcal{X}} P_X(x) H(Y|X=x).$$

$$H(Y) \geq H(Y|X).$$

- Mutual Information:

$$I(X;Y) = H(X) - H(X|Y)$$

$$= H(Y) - H(Y|X).$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)}.$$

$$I(X;Y) \geq 0.$$

$$I(X;Y) = I(Y;X).$$

- Divergence:

$$D(P_Y \,//\, Q_Y)$$

$$= \sum_{y \in \mathcal{Y}} P_Y(y) \log \frac{P_Y(y)}{Q_Y(y)}.$$

$$I(X;Y) = D(P_{X,Y} \,//\, P_X \otimes P_Y).$$

- Condition Mutual Information:

$$I(X;Y|Z).$$

$$= \sum_{\substack{x \in \mathcal{X}, y \in \mathcal{Y} \\ z \in \mathcal{Z}}} P_{XYZ}(x,y,z) \log \frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)\, P_{Y|Z}(y|z)}.$$

$$= H(Y|Z) - H(Y|XZ).$$

$$= H(X|Z) - H(X|YZ).$$

$$= I(X;YZ) - I(X;Z) = H(X) - H(Y|YZ) - [H(X) - H(X|Z)]$$

$$= I(X,Z;Y) - I(Z;Y).$$

$$= E_{Z \sim P_Z}[D(P_{XY|Z} \,//\, P_{X|Z} \otimes P_{Y|Z})].$$

Generalized Information theoretic measures for ERM:

- Entropy: $H_L(Y) = \min_{a \in \mathcal{A}} E_{Y \sim P_Y}[L(Y, a)]$.

  entropy associated with a loss function L.

  minimum loss without knowing X.

- Conditional entropy of Y given $X = x$:

  $$H_L(Y|X=x) = \min_{a \in \mathcal{A}} E_{Y \sim P_{Y|X=x}}[L(Y, a)].$$

  $$= \min_{\psi(x) \in \mathcal{A}} E_{Y \sim P_{Y|X=x}}[L(Y, \psi(x))].$$

- Conditional entropy of Y given X:

  $H_L(Y|X)$

  $$= \min_{\psi \in \Psi} E_{XY \sim P_{XY}}[L(Y, \psi(X))].$$

  $$= \min_{\psi \in \Psi} \sum_{x,y} P_{XY}(x,y) \, L(y, \psi(x)).$$

$$= \min_{\psi \in \Psi} \sum_{x \in X} P_X(x) \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x)).$$

$$= \min_{\psi(x) \in A, \forall x} \sum_{x \in X} P_X(x) \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x))$$

$$= \sum_{x \in X} P_X(x) \min_{\psi(x) \in A} \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x))$$

$$= \sum_{y \in X} P_X(x) \min_{\psi(x) \in A} E_{Y \sim P_{Y|X=x}}[L(Y, \psi(x))]$$

$$= \sum_{x \in X} P_X(x) H_L(Y|X=x).$$

$$H_L(Y|X) = E_{X \sim P_X}[H_L(Y|X=x)]$$

- Mutual information:

$$I_L(Y;X) = H_L(Y) - H_L(Y|X).$$

the reduction of the loss due to knowledge of X.

$$I_L(Y;XZ) = I_L(Y;Z) + I_L(Y;X|Z)$$

$$= [H_L(Y) - H_L(Y|Z)] + [H_L(Y|Z) - H_L(Y|XZ)]$$

Note:

$I_L(Y;X) \neq I_L(X;Y)$ in general.

Shannon's mutual function is a special case.

where. $I(Y;X) = I(X;Y)$.