Lecture 1: Fundamental Concept of Statistical Learning

Tianpei Xie

Jul. 25th., 2015

Contents

1 Fundamental Concepts

1.1 Definitions

- **Definition** (***Data***)
  Define an ***observation*** as a d-dimensional vector x. The *unknown* nature of the observation is called a ***class***, denoted as y. The domain of observation is called an ***input space*** *or* ***feature space***, denoted as $\mathcal{X} \subset \mathbb{R}^d$, whereas the domain of class is called the ***target space***, denoted as $\mathcal{Y}$. For ***classification task***, $\mathcal{Y} = \{1, \ldots, M\}$; and for ***regression task***, $\mathcal{Y} = \mathbb{R}$. Denote a collection of n ***samples*** as $\mathcal{D}_n := \{(x_i, y_i) : 1 \leq i \leq n\}$. $\mathcal{D}_n$ is a finite subset in $\mathcal{X} \times \mathcal{Y}$.

- **Definition** (***Concept Class as a Function Class***)
  A ***concept*** $c : \mathcal{X} \to \mathcal{Y}$ is the *input-output association* from the nature and is *to be learned* by a learning algorithm. Denote $\mathcal{C}$ as *the set of all concepts* we wish to learn as the ***concept class***:

  $$\mathcal{C} := \{c : \mathcal{X} \to \mathcal{Y}\} = \mathcal{Y}^{\mathcal{X}}.$$

  Concept class $\mathcal{C}$ is a *function class.*

- Learning is formalized into two different scenarios:

  1. In ***deterministic*** scenario: Assume that there exist measurable space $(\mathcal{X}, \mathscr{B})$, where $X \in \mathcal{X}$ is the ***random vector*** in $\mathcal{X}$, i.e.

     $$X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{X}, \mathscr{B})$$

     is $\mathscr{F}/\mathscr{B}$ measurable. Let $\mathcal{P}_X$ be *the induced probability distribution* on X.

     Remark (***Sample in Deterministic Scenario***)
     In deterministic scenario, denote a collection of n *independent identically distributed (i.i.d.)* ***random samples*** generated by P_X as $\mathcal{D}_n$, i.e.

     $$\mathcal{D}_n := \{X_i : 1 \leq i \leq n\}.$$

     Note that $\mathcal{D}$ is a finite subset in $\mathcal{X}$.

     Remark (***Learning Task in Deterministic Scenario***)
     Given a collection of *i.i.d. samples* $\mathcal{D}$ generated by $\mathcal{P}_X$, a ***learner*** considers a ***fixed*** *subset of concepts* $\mathcal{H} \subset \mathcal{C}$, which is referred as a ***hypothesis class***, and provides a ***hypothesis*** or a ***classifier*** or a ***decision function*** $g \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ based on $\mathcal{D}$. The task of ***supervised learning*** is to minimize

     Definition (***Generalization Error in Deterministic Scenario***) [?]
     Under *a deterministic scenario*, ***generalization error*** or the ***risk*** or simply ***error*** for the *classifier* $g \in \mathcal{H}$ is defined as

     $$R(g) \equiv L(g) = \mathcal{P}_X \{g(X) \neq c(X)\} \equiv \mathbb{E}_{\mathcal{P}_X} [\mathbb{1} \{g(X) \neq c(X)\}] \tag{1}$$

     with respect to the concept $c \in \mathcal{C}$.

2. In ***stochastic*** scenario: Assume *both* X and Y are random, i.e. there exists a probability space $(\mathcal{X}\times, \mathcal{Y}, \mathscr{B}, \mathcal{P}_{X,Y})$ so that

   $$(X, Y) : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{X} \times \mathcal{Y}, \mathscr{B}, \mathcal{P}_{X,Y})$$

   so that the pair (X, Y) is $\mathscr{F}/\mathscr{B}$ measurable. Let $\mathcal{P}_{X,Y}$ be *the induced* ***joint probability distribution*** on (X, Y).

   Remark (***Sample in Stochastic Scenario***)
   In stochastic scenario, denote a collection of n *independent identically distributed (i.i.d.)* ***random sample pairs*** generated by joint probability $P_{X,Y}$ as $\mathcal{D}$, i.e.

   $$\mathcal{D}_n := \{(X_i, Y_i) : 1 \le i \le n\}.$$

   Note that $\mathcal{D}_n$ is a finite subset in $\mathcal{X} \times \mathcal{Y}$.

   Definition (***Generalization Error in Stochastic Scenario***) [Mohri et al., 2018]
   In *stochastic scenario*, ***generalization error*** or the ***risk*** or simply ***error*** for the *classifier* $g \in \mathcal{H} \subset \mathcal{C}$ is defined as

   $$R(g) \equiv L(g) = \mathcal{P}_{X,Y}\{g(X) \neq Y\} \equiv \mathbb{E}_{\mathcal{P}_{X,Y}}[\mathbb{1}\{g(X) \neq Y\}] \tag{2}$$

   Remark (***Learning Task in Stochastic Scenario***)
   Given *the sample set S* generated from a (joint) probability distribution $P_{X,Y}$. Given a fixed hypothesis class $\mathcal{H}$, the task of learner is to find a hypothesis $g \in \mathcal{H} \subset \mathcal{C}$ so that the generalization error or the *risk* or simply the *error* is minimized.

- **Remark** (***Deterministic*** *vs.* ***Stochastic***)
  The main difference between these two settings is the assumption on Y:

  1. ***In deterministic scenario***, $Y = c(X)$ for some ***unknown*** *but* ***deterministic*** $c \in \mathcal{C}$ and the learning task is to approximate c by some function $g \in \mathcal{H}$.
  2. ***In stochastic scenario***, Y is a ***random variable***, ***generated jointly*** *with the feature* X by some unknown distribution $\mathcal{P}_{X,Y}$.

     The pair (X, Y) *may not follow* a ***function relationship***. Note for a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ to follow function relationship, for *each given* x, there *can only be one* correspoding $y \in \mathcal{Y}$. Under the stochastic assumption, *any pair* $(x, y) \in \mathcal{X} \times \mathcal{Y}$ *would appear* as long as the corresponding measure $\mathcal{P}_{X,Y}(x, y) > 0$.
  3. *An* ***intermediate*** *setting* assumes feature is a random vector $X \sim P_X$, and class $Y = g(X)$ for some ***unknown random function*** g.

     $$g : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{C}, \mathscr{C})$$

     where g is $\mathscr{F}/\mathscr{C}$ measurable, and $g(\omega) \in \mathcal{C} = \mathcal{Y}^{\mathcal{X}}$ for each $\omega \in \mathcal{C}$.

     One may assume that g is generated following an independent generation process from $\mathcal{P}_X$. Or, one may assume that (X, g) are not independent, e.g. $g = g(\cdot|\sigma(X_1, \ldots, X_n))$ is determined by a stochastic process $X_t : t \le n$ of past events.

     In [Devroye et al., 2013], the author defines the random function g as the function of stochastic process $\{(X_t, Y_t) : t \le n\}$ of past events assuming both (X, Y) are random.

Definition (***Random Function by Independent Random Samples***) [Devroye et al., 2013]
Given a collection of samples $\mathcal{D}_n = ((X_i, Y_i), 1 \le i \le n)$, a ***(stochastic) classifier /hypothesis*** is defined as

$$\begin{aligned} g_n(x) &= g_n(x\,;\mathcal{D}_n) \\ &:= g(x|\sigma((X_i, Y_i) : 1 \le i \le n)). \end{aligned}$$

Thus g_n is a *(random) function* determined by σ-algebra $\sigma(\mathcal{D}_n) := \sigma((X_t, Y_t) : t \le n)$, and its output is considered as a *random variable* depended upon data $\mathcal{D}_n$. Note that it should be distinguished with the fixed concept $c \in \mathcal{C}$ or a unknown but fixed hypothesis $g(\cdot) \in \mathcal{H}$, whereas $g_n(\cdot\,;\mathcal{D}_n) \in \mathcal{H}$.

A sequence of hypotheses $\{g_n\}_n$ is called ***a classification rule*** where each g_n is a function of data so is a random mapping

$$g_n : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}.$$

- **Definition** (***Generalization Error of Estimated Hypothesis***) [Devroye et al., 2013]
Given the data $\mathcal{D}_n$, we can define ***the conditional probability of error***:

$$L_n(g) = L(g_n) := \mathcal{P}_{X,Y}\left\{g_n(X; \mathcal{D}_n) \neq Y | \mathcal{D}_n\right\} \equiv \mathbb{E}_{X,Y}\left[\mathbb{1}\left\{g_n(X) \neq Y\right\} | \mathcal{D}_n\right] \quad (3)$$

This is a random variable because it depends upon the data $\mathcal{D}_n$. So, L_n averages over the distribution of (X, Y), but *the* ***data*** *is held* ***fixed***. Averaging over the data as well would be unnatural, because in a given application, one has to live with the data at hand.

- **Remark** (***Learning in Deterministic*** *vs.* ***Stochastic***)

  1. (***Function Approximation***): The learning task in ***derministic scenario*** is to ***approximate*** $c \in \mathcal{C}$ with $g \in \mathcal{H} \subset \mathcal{C}$ given samples $\mathcal{D}$. The ***function approximator*** g should be "***close***" to the unknown c *under the unknown distribution* P_X.

     The theorectial analysis concerns that under ***the worst case scenario***, if it is possible for a function g in function class $\mathcal{H}$ to approximate c so that the generation error approaches to zero.

  2. (***Distribution Approximation***): The learning task in ***stochastic scenario*** is to ***approximate*** the joint probability measure $\mathcal{P}_{X,Y}$ with $\widehat{\mathcal{P}}_{X,Y}$ given samples $\mathcal{D}$. ***The distribution estimator*** $\widehat{\mathcal{P}}_{X,Y}$ should "***converge***" to the unknown $\mathcal{P}_{X,Y}$ asymptotically.

- **Remark** : (***Statistical Decision***) [Berger, 2013]
The *statistical learning theory* is closely related to the *statistical decision theory* in which the terms such as *(Empirical) Risk/Utility*, *decision function* are used as an alternative to the terms like *(Empirical) Error*, *hypothesis/classifier*.

1.2 Bayes Error

- **Definition** (***Bayes Error in Stochastic Scenario***)
Under a given distribution $\mathcal{P}_{X,Y}$, the ***Bayes error*** L^* or ***Bayes risk*** R^* is defined as

$$R^* \equiv L^* = \inf_{g \in \mathcal{C}} \left\{L(g)\right\}, \quad (4)$$

where the infimum is with respect to *all measureable function* $g : \mathcal{X} \to \mathcal{Y}$. And the *hypothesis* g^* such that $L(g^*) = L^*$ is called the ***Bayes classifier***.

- **Remark** ***The Bayes Error is only a function of underlying distribution*** $\mathcal{P}_{X,Y}$ and it does not depend on choice of function g or function class $\mathcal{H}$.

$$L^* = L^*(\mathcal{P}_{X,Y}) := \inf_{g} \left\{\mathcal{P}_{X,Y} \left\{g(X) \neq Y\right\}\right\}.$$

- **Remark** (***Bayes Error in Deterministic Scenario***)
Under ***the deterministic*** setting, the Bayes error is $L^* = 0$ since by assumption $Y = c(X)$ for some $c \in \mathcal{C}$, thus the infimum is zero.

- **Remark** (***Bayes Classifier if*** $\mathcal{P}_{X,Y}$ ***is Known***)
The learning task is concerning about the situation when $\mathcal{P}_{X,Y}$ is ***unknown*** but ***if*** $\mathcal{P}_{X,Y}$ ***is known***, then ***the optimal hypothesis*** is known as ***the posterior conditional expectation***:

$$\eta(X) := \mathcal{P}[Y|X] = \frac{d\mathcal{P}_{X,Y}}{d\mathcal{P}_X}$$

Note that $\mathcal{P}[Y|X]_\omega$ is a function of X given each $\omega \in \Omega$, which means that $g(X, \omega) := \mathcal{P}[Y|X]_\omega$ is a random function itself. For $\mathcal{Y} = \{0, 1\}$ and X be discrete random variables, it can be written as

$$\begin{aligned} \eta(x) &= \mathcal{P}\left\{Y = 1 \middle| X = x\right\} \\ &= \mathbb{E}_{p(y|x)}\left[y \middle| X = x\right]. \end{aligned} \tag{5}$$

and ***the Bayes classifier*** (decision function)

$$\begin{aligned} g^*(x) &= \text{argmax}_{y \in \{0,1\}} P(Y|X = x) \\ &= \begin{cases} 1 & \eta(x) > \frac{1}{2} \\ 0 & \text{o.w.} \end{cases} \end{aligned} \tag{6}$$

with the corresponding ***Bayes error***

$$\begin{aligned} L^* &= \mathbb{E}_{P(X)}\left[\min\left\{P(Y = y|X) \mid y \in \{0, 1\}\right\}\right] \\ &= 1 - \mathbb{E}_{P(X)}\left[\eta(X)\mathbb{1}\left\{\eta(X) > 1/2\right\} + (1 - \eta(X))\mathbb{1}\left\{\eta(X) \le 1/2\right\}\right] \end{aligned} \tag{7}$$

- We summarizes our discussion as follows

Proposition 1.1 ***(Conditional Estimator is Bayes Classifer if Distribution is Known)*** *[Devroye et al., 2013]*
Given the posterior (conditional) probability $\eta(x) = \mathcal{P}(Y = 1|X = x) = \mathbb{E}_{p(y|x)}[Y|X = x]$, *where* $\mathcal{P}(X, Y)$ *is the underlying distribution of data and the Bayes decision function*

$$\begin{aligned} g^*(x) &= \mathbb{1}\left\{\mathcal{P}(Y = 1|X = x) > 1/2\right\} \\ &= \mathbb{1}\left\{\mathbb{E}_{p(y|x)}\left[y \middle| X = x\right] > 1/2\right\}, \end{aligned}$$

for any decision function $g : \mathcal{X} \to \{0, 1\}$,

$$\mathcal{P}\left\{g^*(X) \neq Y\right\} \le \mathcal{P}\left\{g(X) \neq Y\right\}$$

Proof: Given $X = x$, the conditional error probability of any g can be expressed as

$$\begin{aligned}
&\mathcal{P}\left\{g(X) \neq Y | X = x\right\} \\
&= 1 - \mathcal{P}\left\{Y = g(X) | X = x\right\} \\
&= 1 - \left(\mathcal{P}\left\{Y = 1, g(X) = 1 | X = x\right\} + \mathcal{P}\left\{Y = 0, g(X) = 0 | X = x\right\}\right) \\
&= 1 - \left(\mathbb{1}\left\{g(x) = 1\right\} \mathcal{P}\left\{Y = 1 | X = x\right\} + \mathbb{1}\left\{g(x) = 0\right\} \mathcal{P}\left\{Y = 0 | X = x\right\}\right) \\
&= 1 - \left[\mathbb{1}\left\{g(x) = 1\right\} \eta(x) + \mathbb{1}\left\{g(x) = 0\right\} (1 - \eta(x))\right] \qquad (8)
\end{aligned}$$

For any $x \in \mathcal{X}$,

$$\begin{aligned}
&\mathcal{P}\left\{g(X) \neq Y | X = x\right\} - \mathcal{P}\left\{g^*(X) \neq Y | X = x\right\} \\
&= \eta(x)\left(\mathbb{1}\left\{g^*(x) = 1\right\} - \mathbb{1}\left\{g(x) = 1\right\}\right) + (1 - \eta(x))\left(\mathbb{1}\left\{g^*(x) = 0\right\} - \mathbb{1}\left\{g(x) = 0\right\}\right) \\
&= (2\eta(x) - 1)\left(\mathbb{1}\left\{g^*(x) = 1\right\} - \mathbb{1}\left\{g(x) = 1\right\}\right) \\
&\geq 0,
\end{aligned}$$

since $g^*(x) = 1$ if and only if $(2\eta(x) - 1) > 0$ and $(\mathbb{1}\{g^*(x) = 1\} - \mathbb{1}\{g(x) = 1\}) \geq 0$ if and only if $g^*(x) = 1$. ∎

- **Proposition 1.2** ***(Plug-In Estimator)*** *[Devroye et al., 2013]*
Consider a plug-in decision function

$$g(x) = \mathbb{1}\left\{\tilde{\eta}(x) > 1/2\right\},$$

where $\tilde{\eta}(x)$ is an estimate of $\eta(x) = \mathcal{P}(Y = 1 | X = x)$, then for the error probability of plug-in decision function $g(x)$, we have

$$\mathcal{P}\left\{g(X) \neq Y\right\} - L^* = 2\int_{\mathcal{X}} |\eta(x) - 1/2|\, \mathbb{1}\left\{g(x) \neq g^*(x)\right\} \mu(dx) \qquad (9)$$

and

$$\begin{aligned}
\mathcal{P}\left\{g(X) \neq Y\right\} - L^* &\leq 2\int_{\mathcal{X}} |\eta(x) - \tilde{\eta}(x)|\, \mu(dx) \\
&= 2\mathbb{E}_{p(X)}\left[\eta(X) - \tilde{\eta}(X)\right] \qquad (10)
\end{aligned}$$

Proof: If for some $x \in \mathcal{X}$, $g(x) = g^*(x)$, the clearly the difference btw the conditional error probability of g and g^* is zero; i.e.

$$\mathcal{P}\left\{g(X) \neq Y | X = x\right\} - \mathcal{P}\left\{g^*(X) \neq Y | X = x\right\} = 0.$$

Otherwise, $g(x) \neq g^*(x)$, then

$$\begin{aligned}
&\mathcal{P}\left\{g(X) \neq Y | X = x\right\} - \mathcal{P}\left\{g^*(X) \neq Y | X = x\right\} \\
&= (2\eta(x) - 1)\left(\mathbb{1}\left\{g^*(x) = 1\right\} - \mathbb{1}\left\{g(x) = 1\right\}\right) \\
&= |2\eta(x) - 1|\, \mathbb{1}\left\{g(x) \neq g^*(x)\right\}
\end{aligned}$$

Thus

$$\begin{aligned}
\mathcal{P}\left\{g(X) \neq Y\right\} - L^* &= 2\int_{\mathcal{X}} |\eta(x) - 1/2|\, \mathbb{1}\left\{g(x) \neq g^*(x)\right\} \mu(dx) \\
&\leq 2\int_{\mathcal{X}} |\eta(x) - \tilde{\eta}(x)|\, \mu(dx),
\end{aligned}$$

since $g(x) \neq g^*(x)$ implies $|\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2|$. ∎

- **Corollary 1.3** *Consider a plug-in decision function*

$$g(x) = \mathbb{1}\left\{\tilde{\eta}_1(x) > \tilde{\eta}_0(x)\right\},$$

where $\tilde{\eta}_1(x)$ is an estimate of $\eta(x)$ and $\tilde{\eta}_0(x)$ is an estimate of $1-\eta(x)$, then for the error probability of plug-in decision function $g(x)$, we have

$$\mathcal{P}\left\{g(X) \neq Y\right\} - L^* \le \int_{\mathcal{X}} \left|(1-\eta(x)) - \tilde{\eta}_0(x)\right| \mu(dx) + \int_{\mathcal{X}} \left|\eta(x) - \tilde{\eta}_1(x)\right| \mu(dx) \tag{11}$$

In particular, if $\tilde{\eta}_1(x) \equiv \tilde{q}_1\tilde{p}_1(x)$ and $\tilde{\eta}_0(x) \equiv \tilde{q}_0\tilde{p}_0(x)$, where $\tilde{q}_1, \tilde{q}_0$ are estimate of prior distribution for $\mathcal{P}\left\{Y=1\right\} = q$ and $\mathcal{P}\left\{Y=0\right\} = 1-q$ and $\tilde{p}_1(x), \tilde{p}_0(x)$ are estimate of class conditional distribution of x given $Y=1$ and $Y=0$ respectively, then

$$\mathcal{P}\left\{g(X) \neq Y\right\} - L^* \le \int_{\mathcal{X}} \left|(1-q)p_0(x) - \tilde{q}_0\tilde{p}_0(x)\right| dx + \int_{\mathcal{X}} \left|qp_1(x) - \tilde{q}_1\tilde{p}_1(x)\right| dx$$

- **Definition** (***Generative vs. Discriminative Model***)
In stochastic scenario, following the proposition above, we have two ***learning strategies***:

  - A ***generative model*** is an estimate $\widehat{\mathcal{P}}_{X,Y}$ of joint distribution $\mathcal{P}_{X,Y}$. For high dimensional data, an *efficient* estimator is hard to find.

  - A ***deterministic model*** $g : \mathcal{X} \to \mathcal{Y}$ is a function (hypothesis) in $\mathcal{H}$ from input to output. The task of learner is to find $g \in \mathcal{H}$ so that the generalization error is minimized;

    In *probabilistic graphical models* and *Bayesian learning*, e.g. [Koller and Friedman, 2009, Murphy, 2012], a deterministic model is interpreted as an ***estimate*** $\widehat{\mathcal{P}}(Y|X=x)$ of $\mathcal{P}(Y|X=x)$, *the conditional distribution* of Y given the observations $X=x$, so that

$$\begin{aligned} g(x) &= \mathbb{1}\left\{\widehat{\mathcal{P}}(Y=1|X=x) > 1/2\right\} \\ &= \mathbb{1}\left\{\mathbb{E}_{\widehat{\mathcal{P}}(y|x)}\left[y\middle|\, X=x\right] > 1/2\right\} \end{aligned}$$

    is close to the Bayes classifier

$$\begin{aligned} g^*(x) &= \mathbb{1}\left\{\eta(x) > 1/2\right\} \\ &= \mathbb{1}\left\{\mathbb{E}_{\mathcal{P}(y|x)}\left[y\middle|\, X=x\right] > 1/2\right\} \end{aligned}$$

    $\mathcal{P}(Y|X=x)$ is easier to estimate than $\mathcal{P}_{X,Y}$ since Y is of lower dimensionality.

- **Exercise 1.4** ***(Transformation Increases Bayes Error)*** *[Devroye et al., 2013]*
Let $T : \mathcal{X} \to \mathcal{X}'$ be an arbitrary measureable function. If $L^*_{\mathcal{X}}$ and $L^*_{T(\mathcal{X})}$ denote the Bayes error probability for (X,Y) and $(T(X),Y)$, respectively, then prove that

$$L^*_{T(\mathcal{X})} \ge L^*_{\mathcal{X}}.$$

This shows that transformation destroys information, because the Bayes risk increases.

Proof: We see that for any measureable set $B \subset \mathcal{X}'$, $\mathcal{P}(T(X) \in B) = \mathcal{P}\left\{X \in T^{-1}(B)\right\}$. Define the posterior distribution

$$\begin{aligned} \eta_T(t) &\equiv \mathcal{P}\left\{Y=1|T(x)=t\right\} \\ \eta_T(T(x)) &= \mathbb{E}\left[\eta(X)|T(x)\right]. \end{aligned}$$

Use the *F-error* theorem by observing that $L^* = d_F(X,Y)$ with $F(x) = \min\{x, 1-x\}$, thus

$$\begin{aligned}
L^*_{T(\mathcal{X})} &= d_F(T(X),Y) \\
&= \int \min\{1-\eta_T(T(x)), \eta_T(T(x))\}\, p(x)\mu(dx) \\
&= \int \min\{1-\mathbb{E}[\eta(X)|T(x)], \mathbb{E}[\eta(X)|T(x)]\}\, p(x)\mu(dx) \\
&\geq \int \mathbb{E}[\min\{1-\eta(X), \eta(X)\}\,|T(x)]\, p(x)\mu(dx) \\
&= \int \min\{1-\eta(x), \eta(x)\}\, p(x)\mu(dx) \\
&= d_F(X,Y) = L^*_{\mathcal{X}}. \quad \blacksquare
\end{aligned}$$

Note that for any measureable $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{X}', \mathcal{B}')$, let $X' = T(X)$ for $X : \Omega \to \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}[Y|T(X)] &= \mathbb{E}\left[Y|T^{-1}(\sigma(X'))\right] \\
&= \mathbb{E}\left[Y|\,\sigma(X)|_{T^{-1}(\sigma(X'))}\right] \\
&\equiv \mathbb{E}[\mathbb{E}[Y|\sigma(X)]\;|T(X)] \\
\text{for } E \in T^{-1}(\sigma(X')) = \sigma(T^{-1}X') &\subset \sigma(X) \subset \sigma(X,Y) \\
\int_E \mathbb{E}\left[\mathbb{E}[Y|\sigma(X)]\;|\sigma(T^{-1}X')\right] dP_{X,Y} &= \int_E \mathbb{E}[Y|\sigma(X)]\, dP_{X,Y} \\
&= \int_E Y dP_{X,Y} \\
&= \int_E \mathbb{E}[Y|T(X)]\, dP_{X,Y} \quad \blacksquare
\end{aligned}$$

- **Exercise 1.5** *[Devroye et al., 2013]*
  Let X' be independent with (X,Y). Show that

  $$L^*_{X',X} = L^*_X.$$

  Proof: Just need to see that $\eta'(x',x) = \mathcal{P}(Y|(X',X) = (x',x)) = \mathcal{P}(Y|X=x) = \eta(x)$ by independence, the result then follows directly. $\blacksquare$

1.3 Consistency

- **Remark** Without explict statement, we assume stochastic scenario, and the estimated hypothesis is written as

  $$g_n(x) = g(x|\sigma((X_i,Y_i), i \leq n)) = g(x; \mathcal{D}_n)$$

  where $\mathcal{D}_n = ((X_i,Y_i), 1 \leq i \leq n)$, $\sigma((X_i,Y_i), i \leq n)) = \sigma(\mathcal{D}_n)$. For each x, $g_n(x)$ is a random variable itself since it depends on $\mathcal{D}_n$, which is a collection of random variables.

- **Definition** (***Consistent Classification Rules***)
  A classification rule $\{g_n\}$ is ***consistent (asymoptotically Bayes-risk efficient)*** ***for a certain distribution*** $\mathcal{P}_{X,Y}$ if

  $$L_n(g) := L(g_n) = \mathcal{P}_{X,Y}\left\{g_n(X) \neq Y | \mathcal{D}_n\right\} \overset{\mathcal{P}}{\to} L^*, \;\text{ as } n \to \infty$$

  Since $1 \geq L_n \geq L^*$, the above is equivalent to ***convergence in probability***

  $$\lim_{n\to\infty} \mathcal{P}\left\{L_n - L^* \geq \epsilon\right\} = 0.$$

  Also the classification rule is the ***strongly consistent*** if

  $$L_n := L(g_n) \to L^* \;\; a.s.$$

- **Remark** Given (X_i, Y_i) are i.i.d., $\mathcal{P}(\mathcal{D}_n) = \mathcal{P}^n_{X,Y}$. And $\mathcal{P}\{L_n \leq \epsilon\} := \mathcal{P}^n_{X,Y}\{L_n \leq \epsilon\}$.

- **Remark** ***A consistent rule*** $\{g_n\}$ guarantees us that taking more samples essentially suffices to *roughly* ***reconstruct*** *the unknown distribution* of (X, Y) because L_n can be pushed as close as desired to L^*. In other words, *infinite amounts of information can be gleaned from finite samples.* Without this guarantee, we would not be motivated to take more samples.

  We should be careful and ***not impose conditions on*** (X, Y) *for the consistency of a rule*, because such conditions may not be verifiable.

- A stronger version of consistency even if the underlying distribution $\mathcal{P}$ is unknown

  Definition (***Universal Consistency***)
  A sequence of *classification rules* is called ***universally consistent (strongly) consistent*** if it is ***(strongly) consistent*** for ***any distribution*** $\mathcal{P}(X, Y)$, i.e.

  $$\lim_{n\to\infty} \mathcal{P}\left\{L_n - L^* \geq \epsilon\right\} = 0, \quad \forall\, \mathcal{P}$$

  and

  $$\mathcal{P}\left\{\limsup_{n\to\infty}\left\{L_n - L^* \geq \epsilon\right\}\right\} = 0, \quad \forall\, \mathcal{P}.$$

- Recall ***the plug-in rule*** of an estimated posterior conditional probability $\eta_n(x)$

  $$g_n(x) = \begin{cases} 0 & \eta_n(x) \leq \frac{1}{2} \\ 1 & \text{o.w.} \end{cases}$$

  Following Proposition 1.2, we have the following consistency results:

  Remark (**Error Estimate of Plug-In Rule, L^1 norm**) [Devroye et al., 2013]
  The **error probability** of the classifier $g_n(x)$ defined above satisfies the inequality

  $$L(g_n) - L^* \leq 2\int |\eta(x) - \eta_n(x)|\, \mu(dx) = 2\mathbb{E}\left[|\eta(X) - \eta_n(X)|\, |\mathcal{D}_n\right]$$

  where $\eta(x) = \mathcal{P}[Y = 1 | X = x]$ is the Bayes classifer.

  By Cauchy-Schwartz inequality, we have

Corollary 1.6 ***(Error Estimate of Plug-In Rule,*** L^2 ***norm)*** *[Devroye et al., 2013]*
If

$$g_n(x) = \begin{cases} 0 & \eta_n(x) \le \frac{1}{2} \\ 1 & o.w. \end{cases}$$

then its ***error probability*** *satisfies*

$$\begin{aligned} L(g_n) - L^* := \mathcal{P}_{X,Y}\left\{ g_n(X) \neq Y | \mathcal{D}_n \right\} - L^* &\le 2\sqrt{\int \left|\eta(x) - \eta_n(x)\right|^2 \mu(dx)} \\ &= 2\sqrt{\mathbb{E}\left[\left|\eta(X) - \eta_n(X)\right|^2 | \mathcal{D}_n \right]} \end{aligned} \tag{12}$$

Thus if we can show that under any distribution $\mathcal{P}_{X,Y}$, $\eta_n \to \eta$, i.e.

$$\mathbb{E}\left[\left|\eta(X) - \eta_n(X)\right|^2 | \mathcal{D}_n \right] \to 0, \quad \text{as } n \to \infty,$$

we will have ***strong universal consistency***.

- **Remark** (***Weak Convergence for Functions***)
Recall for a function η_n *converges to* η *weakly*, $\eta_n \overset{w}{\to} \eta$ if and only if

$$I(\eta_n) \to I(\eta), \quad \forall I \in \mathcal{H}^*$$

Note that for continuous function $\eta_n \in \mathcal{C}_c(\mathcal{X})$ with compact support on a locally compact Hausdorff space $\mathcal{X}$, the dual space is the space of regular Borel measures on X. In other words, $\eta_n \overset{w}{\to} \eta$ if and only if

$$\int \eta_n d\mathcal{P} \to \int \eta d\mathcal{P}, \quad \forall \mathcal{P} \in \mathscr{P}(\mathcal{X}),$$

which coorresponds to ***the strong consistency definition***.

1.4 No Free Lunch

- **Remark** There are some significant results known to the learning community
  - ***For every fixed*** n ***there exists a distribution where the classifier is arbitrarily bad***.
  For any $\epsilon > 0$ and any integer n and classification rule g_n, there exists a distribution of (X, Y) with Bayes risk $L^* = 0$ such that

  $$\mathbb{E}\left[L(g_n(\cdot|\mathcal{D}_n)) \right] \ge \frac{1}{2}.$$

  - ***Universal rate of convergence guarantees do not exist***. That is, for any rule,

  $$\liminf_{n\to\infty} \sup_{\forall \mathcal{P}_{X,Y}: L^* + \epsilon < 1/2} \mathcal{P}\left\{ L_n \ge L + \epsilon \right\} > 0$$

  Rate of convergence studies must involve certain subclasses of distributions of (X, Y).

  Moreover, *there exists* ***no universally consistent learning algorithm*** such that $L(g_n)$ converges ***uniformly*** over ***all distributions*** to L^*.

  - ***There exists no universally superior learning algorithm***. For every sequence of classification rules f_n, there is a *universally consistent sequence of classification rules* g_n such that for ***some distribution*** on $\mathcal{X} \times \mathcal{Y}$

$$L(f_n) > L(g_n), \quad \forall n > 0$$

- **Remark** In summary, there are two issues:

  1. ***No Restriction on Function Class*** $\mathcal{H}$, i.e. *convergence to* ***Bayes risk*** L^*, i.e. the infimum generalization error *for* ***all possible functions***.
  2. ***No Restriction on Underling Distribution*** $\mathcal{P}_{X,Y}$, i.e. be ***universally consistent*** *for* ***all possible distribution*** $\mathcal{P}_{X,Y}$.

  On the other hand,

  1. ***Restriction*** *of the class of* ***distributions*** on $\mathcal{X} \times \mathcal{Y}$ can lead to *convergence rates to Bayes risk* L^* for ***universally consistent*** *learning algorithms.*

     Problem: Assumptions cannot be tested since $\mathcal{P}_{X,Y}$ is unknown. Performance guarantees are only valid under the made assumptions.

  2. ***Restriction*** *of the* ***function class*** may lead to *no universal consistency* possible.

     Problem: Comparison to the best possible function in the class is possible *uniformly over all distributions.* But ***no performance guarantees*** with respect to ***the Bayes risk***.

1.5 Empirical Risk Minimization

- **Definition** (***Empirical Error/ Risk***)
  Given the data $\mathcal{D}_n$, the ***training error*** or the ***empirical error/risk*** of a hypothesis $g \in \mathcal{H}$ is defined as

$$\begin{aligned} \widehat{L}_n(g) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{g(X_i) \neq c(X_i)\right\}, && \text{(deterministic setting);} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{g(X_i) \neq Y_i\right\}, && \text{(stochastic setting).} \end{aligned}$$

- **Remark** Not to be confused with $L_n(g) := L(g_n) = \mathcal{P}_{X,Y}\left\{g_n(X) \neq Y\right\}$, where the subscript n indicates the dependency of g on $\mathcal{D}_n$.

- **Remark** (***Optimal Rule within A Subclass of Functions***)
  Given a subset of functions/concepts $\mathcal{H} \subset \mathcal{C}$, the ***best possible error probability*** by

$$L = \inf_{g \in \mathcal{H}} L(g). \Rightarrow g^* \in \text{argmin}_{g \in \mathcal{H}} L(g)$$

  Note that $L \geq L^*$. The optimal error rate L is a function of $\mathcal{P}_{X,Y}$ and $\mathcal{H}$.

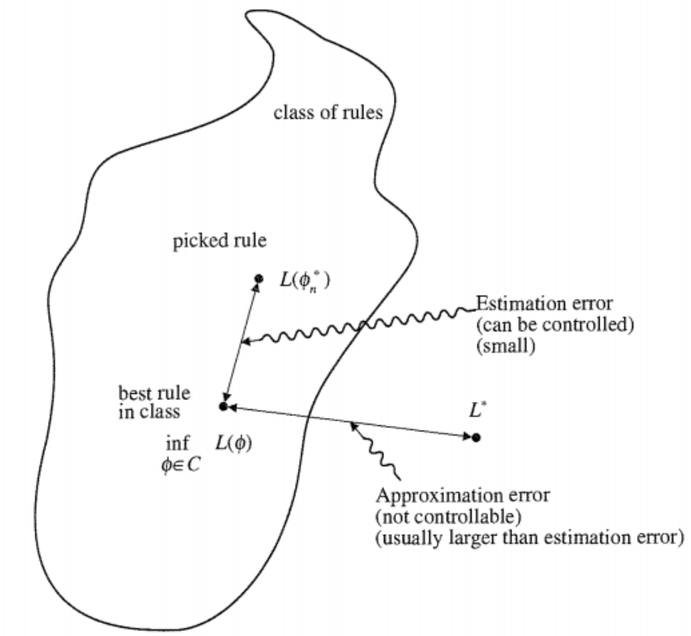

Figure 1: Estimation Error vs. Approximation Error [Devroye et al., 2013]

- **Remark** (***Optimal Rule under Empirical Error Probability***)
  Given a subset of functions/concepts $\mathcal{H} \subset \mathcal{C}$, the ***empirically optimal rule*** g_n^* is given by

$$g_n^* := g_n^*(\cdot|\mathcal{D}_n) \in \arg\min_{g\in\mathcal{H}} \widehat{L}_n(g).$$

- **Remark** (***Estimation Error*** *vs.* ***Approximation Error***)
  Their difference is the quantity that primarily interests us:

$$L(g_n^*) - L := L(g_n^*) - \inf_{g\in\mathcal{H}} L(g)$$

Note that both quantites are generalization error not training error. To compare with *Bayes error*, we have the following decomposition

$$L(g_n^*) - L^* = \left(L(g_n^*) - \inf_{g\in\mathcal{H}} L(g)\right) + \left(\inf_{g\in\mathcal{H}} L(g) - L^*\right).$$

1. The first difference term is called ***the estimation error***;
2. the second difference term is called ***the approximation error***. This latter term may be bounded in a ***distribution-free manner***, and *a rate of convergence results that* ***only depends on the structure of*** $\mathcal{H}$.

When the sub-class of functions $\mathcal{H}$ is ***large***, $L = \inf_{g\in\mathcal{H}} L(g)$ may be close to L^*, but the former error, *the estimation error*, is probably *large* as well. If $\mathcal{H}$ is ***too small***, there is no hope to make the approximation error small.

In empirical risk minimization, the subclass $\mathcal{H}$ is ***fixed***, and we have to live with the functions in $\mathcal{H}$. *The best we may then hope for is to minimize* $L(g_n^*) - \inf_{g\in\mathcal{H}} L(g)$.

- **Remark** (***Overfitting***)
  If $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ is the class of all (measurable) decision functions, then we can always find a classifier in $\mathcal{H}$ with ***zero empirical error***, but it may have ***arbitrary values*** *outside of the points* $X_1, \ldots, X_n$. For example, an *empirically optimal classifier* is

$$g_n^*(x) = \begin{cases} Y_i & x = X_1, \ldots, X_n \\ 0 & \text{otherwise} \end{cases}$$

  This is clearly not what we are looking for. This phenomenon is called ***overfitting***, as the overly large class $\mathcal{H}$ *overfits* the data.

References

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.