Lecture 1: Fundamental Concept of Statistical Learning

Tianpei Xie

Jul. 25th., 2015

Contents

1 Fundamental Concepts and Assumptions

1.1 Categories of Machine Learning

- **Remark** (***Categories of Machine Learning***)
  The field of machine learning has branched into several subfields dealing with different types of learning tasks. We give a rough taxonomy of learning paradigms, aiming to provide some perspective of where the content of this book sits within the wide field of machine learning.

- **Remark** (***Supervised Learning***, ***Unsupervised Learning*** *and* ***Reinforcement Learning***)

  1. **Supervised learning**. Learning to ***predict*** and ***generalize***. In other words, the task of learning is to infer a mapping between covariates (data) and responses (target) so that the error/cost is minimized. Each example is a description of situation (sample) together with a specification (label) of the correct action the system should take to that action. Supervised learning is an **error-correction** process. It is an one-step prediction. Human label is used as ***instructive feedbacks***.

  2. **Unsupervised learning**. Learning to ***represent*** and ***discover*** the ***hidden structure*** of data. A proper representation of data facilitates knowledge discovery and improves the performance of prediction. It also helps in visualization, storage and communication.

  3. **Reinforcement learning**. Learning from ***interaction***. As compared to above approaches, reinforcement learning studies *goal-directed learning from interaction*. The term 'learning' means learning to map *situations* to *actions* so as to maximize the reward. Also it is often unrealistic to obtain all examples of desired behavior that are both correct and representative of all situations in which the agents have to act. Reinforcement learning is a ***trial-and-error*** process and it is a multi-step prediction. It cares about future rewards in multiple steps. On the other hand, reinforcement learning only optimize the future rewards, where the rewards are used as ***evaluative feedbacks***.

- **Remark** (***Active Learning***, ***Passive Learning***)
  Learning paradigms can vary by ***the role*** *played by the* ***learner***. We distinguish between "*active*" and "*passive*" learners.

  1. An ***active*** learner interacts with the environment at training time, say, by posing queries or performing experiments;

  2. A ***passive*** learner only observes the information provided by the environment (or the teacher) without influencing or directing it.

- These notes are mainly about ***supervised learning*** *tasks.*

1.2 Data, Concept and Hypotheses

- **Remark** (***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

$$\mathcal{D} \equiv \mathcal{D}_n = \left((X_1, Y_1), \ldots, (X_n, Y_n) \right).$$

Note that $\mathcal{D}_n$ is a finite ***sub-sequence*** in $(\mathcal{X} \times \mathcal{Y})^n$.

- **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***. That is, $\mathcal{C} \subseteq \{c : \mathcal{X} \to \mathcal{Y}\} = \mathcal{Y}^{\mathcal{X}}$. Concept class $\mathcal{C}$ is a *function class*.

- **Definition** (***Hypothesis and Hypothesis Class***)
  The learner is requested to output a *prediction rule*, $h : \mathcal{X} \to \mathcal{Y}$. This function is also called a ***predictor***, a ***hypothesis***, or a ***classifier***. The predictor can be used to predict the label of new domain points.

  Note that $\mathcal{H}$ and $\mathcal{C}$ may not overlap, since the concept class is unknown to learner.

1.3 Deterministic and Stochastic Scenario

Learning is formalized into two different *scenarios*:

1. ***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, let $X = (X_1, \ldots, X_n)$ be a sequence of n ***independent identically distributed*** *(i.i.d.)* ***random samples***, where $X_i \sim \mathcal{P}_X$. Then $Y_i = c(X_i)$ for $i = 1, \ldots, n$ and the sample set

   $$\mathcal{D} \equiv \mathcal{D}_n = \left((X_1, Y_1), \ldots, (X_n, Y_n) \right) \equiv \left((X_1, c(X_1)), \ldots, (X_n, c(X_n)) \right).$$

   Note that the probability distribution for $\mathcal{D}_n$ is $\mathcal{P}_X^n := \otimes_{i=1}^n \mathcal{P}_X$.

2. ***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, the sample set

   $$\mathcal{D} \equiv \mathcal{D}_n := \left((X_1, Y_1), \ldots, (X_n, Y_n) \right).$$

   is a collection of n ***independent identically distributed*** *(i.i.d.)* ***random sample pairs*** $(X_i, Y_i) \sim \mathcal{P}_{X,Y}$ Note that the probability distribution for $\mathcal{D}_n$ is $\mathcal{P}_{X,Y}^n := \otimes_{i=1}^n \mathcal{P}_{X,Y}$.

- **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*** $h \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ based on $\mathcal{D}$. The task of ***supervised learning*** is to minimize ***the generalization error*** given a set of training data $\mathcal{D}_n$.

- **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 $h \in \mathcal{H}$ 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 $h \in \mathcal{H}$.

  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 $\mathcal{P}_{X,Y}(x, y) > 0$.

1.4 Generalization Error

- We define ***the error of a classifier*** to be the probability that it does not predict the correct label on a random data point generated by the aforementioned underlying distribution.

  Definition (***Generalization Error in Deterministic Scenario***) [Mohri et al., 2018]
  Under *a deterministic scenario*, ***generalization error*** or the ***risk*** or simply ***error*** for the *classifier* $h \in \mathcal{H}$ is defined as

  $$L(h) \equiv L_{\mathcal{P},c}(h) = \mathcal{P}\left\{h(X) \neq c(X)\right\} \equiv \mathbb{E}_X\left[\mathbb{1}\left\{h(X) \neq c(X)\right\}\right] \tag{1}$$

  with respect to the concept $c \in \mathcal{C}$ and *the feature distribution* $\mathcal{P} \equiv \mathcal{P}_X$.

- Similarly, under the stochastic scenario, the joint distribution $\mathcal{P}_{X,Y}$ has replaced the concept class c:

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

  $$L(h) \equiv L_{\mathcal{P}}(h) = \mathcal{P}\left\{h(X) \neq Y\right\} \equiv \mathbb{E}_{X,Y}\left[\mathbb{1}\left\{h(X) \neq Y\right\}\right] \tag{2}$$

  with respect to the joint distribution $\mathcal{P} \equiv \mathcal{P}_{X,Y}$.

- **Remark** Under both situations, *the generalization error* $L(h)$ is a ***functional*** on both the ***hypothesis*** g and *the underlying* ***data generating process***, defined by distribution $\mathcal{P}$.

1.5 Empirical Risk Minimization

- **Remark** (***Assumption***)
  The learner is blind to the underlying distribution $\mathcal{P}$ over the world and to the labeling concept c.

- Since the learner does not know what $\mathcal{P}$ and c are, *the generalization error* is not directly available to the learner. A useful notion of error that can be calculated by the learner is *the training error* or *empirical error*:

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

  $$\widehat{L}(h) \equiv \widehat{L}_{\mathcal{D}_n}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{h(X_i) \neq Y_i\right\} = \frac{1}{n}\left|\{i : h(X_i) \neq Y_i\}\right| := \widehat{\mathbb{E}}\left[\mathbb{1}\left\{h(X) \neq Y\right\}\right]$$

  where either $Y = c(X)$ or Y is a random variable associated with X.

- **Definition** (***Empirical Risk Minimization (ERM)***)
  The Empirical Risk Minimization (ERM) is a learning paradigm that aims at finding the optimal classifer $h \in \mathcal{H}$ by ***minimizing the empirical error*** $\widehat{L}_{\mathcal{D}}(h)$; i.e.

  $$h^* \equiv h^*_{\mathcal{D}} \in \arg\min_{h\in\mathcal{H}} \widehat{L}_{\mathcal{D}}(h)$$

  We denote $h_{\mathcal{D}} \in \mathrm{ERM}(\mathcal{D})$ as the hypothesis returned by ERM learning algorithms given samples $\mathcal{D}$.

- **Remark** (***Empirical Risk Minimization with Inductive Bias***)
  The hypothesis class $\mathcal{H}$ is usually chosen as a subset of concept class $\mathcal{C}$. Thus *the empirical risk minimization (ERM)* may lead to a *biased* predictor $h \in \mathcal{H}$ that is not optimal in $\mathcal{C}$. Such restrictions are often called ***an inductive bias***. Since *the choice of such a restriction* $\mathcal{H}$ is determined ***before*** the learner *sees the training data*, it should ideally be based on some ***prior knowledge*** about the problem to be learned.

1.6 Bayes Error

- **Definition** (***Bayes Error***)
  Under a given distribution $\mathcal{P}$, the ***Bayes error*** or ***Bayes risk*** is defined as

  $$L^* = \inf_h \left\{L_{\mathcal{P}}(h)\right\}, \tag{3}$$

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

- **Remark** (***Bayes Error as Functional of Distribution***)
  Bayes Error** is **a functional of underlying distribution** $\mathcal{P}$ **only and it does not depend on choice of function h or function class $\mathcal{H}$.

  $$L^* \equiv L^*(\mathcal{P}).$$

Intuitively, it reflects *how hard the learning problem* is ***intrinsically*** since it determines ***the lower bound*** *of generalization error.*

- **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

$$\begin{aligned} \eta(x) &= \mathcal{P}\left\{Y = 1 \middle| X = x\right\} \\ &= \mathbb{E}_{p(y|x)}\left[y \middle| X = x\right]. \end{aligned} \tag{4}$$

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

$$\begin{aligned} h^*(x) &= \text{argmax}_{y\in\{0,1\}} P(Y|X = x) \\ &= \begin{cases} 1 & \eta(x) > \frac{1}{2} \\ 0 & \text{o.w.} \end{cases} \end{aligned} \tag{5}$$

  with the corresponding ***Bayes error***

$$\begin{aligned} L^* &= \mathbb{E}_{P(X)}\left[\min\left\{P(Y = y|X) \mid y \in \{0, 1\}\right\}\right] \\ &= 1 - \mathbb{E}_{P(X)}\left[\eta(X)\mathbb{1}\left\{\eta(X) > 1/2\right\} + (1 - \eta(X))\mathbb{1}\left\{\eta(X) \leq 1/2\right\}\right] \end{aligned} \tag{6}$$

- 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*

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

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

$$\mathcal{P}\left\{h^*(X) \neq Y\right\} \leq \mathcal{P}\left\{h(X) \neq Y\right\}$$

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

$$\begin{aligned} &\mathcal{P}\left\{h(X) \neq Y|X = x\right\} \\ &= 1 - \mathcal{P}\left\{Y = h(X)|X = x\right\} \\ &= 1 - \left(\mathcal{P}\left\{Y = 1, h(X) = 1|X = x\right\} + \mathcal{P}\left\{Y = 0, h(X) = 0|X = x\right\}\right) \\ &= 1 - \left(\mathbb{1}\left\{h(X) = 1\right\}\mathcal{P}\left\{Y = 1|X = x\right\} + \mathbb{1}\left\{h(X) = 0\right\}\mathcal{P}\left\{Y = 0|X = x\right\}\right) \\ &= 1 - \left[\mathbb{1}\left\{h(X) = 1\right\}\eta(x) + \mathbb{1}\left\{h(X) = 0\right\}(1 - \eta(x))\right] \end{aligned} \tag{7}$$

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

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

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

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

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

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 $h(X)$, we have

$$\mathcal{P}\{h(X) \neq Y\} - L^* = 2\int_{\mathcal{X}} |\eta(x) - 1/2|\, \mathbb{1}\{h(X) \neq h^*(x)\}\, \mu(dx) \tag{8}$$

and

$$\begin{aligned}
\mathcal{P}\{h(X) \neq Y\} - L^* &\leq 2\int_{\mathcal{X}} |\eta(x) - \tilde{\eta}(x)|\, \mu(dx) \\
&= 2\mathbb{E}_{p(X)}\left[\eta(X) - \tilde{\eta}(X)\right]
\end{aligned} \tag{9}$$

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

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

Otherwise, $h(X) \neq h^*(x)$, then

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

Thus

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

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

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

$$h(X) = \mathbb{1}\{\tilde{\eta}_1(x) > \tilde{\eta}_0(x)\},$$

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 $h(X)$, we have

$$\mathcal{P}\{h(X) \neq Y\} - L^* \leq \int_{\mathcal{X}} |(1-\eta(x)) - \tilde{\eta}_0(x)|\, \mu(dx) + \int_{\mathcal{X}} |\eta(x) - \tilde{\eta}_1(x)|\, \mu(dx) \tag{10}$$

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 $\mathcal{P}\{Y=1\} = q$ and $\mathcal{P}\{Y=0\} = 1-q$ 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

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

- **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

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

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

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

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

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

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

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

- **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. $\blacksquare$

1.7 Approximation Error and Estimation Error

- **Remark** (***Optimal Rule within Hypothesis Class***)
  Given a hypothesis class $\mathcal{H}$, the ***best possible generalization error*** is obtained by

$$L \equiv L_{\mathcal{P},\mathcal{H}} = \inf_{h\in\mathcal{H}} L_{\mathcal{P}}(h) = \inf_{h\in\mathcal{H}} \mathcal{P}\left\{h(X) \neq Y\right\}$$

  The optimal rule within $\mathcal{H}$ is defined as $h^*_{\mathcal{H}} \in \arg\inf_{h\in\mathcal{H}} L_{\mathcal{P}}(h)$. Note that $L \geq L^*$. *The optimal error rate* $L_{\mathcal{P},\mathcal{H}}$ is a function of $\mathcal{P}$ and $\mathcal{H}$.

- **Remark** (***Optimal Rule under Empirical Error Probability***)
  Given a hypothesis class $\mathcal{H}$, the ***empirically optimal rule*** $h^*_{\mathcal{D}}$ is given by

$$h^*_{\mathcal{D}} \in \arg\min_{h\in\mathcal{H}} \widehat{L}_{\mathcal{D}}(h).$$

  Since the function $h^*_{\mathcal{D}}$ is determined by data $\mathcal{D}$, it can be seen as a ***random function***

$$h^*_{\mathcal{D}}(x) := h(x|\mathcal{D}) \equiv h(x|((X_1, Y_1), \ldots, (X_n, Y_n))).$$

  In general, we use the notation $h_{\mathcal{D}}$ for a hypothesis returned by a learning algorithm given data $\mathcal{D}$.

- **Remark** (***Estimation Error*** *vs.* ***Approximation Error***)
  The difference between ***the generalization error*** for the *optimal rule* under ERM and ***the optimal generalization error*** *within a hypothesis class* $\mathcal{H}$ is called ***<u>the excess risk</u>***. It is the quantity that primarily interests us:

$$L(h^*_{\mathcal{D}}) - L := L(g^*_n) - \inf_{h\in\mathcal{H}} L(h)$$

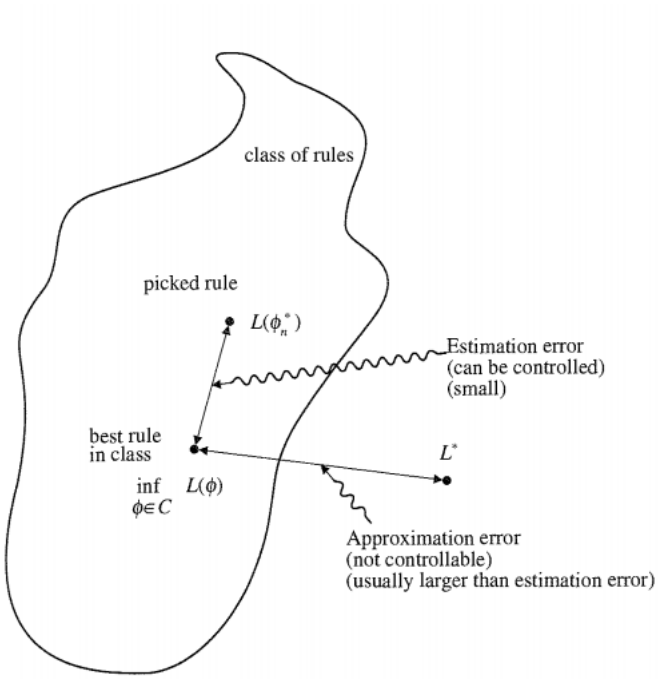

Figure 1: Estimation Error vs. Approximation Error [Devroye et al., 2013]

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

$$L(h_{\mathcal{D}}^*) - L^* = \left(L(h_{\mathcal{D}}^*) - \inf_{h \in \mathcal{H}} L(h) \right) + \left(\inf_{h \in \mathcal{H}} L(h) - L^* \right).$$

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***, $L = \inf_{h \in \mathcal{H}} L(h)$ 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* $L(h_{\mathcal{D}}^*) - \inf_{h \in \mathcal{H}} L(h)$.

- **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

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

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

- **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*.

1.8 Realizability Assumptions

- **Definition** (***Realizability Assumption***)
  There exists a hypothesis $h \in \mathcal{H}$ such that ***the generalization error is zero***.

$$\exists h \in \mathcal{H}, \quad L_{\mathcal{P},c}(h) = 0$$

  Note that this assumption implies that ***with probability*** 1 over random samples, $\mathcal{D}$, where the instances of $\mathcal{D}$ are sampled according to $\mathcal{P}$ and are labeled by c, we have $\widehat{L}(h) = 0$

- **Remark** (***Implications***)
  There are several implications behind ***the realizability assumption***:

  1. $c \in \mathcal{H}$; *The realizability assumption* implies that the concept function c ***is in the hypothesis class*** $\mathcal{H}$. In other words, this assumption states that ***our prior knowledge*** *on the problem* is ***sufficient*** to completely solve the problem.
  2. $h^*_{Bayes} = c$. ***The Bayes optimal classifier*** h^*_{Bayes} is equal to ***the concept function*** c.
  3. ***The approximation error is zero***. To achieve Bayes optimality, one only need to ***minimize the generalization error itself***

$$\inf_{h \in \mathcal{H}} L(h) = L^* = 0.$$

  4. $\widehat{L}(h_{\mathcal{D}}) = 0$ for $h_{\mathcal{D}} \in \mathrm{ERM}(\mathcal{D})$. ***Every ERM hypothesis has zero training error***.

- **Remark** (***Failure under Realizability Assumption***)
  Since the optimal error rate is zero under the realizability assumption, a hypothesis $h \in \mathcal{H}$ with error rate above a certain threshold is considered *a* ***failure*** *of the hypothesis*. If the hypothesis is returned by ERM, then it must have *zero training error*. Therefore, the failure of a hypothesis $h_{\mathcal{D}}$ due to choice of a set of ***bad samples***:

$$\left\{\mathcal{D} \sim \mathcal{P}^n : \exists h \in \mathcal{H}, L_{\mathcal{P},c}(h) > \epsilon \wedge \widehat{L}_{\mathcal{D}}(h) = 0\right\} = \bigcup_{h \in \mathcal{H}_B} \left\{\mathcal{D} \sim \mathcal{P}^n : \widehat{L}_{\mathcal{D}}(h) = 0\right\}$$

  where $\mathcal{H} := \{h \in \mathcal{H} : L_{\mathcal{P},c}(h) > \epsilon\}$ is the set of failure hypotheses.

- **Remark** The realizability assumption is only valid under ***the deteriminstic scenario*** when the Bayes error $L^* = 0$.

  Under ***the stochastic scenario***, when the Bayes error $L^* > \epsilon$, it is impossible to achieve zero generalization error.

2 Asymptotic Analysis of Learning Algorithm

Question (***Low Training Error*** $\Rightarrow$ ***Low Generalization Error ?***)
For a given sequence of data $\mathcal{D}$, the optimal solution $h^*_{\mathcal{D}}$ from ERM minimizes ***the empirical error***, but does it minimize ***the generalization error*** ?

The answer to this question is the motivation behind the analysis of *learning algorithms* and *learning paradigm*. To see why this is complicated, we note that for fixed hypothesis h

$$\widehat{L}_{\mathcal{D}_n}(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{h(X_i) \neq Y_i\right\} := \frac{1}{n}\sum_{i=1}^{n} f_h(X_i)$$
$$L_{\mathcal{P},c}(h) = \mathbb{E}_{\mathcal{P}}\left[\mathbb{1}\left\{h(X) \neq c(X)\right\}\right] := \mathbb{E}\left[f_h(X)\right]$$

Then the absolute deviation of empirical error from generalization error, uniformly over a class of functions $f_h \in \mathcal{F}$ can be formulated as

$$\left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|_{\infty} := \sup_{f_h \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f_h(X_i) - \mathbb{E}\left[f_h(X)\right]\right|.$$

The answer of our question is equivalent to ask when $n \to \infty$, if $\left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|_{\infty} \to 0$ *in probability*. This is so-called ***asymptotic analysis***.

2.1 Consistency Definitions

- **Remark** Since the returned optimal solution from ERM $h^*_{\mathcal{D}}$ is determined by data $\mathcal{D}$, it can be seen as a ***random function***

$$h^*_{\mathcal{D}_n}(x) := h(x|\mathcal{D}_n) \equiv h(x|((X_1, Y_1), \ldots, (X_n, Y_n))).$$

  We simply noted $h^*_{\mathcal{D}_n}$ as h_n to emphasize its dependency on the ***size*** of data.

- **Definition** (***Classification Rule***) [Devroye et al., 2013]
  A *(supervised) learning* process is the process of constructing $h_n \in \mathcal{H}$ from data $\mathcal{D}_n$ and a sequence of classifiers (functions) $\{h_n : n \geq 1\}$ is called ***a (classification) rule***.

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

$$L_n \equiv L_{\mathcal{P}}(h_n) = \mathcal{P}\left\{h_n(X) \neq Y\right\} \xrightarrow{\mathcal{P}} L^*, \text{ as } n \to \infty$$

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

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

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

$$L_n \to L^* \ \ a.s.$$

- **Remark** (***Consistency Within Hypothesis Class*** $\mathcal{H}$)
  For given hypothesis class $\mathcal{H}$, it is ***the optimal error rate within*** $\mathcal{H}$ that are concerned about instead of Bayes error. Thus, we can modify the ***consistency*** definition to become

$$(\textbf{\textit{consistent rule}}) \qquad L_n \xrightarrow{\mathcal{P}} L \equiv \inf_{h\in\mathcal{H}} L(h), \ \text{ as } n \to \infty$$

$$(\textbf{\textit{strong consistent rule}}) \qquad L_n \xrightarrow{a.s.} L \equiv \inf_{h\in\mathcal{H}} L(h), \ \text{ as } n \to \infty$$

- **Remark** (***Generalization Error as Random Variable***)
  In the definition of consistency, one ***assume*** that the function $h_n := h^*_{\mathcal{D}_n}$ from some algorithm given $\mathcal{D}_n$. Since $\mathcal{D}_n$ is random, h_n is ***random*** *not deterministic.* Therefore, *the* ***consistency*** *condition can be written explicitly as*

$$\lim_{n\to\infty} \mathcal{P}_{\mathcal{D}_n} \{L_n - L \geq \epsilon\} = 0, \quad \forall \epsilon > 0,$$

  where L_n is a ***random variable*** depending on $\mathcal{D}_n$

$$L_n \equiv L_{\mathcal{P}}(h_n) := \mathbb{E}\left[\mathbb{1}\left\{h_{\mathcal{D}_n}(X) \neq Y\right\} | \mathcal{D}_n\right] = \mathcal{P}\left\{h_{\mathcal{D}_n}(X) \neq Y | \mathcal{D}_n\right\}.$$

- **Remark** Since samples are i.i.d., $\mathcal{P}(\mathcal{D}_n) = \mathcal{P}^n$ is a product measure.

- **Remark** ***A consistent rule*** $\{h_n\}$ 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*** $\mathcal{P}$, i.e.

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

  and

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

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

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

  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 $h_n(x)$ defined above satisfies the inequality

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

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

By Cauchy-Schwartz inequality, we have

Corollary 2.1 ***(Error Estimate of Plug-In Rule, L^2 norm)*** *[Devroye et al., 2013] If*

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

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

$$\begin{aligned} L(h_n) - L^* := \mathcal{P}_{X,Y}\left\{h_n(X) \neq Y | \mathcal{D}_n\right\} - L^* &\leq 2\sqrt{\int |\eta(x) - \eta_n(x)|^2 \, \mu(dx)} \\ &= 2\sqrt{\mathbb{E}\left[|\eta(X) - \eta_n(X)|^2 \, | \mathcal{D}_n\right]} \end{aligned} \tag{11}$$

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

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

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

2.2 No Free Lunch

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

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

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

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

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

    Moreover, *there exists* ***no universally consistent learning algorithm*** such that $L(h_n)$ converges ***<u>uniformly</u>*** over ***<u>all distributions</u>*** 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* h_n such that for ***<u>some distribution</u>*** on $\mathcal{X} \times \mathcal{Y}$

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

3 Development Paths of Learning Algorithms

- **Remark** *The No-Free-Lunch theorem* reveals that it is hopeless to find an optimal classifier if we have

  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. ***And, No Restriction on Underling Distribution*** $\mathcal{P}$, 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 $\mathcal{P}$ 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***.

- **Remark** (***Development Paths of Learning Algorithms***)

  1. (***Learning Distribution***): The first path is to *relax* the requirement for ***universal learnability***:

     "*Instead of learning a classifier* ***universially*** *good for* ***all possible distributions****, our goal is to learn an optimal classifier for* ***a given famility of distributions***"

     In this setting, we ***assume*** that $\mathcal{P}$ is *from* ***a family of probability distributions*** $\mathscr{P}$. The ***goal*** of learning algorithm is to find *the optimal distribution* $\mathcal{P} \in \mathscr{P}$ that is as close to *the true distribution* as possible. If the approximation is good, the corresponding *conditional distribution* η will approach to the ***Bayes classifier***.

     (a) (***Parameter Estimation***): Assume that the familiy of distributions $\mathscr{P}$ is indexed by some ***finite dimensional parameters***, i.e. $\mathscr{P} := \mathscr{P}_{\theta \in \Theta}$ is a *parametric family of distributions*. The learning task becomes the task of ***parameter inference*** from data $\mathcal{D}$. Many classical statistical analysis tools are useful here.

     (b) (***Distribution Approximation***): Another way is to assume that $\mathscr{P} := \mathscr{P}_{\mathcal{H}}$ is indexed by functions in $\mathcal{H}$. This corresponding to the ***non-parametric setting***. The goal 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*.

     Depending on the distribution of concern is ***the joint distribution*** $\mathcal{P}_{X,Y}$ or ***the conditional distribution*** $\mathcal{P}_{Y|X}$, we can have ***generative models*** or ***discriminative models***. If the assumption is true that the underlying distribution $\mathcal{P} \in \mathscr{P}$, we have guarantee to reach to ***the Bayes optimality***.

2. (***Learning Functions***): An alternative path is to ***restrict the hypothesis class of functions*** $\mathcal{H}$ while *not abandoning* the ***universal learnability***. As a compromise, we abandon *the goal of reaching* ***Bayes optimiality***:

   "*Our goal is to learn an optimal classifier from* ***a given class of functions*** *that is* ***universially*** *good for all possible distributions*"

   In this setting, we ***assume*** that $\mathcal{H}$ is ***a class of functions that are not "too large"***, i.e. ***its expressive power is limited*** so that inference from finite number of samples is possible.

- **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*** $h : \mathcal{X} \to \mathcal{Y}$ is a function (hypothesis) in $\mathcal{H}$ from input to output. The task of learner is to find $h \in \mathcal{H}$ 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

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

    is close to the Bayes classifier

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

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

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.