Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter
Scribed by Douglas Fearing

Lecture 2
Th 2/8

1 Introduction to Learning

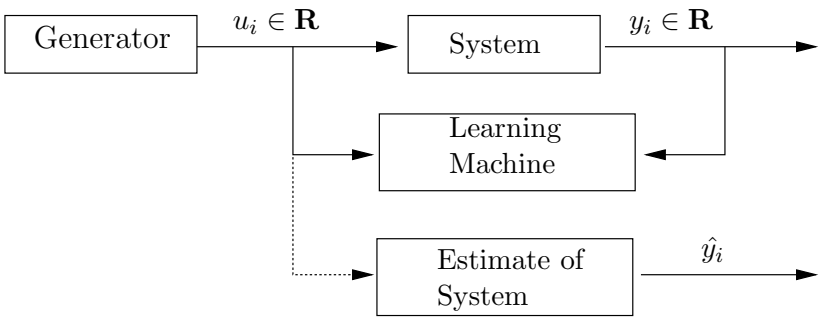


Figure 1: Learning system diagram.

1.1 Ingredients of a Learning Problem

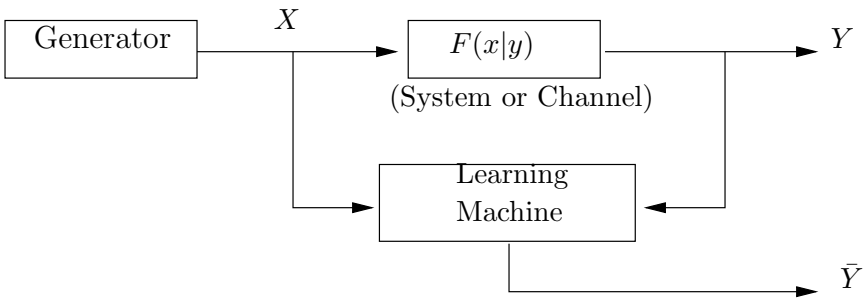
1. **Data** $\{(x_i, y_i),\ 1 \leq i \leq \ell\}$ may be derived in multiple fashions
   - Data composed of i.i.d. samples from a joint distribution $\sim F(x, y)$
   - Data based on a underlying dynamic function $y_i = f(x_i, x_{i-1}, \ldots, x_{i-m}, w)$
   - Data derived from a hidden Markov model (HMM)
2. **Model Set** of alternative models (e.g. i.i.d., linear regression models, or HMMs)
3. **Algorithm** for choosing the best model from the model set based on the data (i.e. $\mathcal{A} : \{(x_i, y_i)\} \rightarrow$ Model Set)
   - **ex**: Empirical Risk Minimization
     Model Set: $\{f(x, \alpha),\ \alpha \in \Lambda\}$
     Empirical Risk: $R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)|^2$
     Algorithm: $\hat{\alpha}^{\ell} = \text{argmin}_{\alpha}\, R_{emp}(\alpha)$
4. **Error Criterion** for evaluation

1.2 Questions We Might Ask

1. Learnability? {consistency; limiting behavior}
2. Rates of convergence? {Entropy of a model set; VC-dimension; asymptotic distribution}
3. Model quality evaluation / Model set selection?
   - Universal coding (MDL)
   - Structured risk minimization
4. Algorithms and computation?
   - Batch vs. recursive

2 Risk Minimization

2.1 Learning System Diagram



1. **Generator** selects the distribution function for the random variable X (i.e. $\sim F(x)$)
   - Values for the random variable X may be produced randomly by the generator or sampled / controlled by the learning machine
2. **System** or **Channel** produces a value for the random variable Y based on the conditional distribution $\sim F(y|x)$, where X and Y are jointly distributed according to $F(x, y)$
3. **Data** are combined from the generator and the channel, $z^\ell = \{(x_i, y_i),\ 1 \leq i \leq \ell\}$
   - Data are i.i.d. samples $\sim F(x, y)$
   - Data are available to the learning machine, though we may only have partial information
4. **Learning Machine** utilizes the data to select a model that produces $\bar{Y}$

2.2 Risk-Based Learning

For this problem, our model class, $\mathcal{C} = \{f(x, \alpha),\ \alpha \in \Lambda\}$, is defined by the singleton function $f(\cdot, \cdot)$ and a set of parameterizations $\alpha \in \Lambda$. Thus, our learning question becomes what parameterization, $\hat{\alpha}$, to select to model the functional dependence of y on x (i.e. $\hat{y} = f(x, \hat{\alpha})$). Note that choosing an appropriate model class is referred to as the Model Selection problem, which we will deal with later in the course. For now, we assume that the model class is provided and we utilize the following function definitions.

- **Cost** $\{L : X \times Y \times \Lambda \to \mathbf{R}\}$ $L(x_i, y_i, \alpha)$ is the cost of selecting parameter α relative to observation (x_i, y_i) (i.e. comparing $f(x, \alpha)$ with y).
- **Risk** $\{R : \Lambda \to \mathbf{R}\}$ is the expected cost of selecting parameter α over the joint distribution of X and Y. Thus, $R(\alpha) \triangleq \int L(x, y, \alpha) dF(x, y)$, where $dF(x, y)$ is the derivative of the distribution function $F(X, Y)$.
- **Empirical Risk** $\{R_{emp} : \Lambda \to \mathbf{R}\}$ is the average cost of selecting parameter α over the observed data. Thus, $R_{emp}(\alpha) \triangleq \frac{1}{\ell} \sum_{i=1}^{\ell} L(x_i, y_i, \alpha)$.

The cost, $L(x, y, \alpha)$, is itself a random variable, and $L(x_i, y_i, \alpha)$ is independent for all i, since we assume the observations (x_i, y_i) are independent. By the Strong Law of Large Numbers, we therefore have that $R^\ell_{emp}(\alpha) \xrightarrow{\text{a.s.}} R(\alpha)$ for all α. Thus, for risk minimization, we try to select an α to minimize $R_{emp}(\alpha)$. This raises the question of whether $\min_\alpha R^\ell_{emp}(\alpha) \longrightarrow \min_\alpha R(\alpha)$. That is, do the pointwise convergence laws in probabibility ensure uniform convergence, and if so, under what conditions? We will now consider a few scenarios where this connection holds.

2.3 Classification / Pattern Recognition

Classification will often be the prototype example used for proofs due to its simplicity, though the same reasoning extends to more complex scenarios. For classification problems, we assume that $Y \in \{1, 2, \ldots, k\}$ with $Y = \phi(X)$. The function $\phi : \mathbf{R}^n \to \{1, 2, \ldots, k\}$ defines a partition of the X space (see Figure 2). Note that ϕ need not belong to the model class as we typically do not know the "true" relationship.

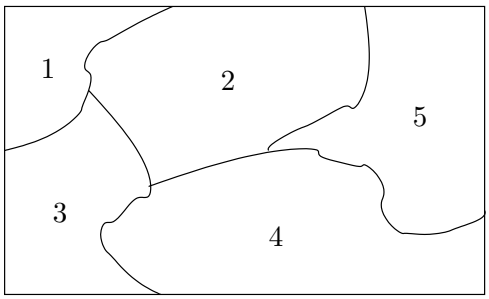


Figure 2: Sample data partition.

For a classification problem, how do we define the cost or risk? One intuitive approach is to "count" the number of misclassifications. For this approach, we define the cost function as follows.

$$L(x, y, \alpha) = \begin{cases} 0 & \text{if } y = f(x, \alpha) \\ 1 & \text{if } y \neq f(x, \alpha) \end{cases}$$

As above, our risk $R(\alpha) = \mathbf{E}[L] = \int L(x, y, \alpha) dF(x, y)$ and our empirical risk $R^{\ell}_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(x_i, y_i, \alpha)$. Thus, our empirical risk equals the number of misclassifications divided by the number of data observations, ℓ. For this problem, using $\min_{\alpha} R^{\ell}_{emp}(\alpha)$ to select the parameter is a reasonable choice.

- ex: ($\mathbf{R}^n = \mathbf{R}$)

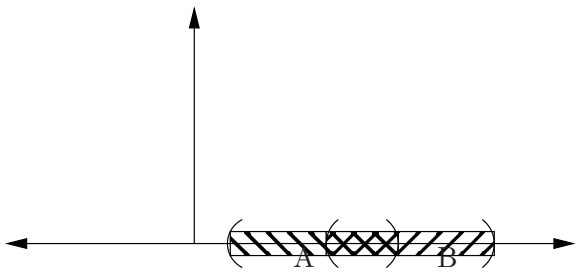


Figure 3: Region of misclassification.

Our (average) risk is the probability that a sample falls into a region of misclassification. That is, $R(\alpha) = \mathbb{P}$[symmetric difference of A and B] (see Figure 3).

2.4 Regression

For regression problems, we estimate Y as a deterministic continuous function of X. Note that we could instead start with $y = f(x, \alpha) + \beta w$, where W is a random variable, but the reasoning is similar. For our cost function, we let $L(x, y, \alpha) = |y - f(x, \alpha)|^2$, thus our risk becomes

$R(\alpha) = \int |y - f(x,\alpha)|^2 dF(x,y)$. Why do we call this regression? As you may have seen in another course, we have that the conditional mean, $\mathbf{E}[Y|X] = \text{argmin}_g \, \mathbf{E}[|y - g(x)|^2]$. That is, the conditional mean minimizes the mean-squared error across all functions. Thus, we let our regression function $r(x) = \text{argmin}_g \, \mathbf{E}[|y - g(x)|^2]$.

Since our model class does not include all functions, we would like to be sure that our parameterization approximates the true regression function. This is in the fact the case, which we show as follows.

- Claim: $\int |y - f(x,\alpha)|^2 dF = \int |y - r(x)|^2 dF + \int |r(x) - f(x,\alpha)|^2$ dF
- Proof:

$$\begin{aligned}\int |y - r(x) + r(x) - f(x,\alpha)|^2 dF &= \int |y - r(x)|^2 dF + \int |r(x) - f(x,\alpha)|^2 dF \\ &\quad + \int [y - r(x)][r(x) - f(x,\alpha)] dF\end{aligned}$$

  We leave it as an exercise to show that $[y - r(x)][r(x) - f(x,\alpha)] = 0$.

Thus, for our regression problem we again minimize the empirical risk, $R^\ell_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - f(x_i,\alpha)|^2$. We still need to show convergence of $R^\ell_{emp}(\alpha)$ to $R(\alpha)$...

2.5 Density Estimation

For density estimation, we only observe X, thus our data is $z^\ell = \{x_1, \ldots, x_l\}$. We could have considered this problem under one of the two previous scenarios with the functional relationship being the identity map. We assume that each independent observation is distributed according to a true, but unknown, probability density function $p^*(x) = \frac{dF^*(x)}{dx}$. Thus, our model class, $C = \{p(x,\alpha),\ \alpha \in \Lambda\}$, defines a set of parameterized probability density functions to choose from. In this case, our cost function is $L(x,y,\alpha) = -\log p(x,\alpha)$, which is referred to as the negative log-likelihood. As expected, our risk $R(\alpha) = -\int p^*(x) \log p(x,\alpha)$ and our empirical risk $R^\ell_{emp}(\alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \log p(x_i,\alpha)$. As before, we minimize our empirical risk in order to choose the optimal parameterization.

2.6 Elements of Information Theory

Entropy. Assume we have $X \in \{1, 2, \ldots, M\} \sim p(x)$, where we use the shorthand p_i for $p(i)$. We define the Entropy of the distribution $H(p) \triangleq -\sum_{i=1}^{M} p_i \log p_i$, though some texts will use the notation $H(X)$. Note that entropy is bounded, that is $H(p) \geq 0$ and $H(p) \leq \log M$. Entropy is a very useful concept in that it provides the fundamental limit of how small a compression can be.

For the continuous case, $X \in \mathbf{R}$, we refer to the Differential Entropy. Differential Entropy is defined as $\mathbf{E}_p[-\log p(x)]$, which equals $-\int p(x) \log p(x) dx$.

Kullback-Leibler Distance. The Kullback-Leibler Distance, $D(f \parallel g)$ measures the difference between how well the density function f fits data composed according to f and how well the density function g fits data composed according to f. The definition and some properties follow.

- Definition: $D(f \parallel g) \triangleq \mathbf{E}_f \log \frac{f(x)}{g(x)} = \mathbf{E}_f \log f(x) - \mathbf{E}_f \log g(x)$
- Properties:
  1. $D(f \parallel g) \geq 0$.

Proof:

$$\begin{aligned} -D(f \parallel g) &= \mathbf{E}_f \log \frac{g(x)}{f(x)} \\ &\leq \log \mathbf{E}_f \frac{g(x)}{f(x)} \quad (1) \\ &= \log \int \cancel{f(x)} \frac{g(x)}{\cancel{f(x)}} dx \\ &= 0 \quad (2) \end{aligned}$$

Step (1) holds due to Jensen's Inequality. Step (2) holds since $g(x)$ is a probability density function and therefore integrates to 1.

2. $D(f \parallel g)$ is jointly convex in f and g.
   Proof:
   By computation.
3. $\int |f(x) - g(x)| dx \leq 2\sqrt{1 - \exp\{-D(f \parallel g)}$ (Bretagnolle-Huber Inequality).
   Proof:

$$\begin{aligned} -D(f \parallel g) &= \int f(x) \log \frac{g(x)}{f(x)} dx \\ &= \int f(x) \left(\log \left[\min \left(\frac{g(x)}{f(x)}, 1 \right) \right] + \log \left[\max \left(\frac{g(x)}{f(x)}, 1 \right) \right] \right) dx \\ &\leq \log \int f(x) \min \left(\frac{g(x)}{f(x)}, 1 \right) dx + \log \int f(x) \max \left(\frac{g(x)}{f(x)}, 1 \right) dx \\ &= \log \int \min \left(f(x), g(x) \right) dx + \log \int \max \left(f(x), g(x) \right) dx \quad (3) \end{aligned}$$

Now, notice that we have the following relationships.

$$\begin{aligned} \min(a, b) &= \frac{a+b}{2} - \frac{|a-b|}{2} \\ \max(a, b) &= \frac{a+b}{2} + \frac{|a-b|}{2} \end{aligned}$$

Thus, we can rewrite equation (3) as follows.

$$\begin{aligned} & \log \int \min \left(f(x), g(x) \right) dx + \log \int \max \left(f(x), g(x) \right) dx \\ = \; & \log \int \left(\frac{f(x) + g(x)}{2} - \frac{|f(x) - g(x)|}{2} \right) dx \\ & + \log \int \left(\frac{f(x) + g(x)}{2} + \frac{|f(x) - g(x)|}{2} \right) dx \\ = \; & \log \left[1 - \int \frac{|f(x) + g(x)|}{2} dx \right] + \log \left[1 + \int \frac{|f(x) + g(x)|}{2} dx \right] \quad (4) \\ = \; & \log \left[\left(1 - \int \frac{|f(x) + g(x)|}{2} dx \right) \cdot \left(1 + \int \frac{|f(x) + g(x)|}{2} dx \right) \right] \\ = \; & \log \left[1 - \left(\frac{1}{2} \int |f(x) - g(x)| dx \right)^2 \right] \quad (5) \end{aligned}$$

Equation (4) holds because $f(x)$ and $g(x)$ are both probability density functions, and therefore so is $\frac{f(x)+g(x)}{2}$. From equation (5) we obtain the desired result by simply rearranging terms.

Now, going back to density estimation, we have that $\min_\alpha R(\alpha) = \min_\alpha D(p^* \parallel p(\cdot, \alpha))$.