Machine Learning

CSE427

Mahbub Majumdar

BRAC University
66 Mohakhali
Dhaka, Bangladesh

April 24, 2019

Inspiring Excellence

Table of Contents

The Statistical Learning Framework

Learner's Input

- **Domain set** X **:**

- For example,

$$
\begin{aligned}
X &= \{\text{set of players Arsenal can buy}\} \\
X &= \{x_1, x_2, \ldots x_n\}
\end{aligned}
$$

The Statistical Learning Framework

Learner's Input

- **Domain set X :**
- $\chi =$ the set of objects to be labeled.
- For example,

$$\begin{aligned} X &= \{\text{set of players Arsenal can buy}\} \\ X &= \{x_1, x_2, ...x_n\} \end{aligned}$$

The Statistical Learning Framework

Learner's Input

- **Domain set X :**
- $\chi =$ the set of objects to be labeled.
- The labels may be a *vector* of features.
- For example,

$$X = \{\text{set of players Arsenal can buy}\}$$
$$X = \{x_1, x_2, ...x_n\}$$

where,

$$x_1 = \begin{pmatrix} speed_1 \\ skill_1 \\ . \\ . \\ . \end{pmatrix}$$

- Here, speed = feature 1, skill = feature 2,...
- Sometimes, the x_1 are called *instances*.

$$\therefore \text{Domain set} \equiv \text{Instance space}$$

The Statistical Learning Framework

Learner's Input

- **Output values** Y

- For example, Binary labels $\{0, 1\}$

$$\begin{aligned} 1 &= \text{player Arsenal should buy} \\ 0 &= \text{player Arsenal should } \textit{not} \text{ buy} \end{aligned}$$

- **Training Data** S**:** Finite sequence of pairs

$$\begin{aligned} S &= \{(x_1, y_1), (x_2, y_2), ..., (x_k, y_k)\} \\ S &\subset X \times Y \end{aligned}$$

- The learner samples S

The Statistical Learning Framework

Goal of the Learner

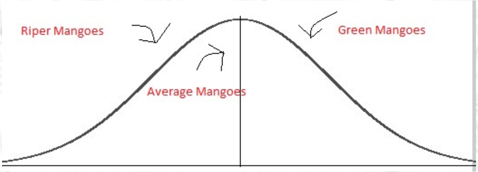
- Obtain a prediction rule

$$h : X \to Y$$

- h is also called a *predictor*, *hypothesis* or *classifier*.
- Predictor h is used to predict *labels* of new *domain points*. For example, suppose $x_i = Lemar$, then h will predict whether Lemar will be a good ($y_i = 1$) or bad ($y_i = 0$) player for Arsenal.
- $A(S) \equiv$ the hypothesis/predictor generated by the data set S.

How to generate the Training Set, S?

- Assume the instances generated by a probability distribution, D
- What does it mean S is generated by D?
  - Suppose X_i = hardness of ith mango
  - Sample n times, then the hardness of the n samples will vary with respect to the same distribution



Mangoes ripeness will not be the same, they will vary w.r.t some distribution D

How to generate the Training Set, S

- Don't assume that the learner knows what D is
- Assume $\exists$ is a correct labeling function

$$f : X \to Y \ \leftrightarrow \ f(x_i) = y_i, \ \forall_i$$

- Labeling function unknown to learner
- Goal of learner is to learn the labeling function
- $(x_i, y_i) \in S$ generated by first sampling points X_i using D, then labeling by f

The Statistical Learning Framework

Error of Classifier

- Error of $h \cong$ probability that the predictor doesn't predict the correct label of a random x_i sampled by D

- Given a set $A \subset X$, then $D(A) =$ probability of observing $x \in A$

$$
\begin{aligned}
A &= \text{An event} \\
\pi &= \text{probability that } A \text{ happens} \\
A &= \{x \in X \mid \pi(x) = 1\} \\
D(A) &\equiv \mathop{\mathbb{P}}_{x \sim D}(\pi(x))
\end{aligned}
$$

The Statistical Learning Framework

Error of Classifier

- Define the prediction error of h as

$$\begin{aligned} L_{D,f}(h) &\cong \underset{x \sim D}{\mathbb{P}}\big(h(x) \neq f(x)\big) \\ &= D\Big(x \mid h(x) \neq f(x)\Big) \end{aligned}$$

- Here f is the correct labeling function

- The error $L_{D,f}$ is the probability of randomly choosing an x for which $h(x) \neq f(x)$

- Subscript *(D,f)* means the error is measured w.r.t. distribution D and correct labeling function, f

Error of Classifier

- **Notation:**

$$L_{D,f}(h) \equiv \begin{cases} risk \\ true\ error \end{cases}$$

$$L \equiv Loss$$

- Underlying Distribution, D is unknown
- Only way for learner to learn D is to sample the environment by observing the training set.

Empirical Risk Minimization

Empirical Risk Minimization

- Learning algorithm does the following :
- Takes input S. Samples generated by distribution D. Each sample x_i labeled by target function f
- Outputs predictor $h_S : X \to Y$
- Goal is to find predictor h_S that minimizes error w.r.t unknown D and f

Empirical Risk Minimization

Empirical Risk Minimization

- To estimate the error, calculate the Training Error, $L_S(h)$

$$L_s(h) \equiv \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m}$$

where, $[m] = \{1, \ldots, m\}$

- **Notation:**: *Empirical Error* $\cong$ *Empirical Risk* $\cong$ *Training Error*

- **Note:** Learner sees the world only through training sample S
  $\Rightarrow$ Look for predictor h minimizing error on S , which is $L_S(h)$

- Minimum error wrt S = **Empirical Risk Minimization(ERM)**

Empirical Risk minimization

Overfitting

- Overfitting occurs when the algorithm tries to mimic the data too closely.
- Often the sample data is imperfect
- Outliers, noise, incorrect measurements, imperfect sampling...
- By trying to reproduce the training data, you may be also reproducing noise.
- Thus, have to worry about overfitting the sample data.

Empirical Risk minimization

Example

Exam Preparation

- Studying for an exam using only past exams
- $\Rightarrow$ Performance on new questions $\rightarrow$ poor
- $\Rightarrow$ Works in Bangladesh *(HSC/SSC)*, Subject *GRE* etc...
  (exams where no creativity is required)

Empirical Risk minimization

Example

Donald Trump predictor

- He observed that countries he visited with white populations like *Norway*, *Sweden*, *Denmark* are good countries.
- Visited only a few non-white countries. There there was corruption. (He is a businessman looking for good deals.)
- Therefore, his training set S is only,

$$S = \{Norway, Sweden, Denmark, USA, corrupt\ brown\ countries\}$$

Empirical Risk minimization

Example

Donald Trump predictor

- His *predictor*

$$h(S) = 1,\ h(everything\ else) = 0$$

- His *hypothesis*: All other countries = shitty countries

- On the sample S,

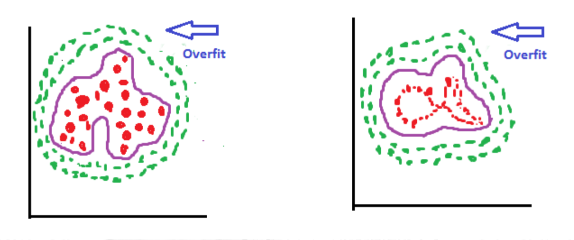
$$L_S(h) = 0\ \Rightarrow \textbf{ERM Estimator}$$

- But the *true error* is quite large, $L_{D,f}(h) \gg 0$.

Overfitting

Standard Model of Particle Physics

- So many *free parameters*...
- Aesthetically, too many *parameters* to parameterize a *fundamental theory*
- **String Theory** is a *unified* theory with 1 free parameter g_s, called the *string coupling* (*not actually true in practice though*)



ERM With Inductive Bias

Preventing overfitting

- **ERM** can lead to *overfitting*
- We want a way to implement **ERM** without *overfitting*

  $\Rightarrow$ *ERM predictor* performs well on S
  $\Rightarrow$ Want it to perform well $\forall x \in X$
  $\Rightarrow$ The x are distributed with distribution D.

- If we apply **ERM** to restricted space of possible hypotheses/predictors, less likely that we will get stuck in a hypothesis very finely tuned to noise.

ERM With Inductive Bias

Preventing overfitting

- Course (not finely tuned) hypotheses will fit the data less well. But, they will not suffer from overfitting.
- Idea: restrict the hypothesis class $\mathcal{H}$.
  - $\mathcal{H} \equiv$ *hypothesis class*
  - $h \in \mathcal{H}$
- For a given class $\mathcal{H}$ and training sample S

  $\Rightarrow$ $ERM_{\mathcal{H}}$ learner uses *ERM rule* to choose a *predictor* $h \in \mathcal{H}$ with lowest possible error over S

  $$\Rightarrow ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\text{argmin}}(L_S(h))$$
- $\underset{h \in \mathcal{H}}{\text{argmin}}(L_S(h)) = \{$the argument h that minimizes $L_S\}$.

ERM With Inductive Bias

Preventing overfitting

- **Introduce bias by restricting to predictors in $\mathcal{H}$**
  $\Rightarrow$ **Inductive Bias**

- Bias formed by knowledge of the problem $\rightarrow$ **Prior Knowledge**

- For example, choose the class $\mathcal{H}$ of players Arsenal Football Club should buy to be the set of *predictors* determined by (*defensive awareness*)

- Choose this way to parameterize the *predictor* because *Arsenal* lack defensively aware players

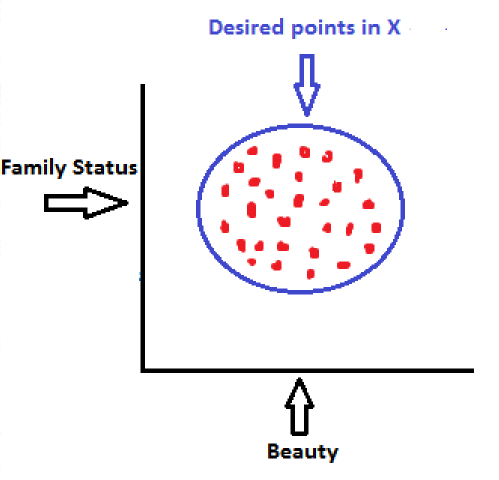
ERM With Inductive Bias

Stupid predictor for a good marriage

- Shallow prior knowledge used by males in the past to construct a predictor for a good marriage
- Males/matchmakers used the class of predictors restricted to the features of (*beauty, family status*)
- Other variables such as, *intelligence, personality,....* were ignored
- Even highly educated males selected spouses based on *beauty, family status*
- Large error in many cases – unhappy/one-sided marriages very common among educated males.

ERM With Inductive Bias

Stupid predictor for a good marriage



ERM With Inductive Bias

ERM With Inductive Bias

- **Fundamental question:** Over which *hypothesis classes* $ERM_{\mathcal{H}}$, will *overfitting* not occur?

- Restricting the *hypothesis classes* via inductive Bias

  $\Rightarrow$ Protects against *overfitting*
  $\Rightarrow$ But causes *stronger inductive bias*

ERM With Inductive Bias

Finite Hypothesis Classes

- Easiest way to restrict *hypothesis class*
  - $\Rightarrow$ Place *upper bound* on number of predictors $h \in \mathcal{H}$
  - $\Rightarrow$ Given a large enough training sample S, a finite class $\mathcal{H}$, we will now find when we can be confident that $ERM_{\mathcal{H}}$ won't overfit
- Recall that the training set S is labeled by $f : X \rightarrow Y$
- h_S is obtained from $ERM_{\mathcal{H}}$ on S, via $h_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ L_S(h)$

Realizability Assumption

Definition:

There exists an $h^* \in \mathcal{H}$ *such that* $L_{D,f}(h^*) = 0$. *This assumption implies that, with probability 1, over random samples* S*, where the instances of* S *are sampled via* D*, and are labeled by* f*, we have* $L_S(h^*) = 0$

Realizability thus means that there is a hypothesis h^* that gives zero true error.

Simplifying Assumptions

Realizability Assumption

- **Realizability's** implications for ERM predictors:
  - The ERM hypothesis has the lowest possible training error
  - The training error will (usually) be less than the true error because the predictor is optimized on S.
  - Thus if the minimum true error is zero, the minimum training error will be zero.
  - The ERM predictor picks the smallest training error predictor

  $\Rightarrow$ $L_S(h_S) = 0$,
- The error of h_S will depend upon how well S captures the information in D

Simplifying Assumptions

Identically Independently Distributed Assumption

- Assume that points in S are obtained by sampling points from D independently of each other
- Sampling point x_i doesn't affect the sample x_j.
- The distribution for point x_j is the same as the distribution for point x_i. This is like sampling with replacement.
- This is called the:
  $\Rightarrow$ *Identically Independently Distributed or iid assumption*

Simplifying Assumptions

Identically Independently Distributed Assumption

- Every $x_i \in S$ is freshly sampled and labeled via f.
- $S \sim D^m$, where $m = |S|$
- D^m is the probability distribution to sample m tuples independently
- The larger S is
  $\Rightarrow$ The more accurately it will reflect the *underlying distribution D*, and *labeling function f.*
- Since S is picked randomly, h_S is a *random variable*.

Representative Sampling

Possible problems

- No guarantee that S will lead to a good predictor h
- $\exists$ a chance that S is non-representative
- For example, Suppose 60% of all £60+ million pound players that are available are *good* players. However, *Arsenal* only samples bad players.

  $\Rightarrow$ $h_S = ERM_{\mathcal{H}}$ will then label every expensive player as a bad investment

  $\Rightarrow$ $\textit{Empirical error} = 0$

  $\Rightarrow$ But, $\textit{true error} = 60\%$

Representative Sampling

Macaulay Example

- Another example: in 1841, the colonialist *Macaulary* wrote:

  "What the horns are to the buffalo, what the paw is to the tiger, what beauty according to old Greek song is to a woman, deceit is to a Bengalee"

  - Maybe this view is because of prejudice
  - Maybe because of bad *training set* data S.
  - Maybe the people Macaulay interacted with exploited fellow Bengalis and were bad people

Accuracy and Confidence Parameters

Definitions

- $\delta =$ *Probability of sampling a non-representative S*
- $1 - \delta =$ *Probability of sampling a representative S*
- $1 - \delta \equiv$ *Confidence parameter*
- Similarly, can't guarantee perfect label prediction
- Introduce parameter for quality of prediction, ϵ
- $\epsilon \ \equiv$ *accuracy parameter*

Accuracy and Confidence Parameters

Definitions

- $L_{D,f}(h_S) > \epsilon \implies$ Failure of learner
- $L_{D,f}(h_S) \leq \epsilon \implies$ Algorithm output approximately correct
- We want to upper-bound the probability to sample *m-tuple* domain points that leads to failure of the learner

Finding the Sample Complexity m

Upper-bounding the Error

- $S|_x \equiv (x_1, \ldots, x_m) \equiv$ instances of the training Set
- Call $S|_x \equiv S_x$
- Want to upper-bound

$$D^m\Big(\{S_x \mid L_{D,f}(h_S) > \epsilon\}\Big)$$

This is the probability of selecting a training set that has large true error

Finding the Sample Complexity m

Upper-bounding the Error

- Let $\mathcal{H}_B$ = *Set of bad hypotheses*

$$\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{D,f}(h) > \epsilon\}$$

- Let M be the set of *misleading samples*

$$M = \{S_x \mid \exists h \in \mathcal{H}_B,\ L_s(h) = 0\}$$

- For every $S_x \in M$, there is a *bad hypothesis* $h \in \mathcal{H}_B$, that looks good from the point of the view of the training sample.

- Want to bound the probability of the event, $L_{D,f}(h_S) > \epsilon$

Finding the Sample Complexity m

Upper-bounding the Error

- Combining *Realizabilty* and the *ERM*,

$\Rightarrow$ $L_S(h_S) = 0$

$\Rightarrow$ $L_{D,f}(h_S) > \epsilon$ can happen only if $\exists h \in \mathcal{H}_B$, for which $L_S(h_S) = 0$

$\Rightarrow$ Will happen only if the sample is in M

$\Rightarrow$ This means

$$\{S_x \mid L_{D,f}(h_S) > \epsilon\} \subseteq M$$

Finding the Sample Complexity m

Upper-bounding the Error

- **Note:** We can rewrite M as

$$M = \bigcup_{h \in \mathcal{H}_B} \{S_x \mid L_S(h) = 0\}$$

- Thus,

$$\begin{aligned} D^m(S_x \mid L_{D,f}(h_S) > \epsilon) &\leq D^m(M) \\ &= D^m\Big(\bigcup_{h \in \mathcal{H}_B} \{S_x \mid L_S(h) = 0\}\Big) \end{aligned}$$

$\uparrow$

(We want to upper bound this)

Finding the Sample Complexity m

Upper-bounding the Error

- **Lemma:** *Union bound*
  For any sets A, B and distribution, D,

$$D(A \cup B) \leq D(A) + D(B)$$

$\Rightarrow$ Clearly follows from known facts

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Then,

$$D^m(S_x \mid L_{D,f}(h) > \epsilon) \leq \sum_{h \in \mathcal{H}_B} D^m(\{S_x \mid L_S(h) = 0\}) \qquad (1)$$

Finding the Sample Complexity m

Upper-bounding the Error

- Now bound each term in equation 1,

$\Rightarrow$ Fix a bad hypothesis $h \in \mathcal{H}_B$

$\Rightarrow$ $L_s(h) = 0$, is equivalent to $h(x_i) = y_i, \forall i$

$\Rightarrow$ Since instances in S are sampled iid,

$$\begin{aligned} D^m\Big(\{S_x \mid L_S(h) = 0\}\Big) &= D^m\Big(\{S_x \mid \forall i, h(x_i) = f(x_i)\}\Big) \\ &= \prod_{i=1}^{m} D\Big(\{x_i \mid h(x_i) = f(x_i)\}\Big) \end{aligned}$$

Since we are sampling x_1 and x_2 and $x_3, \ldots,$ and x_m

Finding the Sample Complexity m

Upper-bounding the Error

- For each individual sampling of an element of the training set

$$\begin{aligned} D(\{x_i \mid h(x_i) = y_i\}) &= 1 - L_{D,f}(h) \\ &\leq 1 - \epsilon \end{aligned}$$

  Since $h \in \mathcal{H}_B$

- Now,

$$\begin{aligned} D^m(S_x \mid L_s(h) = 0) &\leq (1-\epsilon)^m \\ &\leq e^{-\epsilon m} \end{aligned}$$

  Since,

$$1 - \epsilon \leq e^{-\epsilon}$$

Finding the Sample Complexity m

Upper-bounding the Error

- Combining everything,

$$
\begin{aligned}
D^m\Big(\{S_x \mid L_{D,f}(h_S) > \epsilon\}\Big) &\leq \sum_{h\in\mathcal{H}_B} e^{-\epsilon m} \\
&= |\mathcal{H}_B| e^{-\epsilon m} \\
&\leq |\mathcal{H}| e^{-\epsilon m}
\end{aligned}
$$

- We want the probability of a misleading sample to be less than δ.

$$
D^m\Big(\{S_x \mid L_{D,f}(h_S)\Big) < \delta
$$

Finding the Sample Complexity m

Upper-bounding the Error

- Therefore

$$|\mathcal{H}|e^{-\epsilon m} \leq \delta$$

- Solving for the sample complexity m

$$m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Finding the Sample Complexity m

Upper-bounding the Error

Corollary: *Let $\mathcal{H}$ be a finite hypothesis class, and $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer satisfying*

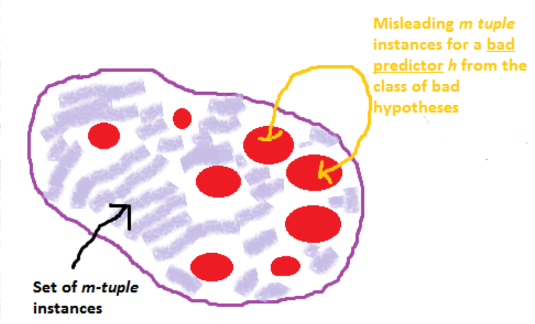
$$m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Then, for any labeling function f, distribution D, for which the realizability assumption holds, (i.e. $\exists h \in \mathcal{H}, L_{D,f}(h) = 0$), with probability of at least $1 - \delta$ over the choice of an iid sample S of size m, we have for every ERM hypothesis h_S:

$$L_{D,f}(h_S) \leq \epsilon$$

- Thus for sufficiently large m, the $ERM_{\mathcal{H}}$ rule over a finite class will probably be correct *(with confidence $1 - \delta$)*.

Graphical Explanation



- For each bad hypothesis, at most $(1-\epsilon)^m$ fraction of the training sets are misleading
- The larger m is, the smaller the red regions are.

Graphical Explanation

- $\Rightarrow$ Area of the space of training sets that are misleading is at most the sum of the areas of the red regions
- $\Rightarrow$ Therefore, it is bounded by $|\mathcal{H}_B|\times$ maximum area of the red regions
- $\Rightarrow$ Any sample outside the red regions doesn't cause overfitting.